Université de Montréal

L'usage des codons régule la présentation des peptides associés aux molécules du CMH-I

par

Tariq Daouda

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.

Janvier 2018

Résumé

Les peptides associés au CMH-I (MIP) sont une pierre angulaire du système immunitaire adaptatif. Ces courtes séquences d'acides aminés présentées à la surface des cellules par les molécules du CMH-I, régulent le développement des lymphocytes T et jouent un rôle majeur dans la reconnaissance par le système immunitaire des cellules cancéreuses et des cellules infectées par des virus. Puisque le système immunitaire utilise les MIP pour identifier les cellules anormales, pouvoir identifier l'ensemble des MIP présents à la surface des cellules permettrait donc de concevoir des vaccins personnalisés contre cancers et virus. Cependant, les règles derrière la genèse des MIP sont encore à découvrir. Les prédicteurs algorithmiques existant ont de hauts taux de faux positifs, et par conséquent, les seules méthodes fiables haut débit pour l'identification des MIP dépendent fortement de la spectrométrie de masse (MS). Ces méthodes nécessitent des mois de culture cellulaire, ce qui les rend inapplicables dans un contexte clinique personnalisé.

Dans cette thèse, nous introduisons d'abord une nouvelle méthode pour l'identification des MIP à travers la MS. Cette méthode intègre les polymorphismes génomiques spécifiques au dans la base de données utilisée pour l'identification. Cette amélioration nous a permis d'augmenter considérablement la précision de l'identification et nous a permis d'identifier 34 nouveaux antigènes mineurs d'histocompatibilité pour l'immunothérapie du cancer. Nous présentons ensuite le logiciel que nous avons écrit pour la génération de bases de données MS personnalisées : pyGeno, une base de données optimisée pour la création et l'exploration de génomes personnalisés.

Enfin, en utilisant un ensemble de données de plus de 117k séquences d'ARN, recueillies en utilisant notre processus de protéogénomique sur les cellules de 18 sujets sains exprimant un total de 33 allèles CMH-I, nous démontrons que l'utilisation des codons influence la présentation MIP. Plus spécifiquement, nous montrons que les codons synonymes ont des effets différents sur la présentation, et que cet effet dépend de la position du codon par rapport à la portion de l'ARNm codant pour le MIP. Nos résultats suggèrent fortement que les règles que nous avons extraites en utilisant des réseaux neuronaux artificiels (ANN) sont conservées à travers les espèces et les types cellulaires. Cette conservation implique à son tour qu'il existe un ensemble de MIP candidats identifiables par l'intermédiaire d'un ensemble conservé de règles, à partir desquelles les MIP liant les allèles du MHC-I sont sélectionnés. Nos résultats montrent que la construction d'un prédicteur MIP à partir de séquences d'ARNm pourrait être possible. Finalement, ils illustrent comment des ANN peut être utilisés comme moyen efficace pour extraire des informations pertinentes biologiques.

Mots-clés : Peptides associés aux molécules du CMH-I, Réseaux de Neurones Artificiels, Immunothérapie du cancer, Prédiction de néoantigènes, Protéogénomique, Spectrométrie de Masse, Intégration de mots, Usage de codons, ARNm, Développement logiciel

Abstract

MHC-I associated peptides (MIPs) are a cornerstone of the adaptive immune system. They are short sequences of amino-acids presented at the surface of cells by the MHC-I molecules. They regulate the development of T lymphocytes and play a major role in the recognition by the immune system of cancerous and virus infected cells. Because the immune system uses MIPs to identify abnormal cells, being able to identify the set of MIPs present at the surface of cells would therefore enable us to design personalized vaccines against cancer and viruses. However, the rules behind the genesis of MIPs are still to be discovered. Current predictors have high false positive rates, and the only reliable, high throughput methods for identifying MIPs heavily depend upon Mass Spectrometry (MS). They require months of cell culture, rendering them inapplicable in a personalized clinical context.

In this thesis, we first introduce a new method for identifying MIPs through MS. This method, by integrating subject-specific genomic polymorphisms into MS databases, allowed us to greatly increase identification precision. This increased precision allowed us to discover 34 new Minor Histocompatibility Antigens for cancer immunotherapy. We then introduce the software we wrote for the generation of personalized MS databases: pyGeno, an optimized database for creating and exploring custom personalized genomes.

Finally, using a dataset of over 117k RNA sequences, gathered using our proteogenomics pipeline from 18 healthy subjects expressing a total of 33 MHC-I alleles, we demonstrate that codon usage influences MIP presentation. More specifically, we show that synonymous codons have different effects on the presentation, and that this effect depends upon the position of the codon relative to the mRNA encoding the MIPs. Our results strongly suggest

that the rules we have extracted using Artificial Neural Networks (ANN) are conserved across species and cell types. This in turn implies that there exists a set of candidate MIPs identifiable through a conserved set of mRNA rules from which MHC-I binders are selected. Finally, they also suggest that building a MIP predictor from mRNA sequences might be possible and illustrate how ANN can be used as effective means for extracting relevant biological information.

Keywords: MHC-I associated peptides, Artificial neural network, Cancer immunotherapy, Neoantigen prediction, Proteogenomics, Mass Spectrometry, Word embeddings, mARN, Codon usage, Software development

Table des matières

Liste des tableaux

Liste des figures

Liste des contributions scientifiques

Outils informatiques (disponibles sur https://github.com/tariqdaouda)

1. Mariana: Cadriciel et langage pour l'élaboration de réseaux de neurons profonds
2. pyGeno: Outil pour la génomique et proteomique personalisé. Optimisé pour fonctionner sur un ordinateur portable.
3. RabaDB: Outils de gestion de base données relationnel écris pour pyGeno et optimisé pour une faible utilisation de la mémoire.

Publications

En tant que premier auteur :

1. Daouda T*, Jeremie Zumer, Perreault C. Lemieux S. "Holographic Neural Architectures". https:/arxiv.org/abs/1806.00931, Submitted to NIPS. 2018
2. Daouda T*, Dumont-Lagacé M, Thibault P, Bengio Y, Lemieux S, Perreault C. "Codon usage regulates biogenesis of MHC-I-associated peptides". Published as part of PhD Thesis. 2017
3. Daouda T*, Perreault C, Lemieux S. "pyGeno: A Python package for precision medicine and proteogenomics". F1000Res. 2016
4. Granados DP*, Sriranganadane D*, Daouda T*, Zieger A, Laumont CM, Caron-Lizotte O, Boucher G, Hardy MP, Gendron P, Côté C, Lemieux S, Thibault P, Perreault C. "Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides". Nat Commun. 2014

En tant que co-auteur :

1. Dumont-Lagacé M, Gerbe H, Daouda T, Laverdure JP, Brochu S, Lemieux S, Gagnon É, Perreault C. "Detection of Quiescent Radioresistant Epithelial Progenitors in the Adult Thymus". Front. Immunol. 2017
2. Pearson H, Daouda T, Granados DP, Durette C, Bonneil É, Courcelles M, Rodenbrock A, Laverdure JP, Côté C, Mader S, Lemieux S, Thibault P, Perreault C. "MHC class I-associated peptides derive from selective regions of the human genome". J Clin Invest. 2016
3. Laumont CM, Daouda T, Laverdure JP, Bonneil É, Caron-Lizotte O, Hardy MP, Granados DP, Durette C, Lemieux S, Thibault P, Perreault C. "Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames". Nat Commun. 2016
4. Granados DP, Yahyaoui W, Laumont CM, Daouda T, Muratore-Schroeder TL, Côté C, Laverdure JP, Lemieux S, Thibault P, Perreault C. "MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements". Blood. 2012
5. Heraz A, Daouda T, Frasson C. "Decision Tree for Tracking Learner's Emotional State predicted from his electrical brain activity". International Conference on Intelligent Tutoring Systems. 2008.

Présentations orales

1. "Neural networks unveil the role of RNA translation in shaping the Immune-self". Canadian Cancer Immunotherapy Consortium. 2017

2. “The Immune-self lost in translation”. Institute for Research in Immunology and Cancer. 2017
3. “Proteogenomic exploration of the Immune-self”. McMaster University, 2017
4. “Mariana, The cutest Deep Learning Framework”, Montreal Institute for Learning Algorithms, private presentation. 2017
5. “pyGeno: A Python package for precision medicine and proteogenomics.” Montreal Bioinformatics User Group. 2016
6. “From RNA to MHC-I Peptides: A Machine Learning approach.” Institute for Research in Immunology and Cancer. 2016
7. “Mariana, The Cutest Deep Learning Framework”, Montreal Institute for Learning Algorithms, Tea Talk. 2016
8. “Mariana, The Cutest Deep Learning Framework”, Machine Learning Paris. 2016
9. “Untranslated, really? The secret role of non-coding RNAs in the adaptive immune system”. Université de Sherbrooke. 2014
10. “What’s the impact of Genomic Polymorphisms on the Immune-self?”. Institute for Research in Immunology and Cancer. 2013

Posters

1. “Codon usage regulates biogenesis of MHC-I associated peptides”. Canadian Cancer Immunotherapy Consortium. 2017
2. “Predicting MHC-I associated peptides from contextual RNA sequences using Deep Artificial Neural Networks”. Summit for Cancer Immunotherapy. 2016

Liste des abréviations

ANN	Réseaux de neurones artificiels ou *Artificial neural networks*
ARNt	ARN de transfert
AUC	*Area under the receiver operator characteristic curve*
B-LCL	*Epstein-Barr virus (EBV)-transformed B cell lines*
CFSE	*Carboxyfluorescein succinimidyl ester*
DC	*Dendritic cell(s)*
DRiP	Produits ribosomaux défectueux ou *Defective ribosomal products*
EBV	*Epstein-Barr virus*
GO	*Gene Ontology*
GVL	Réaction du greffon contre la leucémie ou *graft-versus-leukemia*
INDEL	Insertions/délétions
KL	Kullback-Leibler
MCC	*MIP-coding codons*

MHC-I	Complexe majeur d'histocompatibilité de classe I ou *Major Histocompatibility Complex of class I*
MiHA	Antigènes mineurs d'histocompatibilité ou *Minor histocompatibility antigens*
MIP	Peptides associés aux molécules du MHC-I ou *MHC-I associated peptides*
MLP	Perceptron à multiples couches or *Multi-layer perceptron*
mRNA	ARN messager ou *Messenger RNA*
MS	Spectrométrie de masse ou *Mass spectrometry*
ns-SNP	*Non-synonymous single nucleotide polymorphism*
PBMC	*Peripheral blood mononuclear cells*
RDP	Protéines rapidement dégradées ou *Rapidly degraded protein*
RNA-seq	Séquençage d'ARN/de transcriptome ou *RNA/transcriptome sequencing*
SGD	Descente de gradient stochastique ou *Gradient Stochastique Gradient Descent*
SNP	*Single nucleotide polymorphism*
SUP	Système ubiquitine protéasome
TAA	Antigène associé aux tumeurs ou *Tumor associated antigens*
TCR	Récepteur de cellules T ou *T cell receptor*
TSA	Antigènes spécifiques aux tumeurs ou *Tumor specific antigens*

Two decades ago I was a little boy who dreamt of creating an artificial intelligence and curing cancer, since then very little has changed

Remerciements

A few bytes of text won't suffice to express my gratitude or to summaries what I have learned during the last years. I thank my parents without whom nothing would have been possible, who supported me through unpleasant school years and way more enjoyable university years. My supervisors who wholeheartedly supported my growth as a researcher and who I now consider my friends. Elaine Meunier whose otherworldly efficiency and usefulness so much facilitated every administrative process. And of course, Maude who shares my life and my work, who formatted this thesis and proofreads everything I write.

In no particular order I also thank: Logan Shwartz, Diana Poalo Granados, Assya Trofimov, Patrick Gendron, Jean-Philippe Laverdure, Geneviève Boucher, Éric Audemard, Jonathan Séguin, Céline Laumont, Dev Sriranganadan, Alexandre Rouette, Antione Zieger, Yahya Benslimane (John Solomonson), Mohamed Benhamadi, Olivier Caron-Lizotte, Danièle De Verteuille, Caroline Labelle, Hillary Pearson. Everyone that I had the chance to meet during this few years in the Perreault Lab, and the whole IRIC community. I also thank members of the AÉÉBINUM that I had the honor to preside for two years: Sarah Cherkaoui, Armande Ang Houle, Saraï Mola, Naïm Poonja-Tremblay, Pamela Mehana, Philippe Malric, Emmanuel Noutahi, Marc Legeault. My PhD was punctuated by one unelated C2 fracture and one unrelated ACL surgery. This work herein presented is a sacrifice of myself to myself. Like Odin hanging from the tree of life, six years later I am now reborn, transformed by the process in a way that I could not foresee. It has been a fantastic ride. Thank you all. C >>

Chapitre 1

Introduction

L'efficacité du système immunitaire dépend de la synergie entre le système immunitaire inné et le système immunitaire adaptatif. Le premier est dit inné puisqu'entièrement défini par les caractéristiques génétiques de l'individu[1]. Ce système constitue une première ligne de réponses rapides, mais non spécifiques aux pathogènes rencontrés. Le système adaptatif, de son côté, est développé au cours de l'existence et modelé par l'environnement. Sa réaction initiale est en général plus lente, mais hautement spécifique et précise[1]. C'est cette spécificité du système immunitaire adaptatif qui lui permet de jouer un rôle essentiel dans le repérage et l'élimination des cellules infectées par des virus et des cellules cancéreuses. L'exploitation de l'adaptabilité du système immunitaire adaptatif a permis l'une des plus grandes avancées médicales jamais accomplies : la vaccination[1]. Aujourd'hui, l'immunothérapie du cancer propose à son tour d'exploiter cette adaptabilité afin d'améliorer la réponse immunitaire face au cancer[2,3]. Pour arriver à cette fin, il est nécessaire de comprendre les mécanismes qui permettent au système immunitaire adaptatif de reconnaitre les cellules infectées et les cellules néoplasiques. En d'autres termes, d'élucider les mécanismes moléculaires de la définition du soi biologique. Ceci implique de comprendre la genèse de l'immunopeptidome.

1.1 L'immunopeptidome

Le rôle de l'immunopeptidome est de présenter à la surface des cellules une représentation de leur état interne[4]. Il se compose de l'ensemble des peptides présentés par les molécules du complexe majeur d'histocompatibilité de classe I (*Major Histocompatibility Complex of class I*, MHC-I) à la surface des cellules nucléées. Ces peptides, auxquels nous référerons dorénavant sous le nom de peptides associés aux molécules du MHC-I (*MHC-I Associated Peptides*, MIP), proviennent principalement de la dégradation de protéines endogènes, mais peuvent également provenir de régions dites cryptiques (toutes régions autres que celles correspondant au cadre de lecture canonique)[5]. Les gènes encodant les molécules du MHC-I proviennent des régions les plus polymorphiques du génome humain[6]. De plus, chaque personne possède 6 allèles différents encodant des MHC-I, chacun avec ses propres patrons de liaison[1], et de ce fait, exerçant une sélection forte sur les séquences des MIP présentés.

1.1.1 Rôles des MIP dans la définition du soi

L'immunopeptidome joue un rôle essentiel dans la sélection des lymphocytes T. Les lymphocytes T sont produits dans la moelle des os longs puis migrent vers le thymus pour leur maturation et sélection. Ils possèdent des récepteurs de cellules T (*T-Cell Receptors*, TCR) produits à partir de réarrangements aléatoires[1] et devront subir deux étapes de sélection avant d'atteindre leur état mature et fonctionnel. La première, appelée « *sélection positive* », se passe au moment de la rencontre entre les lymphocytes et les cellules épithéliales du cortex thymique. Cette étape vérifie que les TCR des lymphocytes sont fonctionnels et capables d'interagir avec les molécules du MHC-I[1]. Les lymphocytes aux TCR non fonctionnels ne reçoivent pas de

signal de survie et meurent par négligence[1]. La deuxième étape, appelée « *sélection négative* », se passe au contact des cellules de la médulla thymique. Durant cette étape, les lymphocytes capables de reconnaitre des MIP du soi, et donc susceptibles d'être à l'origine de réactions auto-immunes, sont éliminés[1,7]. Les cellules de la médulla thymique, entre autres par l'expression du gène *AIRE*, sont en effet capables d'exprimer et de présenter des MIP dont l'expression est normalement restreinte à des tissus spécifiques[7–9]. En permettant l'élimination des lymphocytes T auto-réactifs, l'immunopeptidome du thymus joue donc un rôle central dans la prévention des maladies auto-immunes. Il représente le soi biologique du point de vue du système immunitaire adaptatif.

1.1.2 Rôles des MIP dans l'immunothérapie du cancer

En temps normal, l'immunopeptidome est composé intégralement de MIP issus de séquences normales du soi. En revanche, lorsqu'une cellule est infectée par un virus, le rôle des molécules du MHC-I est de présenter des MIP originaires des protéines virales. De même, lorsque des cellules deviennent cancéreuses, les MIP issus de protéines mutées permettent aux lymphocytes T de repérer et d'éliminer ces cellules. C'est ce mécanisme d'identification et d'élimination des cellules néoplasiques qui fait des MIP d'excellentes cibles de vaccins anti-cancers[10]. Trois catégories de MIP sont particulièrement importantes pour l'immunothérapie :

1. Les antigènes associés aux tumeurs (*Tumour Associated Antigens, TAA*) : MIP présents chez les cellules saines, mais plus fortement exprimés chez les cellules cancéreuses.
2. Les antigènes mineurs d'histocompatibilité (*Minor Histocompatibility Antigens, MiHA*): MIP issus de régions polymorphiques saines.

3. Les antigènes spécifiques aux tumeurs (*Tumour Specific Antigens, TSA*) : MIP issus de polymorphismes propres aux tumeurs et ne sont donc pas présentés par les cellules saines.

En raison de leur présence chez les cellules saines, les TAA ne confèrent dans les meilleurs des cas qu'une très faible immunité[11]. Les MiHA en revanche ont été démontrés chez la souris comme étant capables de conférer une immunité forte face aux leucémies[11]. Les MiHA possèdent néanmoins certains inconvénients notables :

1. Leur utilisation est restreinte par le génome du patient. En effet, pour produire une réaction du greffon contre la leucémie, le donneur et le receveur doivent à la fois posséder les mêmes allèles de molécules du MHC-I et posséder des différences polymorphiques dans des régions du génome correspondant au MiHA.
2. Même si aucune toxicité n'a été observée à ce jour[12], les MiHA peuvent en théorie être à l'origine de réactions auto-immunes, puisque les polymorphismes à l'origine du rejet tumoral ne sont pas spécifiques aux cellules leucémiques, mais retrouvés dans le génome de toutes les cellules du patient.

De ces trois catégories, les TSA sont donc les MIP qui semblent *a priori* les plus prometteurs pour l'immunothérapie personnalisée du cancer. Puisque les TSA sont spécifiques à la tumeur, ils seraient à la fois fortement immunogéniques et incapables d'être à l'origine de réactions auto-immunes. De plus, contrairement aux MiHA, leur utilisation n'est pas restreinte par les polymorphismes génétiques du patient. Les TSA possèdent également certains désavantages :

1. L'hétérogénéité des tumeurs peut faire en sorte que toutes les cellules tumorales ne possèdent pas les mêmes polymorphismes[13–15] et donc présentent des TSA différents. Pour être efficace, un TSA doit donc être partagé par un nombre élevé de sous-clones.
2. Les cellules cancéreuses présentant les TSA les plus immunogéniques sont peut-être déjà reconnus et éliminés par le système immunitaire. Ce phénomène est communément appelé « *immuno-editing* » en anglais[16].

Ces points suggèrent qu'une stratégie de vaccination efficace anti-cancer devrait sans doute combiner MiHA et TSA. Ils expliquent également pourquoi l'identification de TSA reste pour le moment une tâche extrêmement ardue qui se rapproche à chercher une aiguille dans une botte de foin. Historiquement, la découverte de TSA a nécessité des procédures laborieuses et n'ont mené la plupart du temps à l'identification que d'un seul TSA à la fois. Récemment, deux équipes indépendantes ont identifié un total de 10 TSA provenant de mélanomes de quatre sujets humains. Malgré cette amélioration, le nombre de TSA identifiés reste surprenamment bas, surtout considérant que le mélanome est le type de cancer le plus muté[17]. De plus, les seules méthodes permettant l'identification haut-débit de MIP dépendent de la spectrométrie de masse (MS, *mass spectrometry*). Ces méthodes requièrent des centaines de millions de cellules par réplicas et, par conséquent, nécessitent des mois de culture cellulaire. Ces contraintes les rendent difficilement applicables aux tumeurs solides (ce qui limite considérablement les avenues de recherche) et inapplicables dans un contexte clinique. S'affranchir de ces contraintes entraine donc la nécessité d'élucider les mécanismes derrière la genèse de l'immunopeptidome afin d'élaborer des méthodes d'identification de MIP haut-débit, rapides et précises.

1.1.3 Mécanismes de genèse de l'immunopeptidome

Le processus à l'origine de la présentation des MIP à la surface des cellules peut être modélisé comme le résultat de deux phases successives : (i) la génération des MIP, (ii) la présentation des MIP à la surface par les molécules de MHC-I (Figure 1).

La deuxième phase est celle qui est le mieux caractérisée, puisqu'elle dépend seulement de l'affinité du peptide aux molécules du MHC-I du sujet. Les patrons d'affinité de plusieurs molécules du MHC-I ont été décrites, et on est capable aujourd'hui d'obtenir des prédictions fiables de l'affinité de séquences peptidiques pour ces molécules grâce à des algorithmes d'apprentissage machine[18]. D'ailleurs, l'utilisation de ces algorithmes de prédictions s'est depuis imposée comme pratique standard dans l'étude de l'immunopeptidome[6,18–20]. La première phase, quant à elle, implique plusieurs étapes complexes de dégradation et de transformation des chaines d'acides aminés. Si les étapes générales de cette phase sont connues, les impacts exacts de chacune d'elles sur la présentation restent encore à découvrir.

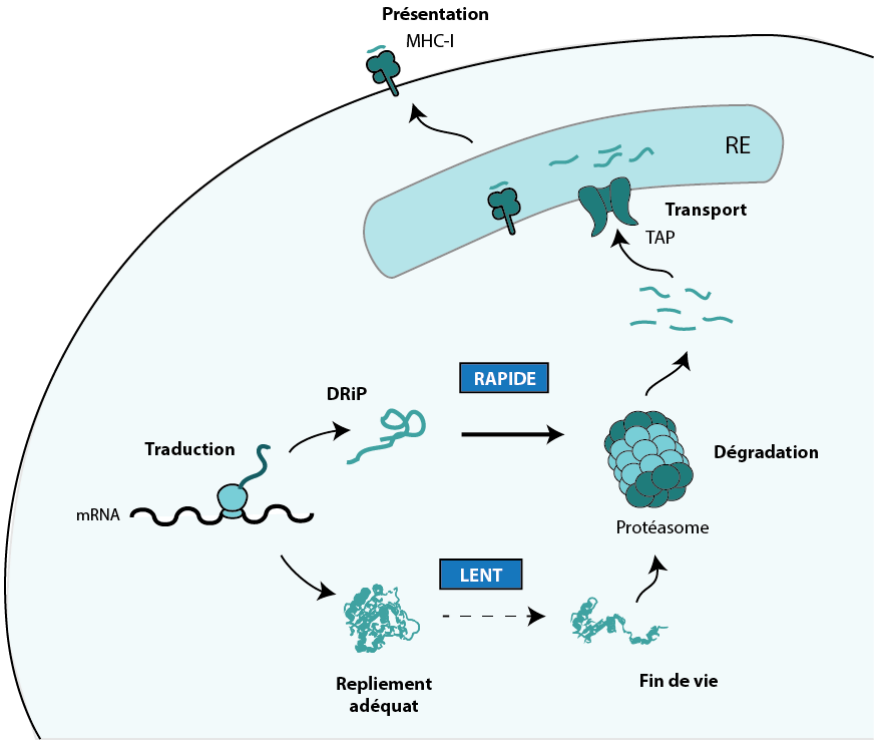


Figure 1. Présentation des MIP à la surface cellulaire. Le processus commence par la traduction de mRNA et le repliement de la protéine. Les protéines en fin de vie sont ensuite dégradées en peptides par le protéasome, qui seront ensuite transportés dans le réticulum endoplasmique (RE) par TAP, où ils se lient aux molécules du MHC-I en fonction de l'affinité de celles-ci pour leurs séquences. Les complexes formés par les peptides et les molécules du MHC-I migrent ensuite vers la membrane. La théorie des DRiP stipule que si une erreur provient durant la traduction ou le repliement de la protéine, la chaine naissante passe directement à la phase de dégradation, ce qui cause une présentation rapide des MIP issues de cette dégradation.

1.1.4 La génération de MIP n'est pas aléatoire

Malgré l'importance primordiale de la liaison des peptides aux molécules du MHC-I, cela ne permet pas d'expliquer complétement la sélection des régions protéiques générant des MIP. Il n'existe à ce jour aucune méthode permettant la prédiction des MIP présents à la surface d'une cellule à partir de son génome, de son transcriptome ou de son protéome[18].

L'immunopeptidome n'est pourtant pas aléatoire, comme le montre les résultats de Pearson *et al.*[20] (article disponible en Annexe 1). Cette étude, réalisée sur 25,270 MIP issus de 18 sujets exprimant au total 27 allèles du MHC-I, indique que 41% des gènes ne produisent aucun MIP. Comme le montrent les auteurs, la sous-représentation de certains gènes ne peut être expliquée par un échantillonnage aléatoire non biaisé du génome. En effet, le nombre de gènes sources prédits par une distribution binomiale est largement supérieur à celui observé ($p < 1 \times 10^{-250}$). Les résultats montrent également que le nombre de MIP uniques augmente quasi-linéairement avec le nombre d'allèles du MHC-I, alors que l'augmentation du nombre de gènes sources décroit avec l'ajout de nouveaux allèles[20]. La conclusion est donc que malgré la quasi-orthogonalité des patrons de fixation des allèles du MHC-I, il existe un sous-ensemble restreint de gènes surreprésentés dans l'immunopeptidome. Plus intriguant encore, cet article rapporte également que seulement 10% de l'exome est représenté dans l'immunopeptidome. Les MIP ont tendance à provenir des mêmes régions exomiques au sein des gênes sources, que les auteurs nomment « *hot spots* », indépendamment de l'allèle du MHC-I qui les présente.

Ces résultats montrent que connaitre les patrons de liaisons des molécules du MHC-I est donc nécessaire, mais insuffisant pour expliquer la présentation des MIP et sugèrent l'existence de règles de présentation indépendantes des allèles de MHC-I. Ceci implique que les règles complètes de présentation des MIP ne peuvent être déduites avec certitude en observant les MIP présenté par un allèle du MHC-I à la fois. Il est donc nécessaire d'étudier des ensembles de données contenant des dizaines de milliers de MIP présentés par plusieurs allèles du MHC-I différents.

1.1.5 Peut-on prédire l'immunopeptidome?

L'étude de Pearson *et al.*[20] répond également à une question fondamentale : L'immunopeptidome est-il prédictible? Selon les résultats exposés plus haut, la réponse théorique est indubitablement oui, puisque l'immunopeptidome n'est pas le résultat d'un échantillonnage aléatoire. Les résultats de l'article vont aussi plus loin en prédisant les gènes sources de peptides au moyen d'une régression logistique qui prend en entrée plusieurs indices sur les gènes. Les deux indices les plus importants pour la prédiction sont directement mesurables : l'expression du gène (résultat également rapporté dans des études précédentes[19,21], et la longueur de la protéine (les protéines plus longues ayant plus de chance de produire des peptides). Les indices suivants sont quant à eux composés de la présence de *Gene Ontology* (*GO)-terms* tel que « *Nucleolus* » ou « *Protein complex* », et de résultats issus d'algorithmes de prédictions divers, prédisant par exemple la structure secondaire des protéines ou le désordre intrinsèque des acides aminés. Les *GO-terms* sont par nature vagues. Au moment de l'écriture de cette thèse, « *Nucleolus* » réfère à 1,259 gènes chez l'humain, « *Protein complex* » à 6,744. Les algorithmes de prédiction, quant à eux, ont une imprécision intrinsèque, et introduisent invariablement un biais pour lequel il est difficile de contrôler. Ces indices se rapportent également aux gènes complets et par conséquent ne sont pas spécifiques aux « *hot-spots* », ces régions exomiques dont dérivent la plupart des MIP.

L'existence de ces *hot-spots* suggère fortement la présence de patrons localisés dans la séquence qui prédisposent ces régions à être source de MIP. Étant donné que ces régions sont communes aux 18 sujets de l'étude, et que les MIP ne proviennent pas de régions particulièrement polymorphiques[17,19], ces patrons devraient donc être conservés à travers

l'espèce humaine. Une question importante reste donc à résoudre : qu'est-ce qui, dans la séquence des gènes sources, les prédisposent à être générateurs de MIP?

1.2 La théorie des DRiP

L'arrêt de la synthèse protéique provoque une diminution de la disponibilité des peptides intracellulaires interagissant avec TAP, et ce, aussi rapidement que 30 minutes après l'ajout d'inhibiteurs de traduction[21–23]. Une proportion substantielle des MIP proviendrait donc de protéines rapidement dégradées (*Rapidly Degraded Protein*, RDP). De plus, si l'immunopeptidome était composé exclusivement de peptides issus de la dégradation de protéines en fin de vie, la vitesse de présentation des MIP devrait refléter la demi-vie des protéines sources. Or, les lymphocytes T CD8+ cytotoxiques sont capables de reconnaitre des MIP issus de protéines virales quelques dizaines minutes après infection, alors que les demi-vies des protéines sources sont de l'ordre de quelques jours[22,23]. Ces observations impliquent que ces MIP résultent d'un processus de dégradation de polypeptides se produisant très tôt dans la vie des protéines. Pour expliquer cette présentation rapide, Yewdell *et al*. proposent en 1996 la théorie selon laquelle la plupart des MIP proviendrait de Produits Ribosomaux Défectueux (*Defective ribosomal products*, DRiP)[24]. Les DRiP sont des polypeptides qui n'atteignent pas leur conformation native et sont dégradés pendant ou immédiatement après leur traduction[21]. Les DRiP constitueraient une part importante du polypeptidome puisqu'entre 25 et 30% des protéines seraient dégradées seulement quelques minutes après leur synthèse[21,25].

L'hypothèse des DRiP fait encore l'objet de débats, et aucun DRiP n'a encore été isolé *in vivo*. Les protéines mal repliées peuvent être réhabilitées par des chaperonnes[26–29], et certaines

études suggèrent que les protéines matures sont autant susceptibles de produire des MIP que les protéines nouvellement synthétisées[26]. De plus, pour que la présentation des MIP puisse se faire, les polypeptides sources doivent être traités par le Système Ubiquitine Protéasome (SUP)[21,30,31]. Or, l'ubiquitination d'un polypeptide n'implique pas nécessairement sa dégradation, l'ubiquitination étant un phénomène réversible[28]. Tous les polypeptides ubiquitinés ne sont pas dégradés par le protéasome, puisque les protéines mal repliées peuvent former des agrégats dégradés par autophagie, également initiée par ubiquitination[28,32].

L'existence des DRiP impliquerait une dégradation rapide des polypeptides synthétisés, ce qui nécessiterait une ubiquitination pendant, ou immédiatement après la synthèse. Or, il existe plusieurs mécanismes protégeant les protéines d'une ubiquitination hâtive, et par conséquent la majorité des protéines ne sont pas ubiquitinées pendant leur synthèse[29]. *A priori*, ces résultats peuvent être interprétés comme étant contraire à l'hypothèse des DRiP. Néanmoins, deux des quatre critères identifiés comme favorisant l'ubiquitination des chaines naissantes (longueur de la protéine et forte hydrophobicité) sont également enrichis dans les gènes sources de MIP[20] et sont donc en accord avec l'observation selon laquelle une majorité de MIP proviendraient d'un sous-ensemble restreint de gènes.

La dégradation des polypeptides via le SUP est un processus plus ancien que la présentation des MIP puisqu'il est retrouvé même chez la levure. Serait-il possible que le SUP ait évolué à la fois comme un mécanisme central de régulation de la protéostasie et une première ligne de défense contre les infections qui dégraderait en priorité les polypeptides d'origine infectieuse? Selon cette théorie, le SUP posséderait une préférence pour les polypeptides possédant des caractéristiques différentes de celles de l'hôte. Un tel bais induirait inévitablement un biais au niveau des sources de MIP présentés à la surface.

Remarquablement, la vitesse de traduction de l'ARN messager (RNA) a aussi été identifié comme un facteur important participant à l'ubiquitination des chaines naissantes [29]. En outre, la présentation de MIP corrèle avec l'expression du mRNA[6,19,20], mais non avec l'abondance des protéines sources[21]. Ces caractéristiques pourraient donc être liées aux chaînes de polypeptides dégradées tout comme aux mRNA dont elles découlent.

1.2.1 Le biais de codons

La traduction et l'élongation des protéines est un processus complexe impliquant plusieurs acteurs intervenant à diverses étapes. Le code de l'ADN/ARN est composé de mots de 3 nucléotides appelés codons. Il existe donc 64 codons possibles qui ne sont traduisibles qu'en 20 acides aminés et un stop marquant la fin de la séquence à traduire. Le code de l'ADN est donc redondant (dit dégénéré), un même acide aminé pouvant être traduit par jusqu'à 6 codons différents. Mais si ces codons sont synonymes, ils ne sont pas forcément équivalents. La fréquence d'utilisation de chaque codon, aussi appelée *biais de codons*, est aujourd'hui reconnue comme un mécanisme important de régulation de l'expression génique, influençant le traitement de l'ARN, la traduction et le repliement des protéines[29,33–37].

1.2.2 Le biais de codon régule l'expression génique

Le concept central derrière cette influence est l'optimalité des codons. De manière générale, les protéines fortement exprimées sont encodées par des gènes avec une proportion relativement élevée de codons reconnus par des ARN de transfert (ARNt) abondants[36,37]. L'utilisation de codons fréquents augmente par conséquent la vitesse et l'efficacité de

traduction[38]. Optimiser la séquence de gènes hétérologues en remplaçant des codons peu fréquents chez l'hôte par des synonymes plus fréquents peut dans certains cas grandement augmenter l'expression de ces gènes[33,36].

D'un côté, les codons particulièrement adaptés au bassin d'ARNt seraient préférentiellement utilisés dans les gènes fortement exprimés, parce que ceux-ci sont soumis à une pression élevée pour une traduction précise et efficace[33,36,39]. Une traduction efficace pourrait fournir un bénéfice global à la cellule en augmentant le nombre de ribosomes disponibles. Une traduction précise permettrait de réduire les effets néfastes causés par une production de protéines mal traduites[33,39]. D'un autre côté, les codons non-optimaux peuvent être utilisés dans des sites nécessitant une pause du ribosome pour permettre à la protéine de se replier correctement[33]. L'utilisation de codons non-optimaux peut avoir d'importantes conséquences. Chez *E. coli*, des répétitions AGA ou AGG peuvent entrainer une suite de phénomènes identifiés comme possible source de DRiP : arrêt du ribosome, un clivage du mRNA, un changement de cadre de lecture du ribosome ou des erreurs de traduction[33].

1.2.3 La nature du biais

La fréquence d'utilisation de codons synonymes n'est pas le seul biais observable, il existe aussi un biais dans l'assortiment des codons dans les séquences de gènes[36,37]. Les codons synonymes reconnus par les mêmes ARNt ont tendance à s'agglutiner[34,36]. Cet effet de co-occurrence touche à la fois les codons les plus fréquents et les plus rares. Il est néanmoins particulièrement important dans les gènes fortement exprimés qui doivent être rapidement induits, comme par exemple les gènes en lien avec des réponses au stress[36]. Il existe également

un biais important au niveau de l'utilisation de paires de codons. Au vu de leur fréquence d'utilisation, on s'attendrait à ce que la paire d'acides aminés Ala-Glu soit encodée autant par la paire codons GCC-GAA que par GCA-GAG. Or, chez l'humain, la paire GCC-GAA est fortement sous-représentée, même si GCC est le codon le plus fréquent encodant pour l'alanine[36]. Récemment, Boël *et al.* ont utilisé une régression logistique prédisant l'expression des protéines chez *E.Coli* au moyen de la fréquence de codons dans les gènes qui les encodent. L'influence des codons qu'ils rapportent corrèle légèrement avec la fréquence des codons dans le génome, mais fortement avec la concentration des protéines, ainsi que la concentration et la demi-vie des mRNA *in vivo*[40]. Les auteurs rapportent que modifier les séquences des gènes en accord avec cette influence améliore l'efficacité de traduction *in vitro*[40]. Ces résultats, ainsi que la faible corrélation avec les fréquences des codons dans le génome, suggèrent que l'usage des codons possède une influence plus locale que globale.

1.2.4 Régulation de la stabilité du mRNA

L'usage des codons a aussi été démontré comme influant sur la stabilité des mRNA. Les séquences des mRNA stables sont enrichies en codons optimaux, alors que les séquences instables son principalement composées de codons non-optimaux[35]. Remplacer des codons optimaux par des synonymes non-optimaux diminue grandement la stabilité du mRNA, et inversement[35].

Étant donné que les MIP doivent être traités par le SUP, une source de MIP serait potentiellement les protéines ubiquitinées durant leur synthèse. Plusieurs facteurs peuvent augmenter la probabilité qu'une protéine soit ubiquitinée pendant sa traduction, un des plus

important étant le mauvais repliement de la chaine naissante de polypeptides. Les protéines les plus longues sont aussi les plus susceptibles d'être ubiquitinées durant leur traduction[29]. Les protéines courtes qui se retrouvent ubiquitinés durant leur traduction sont quant à elles enrichies en régions hydrophobes[29]. C'est d'ailleurs dans ces régions particulièrement susceptibles de former des agrégats que l'on retrouve un enrichissement en codons optimaux[37,39].

L'usage des codons influe donc fortement sur la traduction du mRNA en régulant notamment la stabilité du mRNA, le repliement des protéines et l'ubiquinitation des chaînes de polypeptides durant la traduction. Se pourrait-il donc que l'usage des codons régule également la genèse de l'immunopeptidome ?

1.3 L'analyse par Apprentissage Machine

Durant la dernière décennie, les réseaux de neurones artificiels (*Artificial Neural Networks*, ANN) ont connu un important regain d'intérêt grâce aux succès rencontrés par l'apprentissage profond[41,42]. Grâce à l'application de ces méthodes basées sur des ANN, un certain nombre de problèmes considérés comme traditionnellement difficiles en intelligence artificielle ont pu être résolus[41–43]. Ces avancements ne sont pas passés inaperçus dans le monde des sciences de la santé puisque les applications de réseaux de neurones aux problèmes de bio-informatiques se sont multipliées dans les dernières années[44–46]. L'utilisation de méthodes d'apprentissage machine en immunologie n'est pourtant pas chose nouvelle. L'exemple le plus connu d'application étant sans doute NetMHC[47], un ANN qui prédit l'affinité de séquences peptides aux molécules du MHC. Aujourd'hui, les principaux domaines de recherche et d'applications des réseaux de neurones sont les suivants :

1. Les problèmes de régression, où il s'agit de modéliser une fonction continue à partir de son ensemble de définition. Cette classe comprend la modélisation de structures de molécules et l'estimation de l'affinité de liaison entre molécules.

2. Les problèmes de classification, où il s'agit d'attribuer à des exemples une classe adéquate. Cette catégorie comprend les méthodes de diagnostic automatisé, comme par exemple la classification de tumeurs à partir d'images.
3. Les modèles génératifs, où il s'agit d'apprendre à générer des sorties similaires aux exemples de l'ensemble d'entrainement.
4. L'apprentissage par renforcement, où un agent apprend à effectuer des séries d'actions afin de maximiser sa récompense.

Au-delà de ces domaines, les ANN possèdent des caractéristiques fondamentales qui font d'eux des méthodes hautement efficaces pour l'analyse de gros jeux de données :

1. Ils sont capables de s'adapter aux ensembles de données et donc d'en extraire de l'information.
2. Cette adaptation est automatique et peut se faire indépendamment des connaissances *a priori* du chercheur.
3. Une fois qu'un ANN a été entrainé à effectuer une tâche, il est possible d'analyser et d'interpréter les structures internes du réseau afin d'en extraire des règles qui permettent à l'ANN de résoudre cette tâche. Bien que l'interprétation des ANNs soit encore à ses débuts, des techniques comme celles fondés sur l'analyse des poids du réseaux et l'ajout de mécanismes d'attentions, permettent d'identifier certaines des règles utilisées par un réseau entrainé[42].

Ce sont ces raisons qui nous ont poussés à utiliser des ANN pour découvrir le lien entre l'usage des codons et la présentation des MIP.

1.3.1 Introduction sur les réseaux de neurones artificiels

Un ANN est une fonction mathématique déterministe, déduite automatiquement à partir d'un ensemble de données, afin de répondre à une tâche précise. Dans un mode d'apprentissage supervisé, l'ensemble de données est composé d'entrées auxquelles ont été associées des valeurs cibles. Lorsqu'un ANN est utilisé pour répondre à une tâche, il est aussi appelé modèle, puisque

considéré comme une fonction modélisant un phénomène, dont en général seul le résultat est observable.

Le modèle obtenu suite à l'apprentissage dépend principalement de 4 facteurs :

1. L'architecture du réseau, autrement dit, la structure intrinsèque de la fonction de l'ANN.
2. La fonction de coût choisie pour l'apprentissage, qui mesure une divergence entre la sortie du réseau et la sortie attendue.
3. L'ensemble de données sur lequel est entrainé le réseau.
4. Les divers choix d'implémentation, tels que le choix d'algorithmes d'optimisation, ou l'addition de méthodes de régularisation.

1.3.2 Structure d'un ANN

Il existe plusieurs types d'ANN[48]. Par souci de concision, nous ne nous intéresserons qu'au type utilisé dans cette thèse. Ces réseaux sont appelés perceptrons à multiples couches (*Muli-Layer Perceptron*, MLP) pour des raisons historiques. *Perceptron* étant le premier ANN de ce type, qui à l'époque ne comportait que deux couches (une entrée et une sortie)[49].

Un MLP est, comme son nom l'indique, composé de multiples couches. Une couche d'un MLP est composée d'un ensemble de neurones non connectés entre eux. Les couches sont connectées et organisées hiérarchiquement (Figure 2). Elles sont en général *complètement connectées*, ce qui veux dire que chaque neurone d'une couche est connecté à tous les neurones de la couche précédente. Les connections qui lient les neurones sont appelées *poids*, ou *paramètres*. Ils sont initialisés aléatoirement et représentent les seules valeurs modifiées durant l'apprentissage.

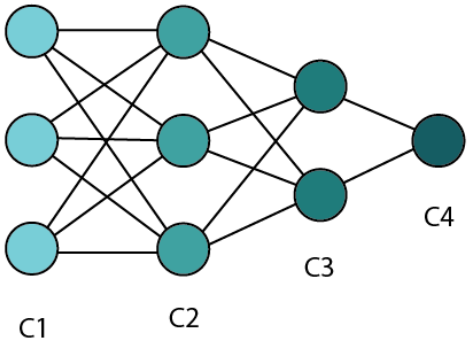


Figure 2. Architecture d'un MLP. Un MLP est composé de couches successives organisées hiérarchiquement. Ici, l'entrée est la couche C1, et la sortie la couche C4. Chaque neurone est connecté à l'ensemble des neurones de la couche précédentes par des poids (lignes noires). L'apprentissage consiste en un ajustement graduel de ces poids. Chaque couche intègre l'information de la couche précédente, qui elle-même intègre celle de la suivante. Ceci permet au réseau de développer de représentations plus en plus abstraites de l'entrée.

Le rôle d'un neurone dans un MLP est d'intégrer l'ensemble des valeurs des neurones de la couche précédente en une seule valeur appelée *activation*. Plus concrètement un neurone effectue une somme pondérée des valeurs des neurones de la couche précédente. Son activation est calculée selon la formule :

$$a_i = f(\sum_j w_{ij} a_i + b_i)$$

Ou a_i est l'activation du neurone i, w_{ij} est le poids entre le neurone i et le neurone j de la couche précédente, b_i est un paramètre supplémentaire d'ajustement appelé *biais*, f est une fonction de *non-linéarité*. Le rôle de f est de permettre au neurone de retourner une valeur qui n'est pas nécessairement proportionnelle à la valeur qu'il reçoit en entrée, ce qui confère au réseau une plus grande adaptabilité.

Dans un MLP, chaque couche intègre donc l'information de la couche précédente, qui elle-même intègre l'information de la couche qui la précède, etc… Cette organisation hiérarchique permet aux ANN de développer une représentation interne de plus en plus abstraite de l'entrée, au fur et à mesure que l'information remonte vers la sortie (Figure 2)[48].

1.3.3 Architecture

L'architecture désigne le nombre de couches d'un MLP ainsi que le nombre de neurones qui constitue une couche. L'architecture d'un MLP est fixe et n'évolue donc pas durant l'apprentissage. Pour cette raison, l'architecture donnée au réseau conditionne sa capacité à modéliser le phénomène en question. Si le réseau ne possède qu'une couche d'entrée et une couche de sortie, il est alors incapable de modéliser tout phénomène qui demanderait plus de complexité, et donc des couches intermédiaires.

1.3.4 Comment est entrainé un ANN

Les réseaux de neurones artificiels sont en général entrainés à l'aide d'un algorithme d'optimisation appelé descente de gradient stochastique (*Stochastique Gradient Descent*, SGD). Il existe plusieurs versions de la SGD, certaines utilisant un momentum[48], d'autres adaptatives[48], d'autres encore combinant adaptation et momentum[48]. Ici, nous ne nous intéresserons qu'à la version classique de la SGD. D'une part, parce qu'indépendamment de l'algorithme utilisé, les principes fondamentaux restent ceux de la SGD classique. D'autre part, parce que seulement la version classique a été utilisée pour produire les résultats de cette thèse.

La SGD est une méthode d'optimisation itérative de premier ordre. En d'autres termes, c'est un algorithme dont le but est de trouver le minimum d'une fonction en appliquant une succession d'étapes basées sur une dérivée première. Dans le cas d'un ANN, la fonction à minimiser est la divergence entre la sortie du réseau et la sortie cible attendue. Plus la valeur de cette divergence est petite et plus la sortie de l'ANN est proche de la cible. Le côté stochastique (« *aléatoire* ») de la méthode vient du fait que les optimisations ne sont calculées qu'à partir d'un sous-ensemble aléatoire d'exemples d'entrainement à la fois. Afin de simplifier les explications, nous considérerons que le réseau ne reçoit qu'un exemple à la fois. Les principes restent identiques indépendamment du nombre d'exemples présentés. En pratique toutefois, ce nombre est arbitraire et défini par l'expérimentateur.

L'entraînement d'un ANN par SGD se fait en deux étapes :

1. Propagation avant : le réseau reçoit des valeurs d'entrée issues de l'ensemble d'entrainement et sa sortie *Y* est récupérée.
2. Propagation arrière ou *rétro-propagation du gradient* : la dérivée de la divergence entre *Y* et la sortie cible *Z* par rapport à chaque paramètre du réseau est calculée. Chaque paramètre est ensuite ajusté de façon inversement proportionnelle à cette dérivée. En d'autres termes, chaque paramètre est ajusté proportionnellement à sa « responsabilité » dans l'erreur.

La propagation est dite arrière puisque l'on calcule d'abord la dérivée par rapport aux paramètres de la dernière couche, puis par rapport à ceux de la suivante, etc… en utilisant la règle des dérivées en chaines[42,49]. Selon cette règle, les dérivées des paramètres par rapport à la fonction de coût sont entièrement dépendantes des dérivées par rapport aux paramètres de la couche précédente. De la même façon que pour calculer la sortie du réseau, il est nécessaire de calculer les sorties des couches dans l'ordre (la sortie de chaque couche dépendant de la sortie de la précédente). Calculer les ajustements à appliquer aux paramètres nécessite de calculer ces modifications dans l'ordre inverse.

Ceci nous amène à une importante limitation de la SGD. Étant donné que les modifications appliquées par la SGD sont entièrement dépendantes de la dérivée première de la fonction de coût, cela implique que si cette dérivée est nulle (i.e. égale à zéro), aucune modification des paramètres n'est effectuée. La SGD ne peut donc différencier entre une sortie juste et une sortie saturée (ex : proche de 1, quelle que soit la valeur d'entrée). Or, d'après la règle des dérivés en chaine, si la dérivée pour les paramètres d'une couche est proche de 0, ceci implique également une dérivée proche de 0 pour les paramètres des couches précédentes. Par conséquent, plus les neurones qui saturent sont proches de la sortie et plus l'effet de la saturation se fait sentir dans l'ensemble du réseau. Ce phénomène est communément appelé la *disparition du gradient* (*vanishing gradient*)[50]. Afin d'éviter de faire face à ce problème, il est conseillé d'éviter les fonctions d'activations saturantes telles que *tanh* et *sigmoid* pour les réseaux profonds, et de les remplacer par des fonctions d'activation non saturantes telles que la *ReLU*[42,48].

1.3.5 Mesurer les performances d'un ANN

Comment s'assurer des performances d'un ANN entrainé? Un ANN est capable d'apprendre par cœur un ensemble d'entrainement, mais les performances sur cet ensemble ne nous renseignent pas sur les capacités de généralisation du réseau. En effet, 100% de réponses justes sur l'ensemble d'entrainement peut être dû au fait que le réseau a appris des caractéristiques propres à l'ensemble d'entrainement qui ne se transposent pas en dehors de celui-ci. Les performances d'un ANN sont donc rapportées sur un autre ensemble, appelé ensemble de test, qui contient exclusivement des exemples qui ne sont pas présents dans l'ensemble d'entrainement.

1.4 Introduction des articles

Le premier chapitre de cette thèse introduit les méthodes de protéogénomique qui nous ont permis d'acquérir notre ensemble de données regroupant la fois MIP et informations transcriptionnelles.

Les seules méthodes fiables permettant l'identification de MIP à haut débit dépendent de la Spectrométrie de Masse (MS)[17]. Ces méthodes calculent des spectres théoriques à partir d'une base de données, et comparent ensuite ces spectres à ceux observés durant le séquençage afin de déterminer les séquences des MIP[17,51]. Afin d'améliorer la précision de l'identification, il est donc primordial d'améliorer la précision de l'identification par MS. Le second chapitre de cette thèse présente l'article qui a introduit l'utilisation de génomes personnalisés pour l'identification de MIP. Cette méthode est à l'origine de l'ensemble de données utilisé dans le chapitre 4, ainsi que de celui utilisé dans l'étude de Pearson *et al*.[20] (voir Annexe 1). Elle

remplace le protéome de référence traditionnellement utilisé pour les séquençages par MS par un protéome spécifique construit à partir du génome du sujet. Cette amélioration nous a permis d'obtenir des identifications plus précises et d'identifier 34 nouveaux antigènes mineurs (MiHA).

Le troisième chapitre présente pyGeno, une librairie python développée par l'auteur de cette thèse. PyGeno est une base donnée génomique et protéique qui permet d'intégrer des génomes de référence issus des séquences et annotations fournies par Ensembl[52], ainsi que des ensembles de polymorphismes arbitraires. De cette façon, pyGeno permet la création simple et rapide de génomes et protéomes personnalisés.

Le quatrième chapitre présente notre travail sur l'influence de la traduction sur la présentation des MIP. En utilisant diverses méthodes statistiques, dont des ANN, sur un ensemble de données contenant à la fois les MIP et les transcrits issus de 18 sujets sains, nous démontrons que l'usage des codons influence la présentation des MIP.

En annexe se trouve l'article de Pearson et al.[20] introduit dans les sections 1.1.4 et 1.1.5, ainsi que l'article de Laumont et al.[5] qui introduit une méthode de proteogénomique pour l'identification de MIP provenant de région traditionnellement considérées comme non-codantes. Cette méthode évite l'étape d'alignement de séquences en créant une base de données de MS construite directement à partir des résultats du séquenceurs. Cet article est présenté conjointement à ceux de cette thèse car il présente une méthode d'identifications qui permettrait d'étendre les résultats cette thèse aux séquences traditionnellement considérés comme non codantes.

Chapitre 2

Impact of Genomic Polymorphisms on the Repertoire of Human MHC Class I-Associated Peptides

2.1 Résumé

Pour plusieurs décennies, l'impact global des polymorphismes génomiques sur le répertoire des peptides présentés par les complexes majeurs d'histocompatilibité (MHC) est resté l'objet de spéculations. Nous avons développé une nouvelle approche permettant l'identification haut débit de peptides polymorphiques associés aux molécules du MHC de classe I (MIP), qui jouent un rôle majeur dans la reconnaissance allogénique. Par l'analyse globale des MIP élués de lymphoblastes B provenant de deux sœurs MHC-identiques, nous avons montré que seulement 0.5% des variations nucléotidiques non-synonymes sont représentées dans le répertoire des MIP. Les 34 MIP polymorphiques identifiés chez nos sujets sont encodés par les loci bi-alléliques possédant des allèles dominants et récessifs. Nos analyses montrent qu'au niveau de la population, 12% de l'exome encodant des MIP est polymorphique. Notre méthode démontre une relation fondamentale entre le soi génomique et le soi immunitaire et accélère la découverte de MIP polymorphiques (aussi appelé antigènes mineurs d'histocompatibilité). En introduisant l'utilisation de génomes personnalisé, ce travail a mis en évidence la nécessité d'avoir une base de données précise pour l'identification de MIP par

spectrométrie de masse. Ceci lui a permis d'avoir une influence importante sur le développement des technologies de spectrométrie de masse pour l'identification des MIP[5,51,53].

2.2 Contributions des auteurs

D.P.G : Conception de l'étude. Préparation des figures. Rédaction de la première ébauche du manuscrit. Analyse de données et réalisation des expériences.

D.S : Conception de l'étude. Préparation des figures. Rédaction de la première ébauche du manuscrit. Réalisation des expériences de spectrométrie de masse. Analyse de données et réalisation des expériences (Spectrométrie de masse).

T.D : Conception de l'étude. Création des génomes personnalisés. Identification des sources génomiques des MiHA. Développement de pyGeno. Contribué à la rédaction et à la préparation des figures 1 et 6. Analyse de données et réalisation des expériences (bio-informatique).

O.C.L et A.Z : Développement d'outils bio-informatiques pour l'analyse de données de spectrométrie de masse. Préparation de figures.

C.L. : Effectué des analyses et préparation d'une figure.

C.C. et M.P.H. : Effectué des expériences.

G.B. : Préparation des figures de Circos et les analyses bio-informatiques.

P.G. : Séquençage et alignement des séquences génomiques et transcriptomiques.

S.L. : Conception de l'étude, discussion à propos des analyses statistiques et des résultats.

P.T. et C.P. : Conception de l'étude, analyse de données, discussion à propos des résultats. Rédaction du manuscrit et contribution également en tant qu'auteurs principaux.

2.3 Référence de publication

Granados, D. P., Sriranganadane, D., Daouda, T., *et al*. (2014). Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nature communications*, 5.

2.4 Article

Impact of Genomic Polymorphisms on the Repertoire of Human MHC Class I-Associated Peptides

Diana Paola Granados[1,2,†], Dev Sriranganadane[1,3,†], Tariq Daouda[1,4,†], Antoine Zieger[1,4], Céline M. Laumont[1,2], Olivier Caron-Lizotte[1], Geneviève Boucher[1], Marie-Pierre Hardy[1], Patrick Gendron[1], Caroline Côté[1], Sébastien Lemieux[1,4], Pierre Thibault[1,3,*] & Claude Perreault[1,2,*]

[†] These authors contributed equally to this work

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

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

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

[4]Department of Informatics and Operational Research, Université de Montréal, P.O. Box 6128, Downtown Station, Montreal, QC, Canada H3C 3J7

* Corresponding authors:

Pierre Thibault (pierre.thibault@umontreal.ca) and

Claude Perreault (claude.perreault@umontreal.ca)

2.5 Abstract

For decades, the global impact of genomic polymorphisms on the repertoire of peptides presented by major histocompatibility complex (MHC) has remained a matter of speculation. Here we present a novel approach that enables high-throughput discovery of polymorphic MHC class I-associated peptides (MIPs), which play a major role in allorecognition. Based on comprehensive analyses of the genomic landscape of MIPs eluted from B-lymphoblasts of two MHC-identical siblings, we show that 0.5% of non-synonymous single nucleotide variations are represented in the MIP repertoire. The 34 polymorphic MIPs found in our subjects are encoded by biallelic loci with dominant and recessive alleles. Our analyses show that, at the population level, 12% of the MIP-coding exome is polymorphic. Our method provides fundamental insights into the relation between the genomic self and the immune self and accelerates the discovery of polymorphic MIPs (also known as minor histocompatibility antigens).

2.6 Introduction

Classic adaptive CD8 T cells recognize MHC class I-associated peptides (MIPs), and the ensemble of MIPs presented on the surface of a cell (the "immunopeptidome") establishes its immunologic identity[1-3]. CD8 T cells are eminently self-referential and highly discriminant: they are selected on self-MIPs, sustained by self-MIPs, and must swiftly react when confronted with nonself MIPs interspersed in a sea of self-MIPs[4, 5]. Understanding the molecular definition of self for CD8 T cells has been made possible by high-throughput mass spectrometry (MS) analyses of MIPs[6-12]. Progress in this field has been heralded by the development of MS instruments whose sensitivity, dynamic range and mass accuracy are orders of magnitude superior to those of analyzers available a decade ago[13]. High-throughput MS studies have revealed that the immunopeptidome is highly complex and that its composition (i.e., the source of MIPs) cannot be inferred solely from transcript or protein abundance[7, 9, 12, 14-16].

The MHC I region contains two major classes of genes: modern classical MHC Ia genes (e.g., *HLA-A*, *HLA-B* and *HLA-C* in humans) and more ancient MHC Ib genes (e.g., *HLA-E* and *HLA-G*). MHC Ia molecules play a dominant role in adaptive immunity. They bind MIPs and are encoded by the most polymorphic genes known[17, 18]. Since MHC Ia allotypes display distinct peptide binding motifs, the HLA genotype has a major impact on the MIP repertoire[19]. Notably, almost all genetic polymorphisms in HLA Ia alleles are located in exons 2 and 3, which encode the MIP-binding pocket. Besides, the 1000 Genomes Project Consortium has identified 38 million single nucleotide polymorphisms (SNP), 1.4 million short insertions and deletions, after comprehensive studies on 1,092 subjects[18]. This raises the fundamental question: what might be the impact of the numerous polymorphisms outside of the MHC on the MIP repertoire?

In other words, to what extent do genomic polymorphisms translate into differences in the immunopeptidome?

Several MIPs have been found to derive from polymorphic genomic regions[20, 21]. For historical reasons, these polymorphic MIPs are referred to as minor histocompatibility antigens (MiHAs). MiHAs are essentially genetic polymorphisms viewed from a T-cell perspective. MiHA-coding alleles can be dominant or recessive at the peptide level. Thus, a non-synonymous single nucleotide polymorphism (ns-SNP) in a MIP-coding genomic sequence will either hinder MIP generation (recessive allele) or generate a variant MIP (dominant allele)[22-24]. MiHAs are generally defined according to three criteria: they are present in some but not in all subjects bearing a given HLA allele, their presence/absence is linked to a well-defined genetic polymorphism, and they can elicit allo-immune T-cell responses[22-24]. Three decades of research have led to the discovery of about 35 human MiHAs encoded by autosomes and presented by HLA class I molecules[23, 24]. The discovery of each MiHA has been a major endeavor, if not a technical tour de force[25-30]. However, due to the lack of a suitable systems level approach, we ignore the global impact of non-MHC genomic polymorphisms on the immunopeptidome (i.e., what proportion of MIPs are MiHAs). Based on various theoretical premises, it has been speculated that the number of MiHAs expressed by an individual might be very low (less than 10) or very high (greater than 1,000)[21, 24]. In addition to its conceptual importance, the impact of genetic polymorphisms on the immunopeptidome is of considerable medical relevance because MiHAs are the targets of three allo-immune processes: graft rejection, graft-versus-host disease and graft-versus-tumor reaction[24, 31-36].

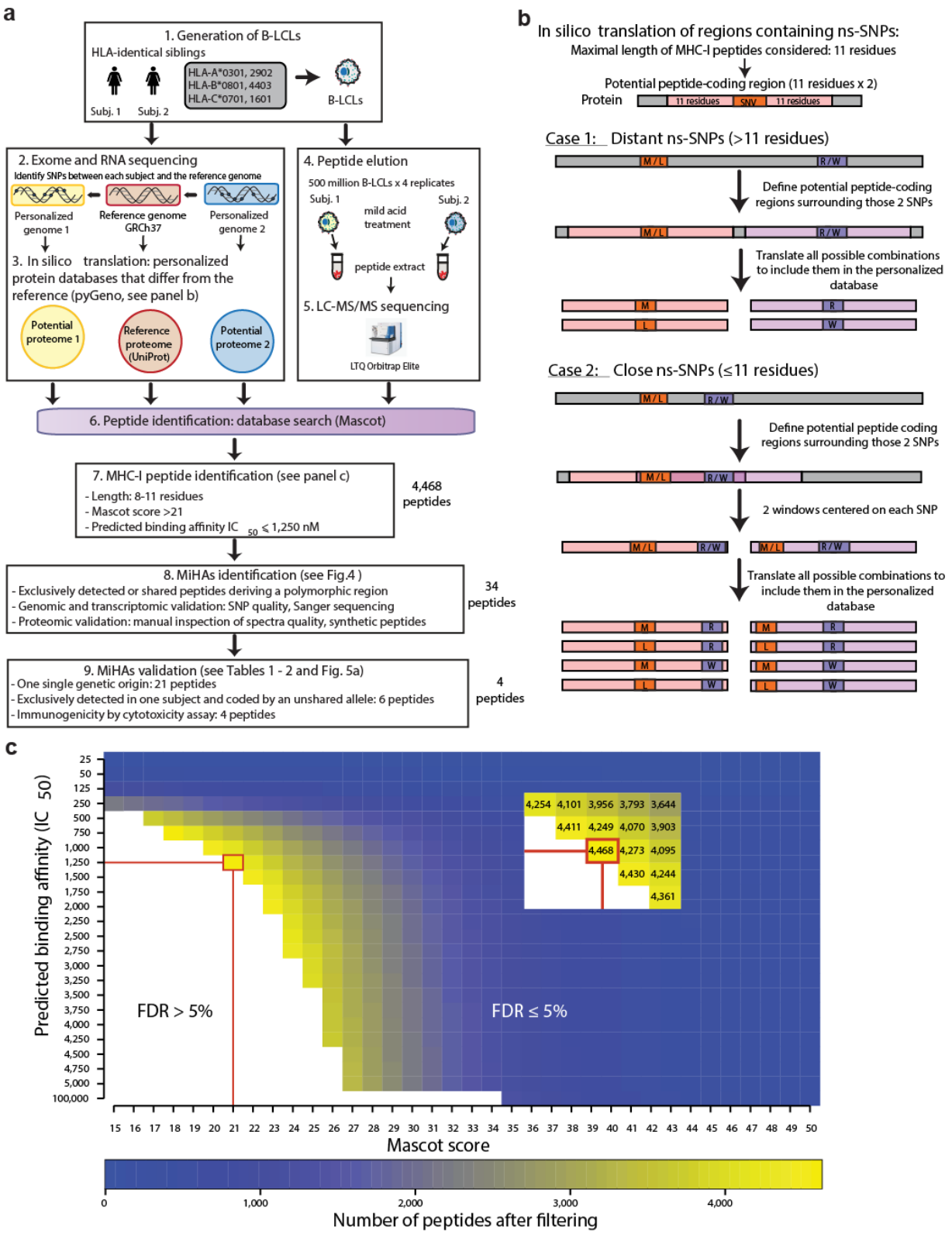
Systems-level molecular definition of the immunopeptidome can be achieved only by MS studies. However, since current MS approaches cannot reliably detect polymorphic peptides, they are inadequate for MiHA discovery[37]. Furthermore, since several steps of MIP processing cannot be modeled with available algorithms[38], MiHA identification using prediction tools is a daunting task fraught with high false discovery rates[37]. To resolve this conundrum, we have developed a genoproteomic strategy that hinges on a combination of next-generation sequencing and high-throughput MS peptide identification. Our personalized platform provides unprecedented insights into the genomic landscape of human MIPs and enables high-throughput identification of MiHAs and of their underlying genomic polymorphisms.

2.7 Results

2.7.1 Novel approach for the identification of MIPs

To evaluate the impact of non-HLA genetic polymorphisms on the MIP repertoire, we analyzed the immunopeptidome of Epstein-Barr virus (EBV)-transformed B cell lines (B-LCLs) from two non-twin HLA-identical female siblings (Fig. 1a). The success of our endeavor hinged on two factors: the need to reliably identify MIPs encoded by polymorphic genomic regions and to maximize the coverage of the immunopeptidome (the number of unique MIPs identified).

Detection of MIPs encoded by polymorphic genomic sequences using personalized proteomic databases. Large-scale MS-based analyses represent the sole approach enabling comprehensive molecular definition of the MIP repertoire[1, 12, 39]. However, standard high-throughput MS is blind to a whole universe of polymorphic peptides. Indeed, sequencing (or assignation) of peptides by tandem MS is done using engines (e.g., Mascot) that attempt to correlate tandem MS fragment ions from a sample under study with those predicted from available protein databases (e.g. UniProt). Unfortunately, most polymorphic peptides are absent from these databases and tandem MS spectra from unlisted polymorphic peptides will inevitably remain unassigned or misassigned. We reasoned that the most straightforward solution to this conundrum would be to use next generation sequencing data to create subject-specific proteomic databases that would serve as a reference for MS sequencing.



(Figure legend on next page)

Figure 1. High-throughput genoproteomic strategy used for the identification of polymorphic MIPs on B-LCLs from 2 HLA-identical siblings. (a) General overview of the personalized approach, which combines next-generation sequencing, MS and bioinformatics. (b) Schematic representation of the combinatorial method used to translate *in silico* polymorphic regions containing ns-SNPs. (c) Combining the predicted MHC binding affinity and Mascot score enables to discriminate between MIPs and contaminant peptides. The dataset of peptides identified with an FDR ≤ 5% was filtered according the Mascot score (which represents the confidence level of a peptide assignation), and the predicted MHC binding affinity. The red rectangle and lines indicate the combination of values (IC_{50} ≤ 1,250 nM and Mascot score ≥ 21) that allowed identifying the maximum number of MIPs with a 5% FDR threshold.

Transcriptome sequencing (or RNA-seq) provides information about gene expression and can reveal sequence variation such as SNPs or RNA editing events[40]. However, lowly expressed genes might be missed by RNA-seq depending on the depth of coverage. Exome sequencing is the method of choice to capture RNA coding or exonic regions including SNPs as it tends to be less noisy than RNA-seq for variant calling and mapping[40]. Nevertheless, exome capture is limited to regions that are targeted by the probe set and not all exons are indeed transcribed in a particular cell type. Accordingly, the immunopeptidome is cell type-specific[1] and preferentially derives from abundant transcripts[8, 19], and hence it is more likely to reflect transcriptome sequences rather than genomic sequences. To combine the benefits of both sequencing technologies and to cover as much as possible each individual's coding genome[40], we sequenced both the exome and the transcriptome of B-LCLs from each subject (Fig. 1a). Annotated exons were covered on average at a depth of 130-131x in the RNA-seq and a coverage depth of 66-158x of exonic targets was achieved in the exome capture, with 98% of targets covered at a minimum depth of five reads (Supplementary Data 1**)**. In total, more than

53 and 50 mega base pairs (Mb) of annotated exons were covered in subjects 1 and 2, representing 76-81% of the human annotated exome **(**Supplementary Data 1**)**.

Next-generation sequencing data were used to build *in silico* the proteome of B-LCLs from our subjects using the in-house developed python module pyGeno[19] (Fig. 1b). Following integration of exome and transcriptome sequencing, similar number of base pairs and proportions of the human exome were covered in both siblings **(**Fig. 2 track 3 blue vs. orange and Supplementary Data 1**)**. Exome and transcriptome sequencing data of each subject were used to identify SNPs with respect to the reference genome (GRCh37.p2, NCBI), which were then filtered according to their quality (see Methods). The majority (93.2 – 97.7%) of the identified SNPs are reported in the dbSNP database[41] (Supplementary Data 2). SNPs were combined into a single set and integrated at their respective position on the reference human genome to obtain two "personalized genomes", from which we extracted and translated every transcript (see Methods section). The translations were then compiled in two "personalized protein databases", one for each subject.

Maximizing the level of coverage of the immunopeptidome. MIPs were eluted from the cell surface by mild acid elution performed on four biological replicates of 500 million cells for each subject. Eluted peptides were desalted and separated on strong cation exchange chromatography prior to LC-MS/MS analyses using high resolution precursor and product ion spectra. Compared to other methods such as MHC I immunoprecipitation, acid elution has the advantage of harvesting almost all MIPs, irrespective of their MHC binding affinity[42]. However,

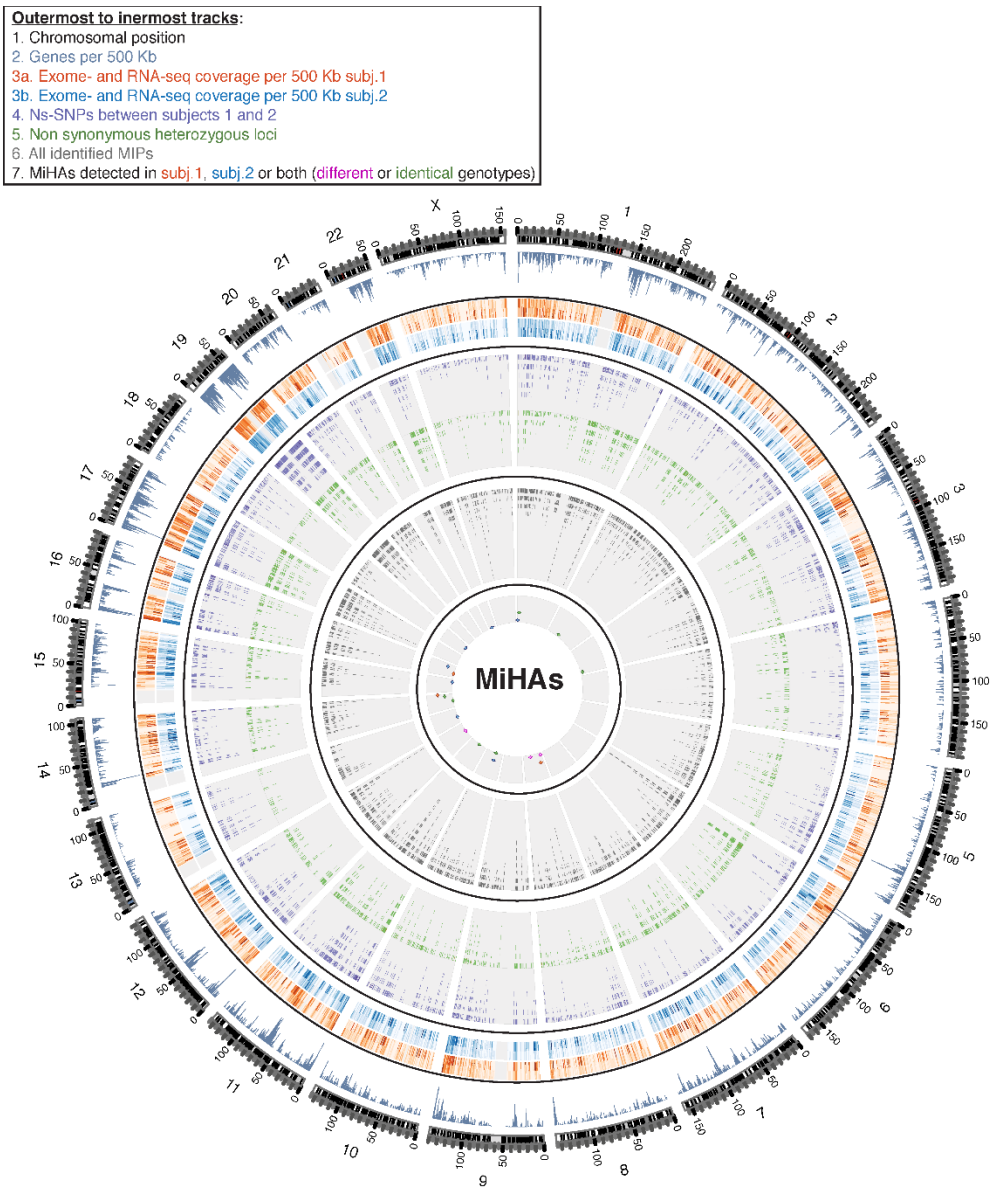


Figure 2. Integrative view of the genomic landscape of the MIP repertoire of HLA-identical siblings. Circos plot showing similar proportions of sequenced genomic and transcriptomic regions in both siblings (tracks 1-3) and the small number of identified MiHAs (track 7) relative to the number of MS-detected MIPs (track 6) and sequenced polymorphic regions (tracks 4-5). From outermost to innermost tracks: 1) ideogram indicating chromosomal positions for each chromosome, 2) histogram depicting the number of genes for 500-Kb windows, 3) heat map showing the fraction of bases of 500-Kb windows covered by exome (outer circle) or transcriptome (inner circle) sequencing of subjects 1 (orange) and 2 (blue), 4) tile graph of 4,833 ns-SNP between siblings (purple), 5) tile graph of 3,774 heterozygous loci where both alleles are shared by the two subjects and lead to non-synonymous amino

acid changes (green), 6) tile graph representing genomic regions that give rise to 4,468 MIPs, 7) Each dot represents one single gene-encoded MiHA deriving from regions containing ns-SNPs and detected by MS in subjects 1 (orange), 2 (blue) or both (green).

direct acid elution can increase the amount of non-MHC contaminant peptides that are recovered[8]. In order to maximize the sensitivity and specificity of MIP detection, we have therefore developed an analysis pipeline that relies on a combination of four parameters: i) the canonical MIP length of 8 to 11 amino acids, (ii) the predicted MHC binding affinity given by the NetMHCcons algorithm[43], (iii) the Mascot score, which reflects the quality of peptide assignation, and (iv) the false discovery rate (FDR), which indicates the proportion of decoy (false) vs. target (true) identifications (see Methods section). We found that for an FDR of 5%, the best coverage of the immunopeptidome was obtained by combining a Mascot score ≥ 21 and an MHC binding affinity $\leq 1{,}250$ nM (Fig. 1c and Supplementary Fig. 1-2).

Next, we compared the number of peptide identifications obtained by Mascot using the regular human protein database (UniProt) and personalized databases based on exome and transcriptome sequencing (Supplementary Fig. 3a). We identified 4,468 unique MIPs from the two personalized databases (Supplementary Data 3). The numbers of MIPs identified with the reference database vs. personalized databases were similar with a 96% overlap (Supplementary Fig. 3a). Notably, replacement of reference with the personalized databases had no impact on the quality (Mascot score) of identified peptides (Supplementary Fig. 3b).

2.7.2 The MIP repertoire of HLA-identical siblings

We have previously shown that the HLA genotype has a major impact on the MIP repertoire of MHC-mismatched individuals[19]. Here, we compared the MIP repertoire of HLA-identical siblings to evaluate the impact of non-HLA genetic polymorphisms on the immunopeptidome. In addition to having identical HLA genotypes, the two siblings showed similar expression levels of the *HLA-A*, *HLA-B* and *HLA-C* genes (Supplementary Data 4) and of the total amount of MHC class I molecules at the cell surface (Supplementary Fig. 4). Following mild acid elution of peptides of comparable efficacy between subjects (Supplementary Fig. 4), we identified a total of 4,468 MIPs encoded by genes from all chromosomes (Fig. 2 track 6 and Supplementary Data 3), detected in a variable number of biological replicates (Fig. 3a) and associated to HLA-A*03:01, -A*29:02, -B*08:01, -B*44:03 or –C*16:01. Similar numbers of MIPs were identified from the two subjects (4,114 in subject 1 and 4,186 in subject 2). As expected, the majority of the MIPs (86%) were detected in both subjects (Fig. 3a). Most MIPs (75%) had a predicted binding affinity < 500 nM (Fig. 3b). We found no significant difference in the average binding affinity of 282 peptides exclusively detected in subject 1 vs. 351 peptides exclusively detected in subject 2 (Fig. 3b). Furthermore, the number of peptides predicted to bind each of the HLA molecules was similar between the 2 subjects, suggesting that both siblings had comparable surface expression of each of the 5 HLA allelic products tested (Fig. 3c). Collectively, these results show that the MIP repertoire of HLA-identical subjects is similar yet not identical.

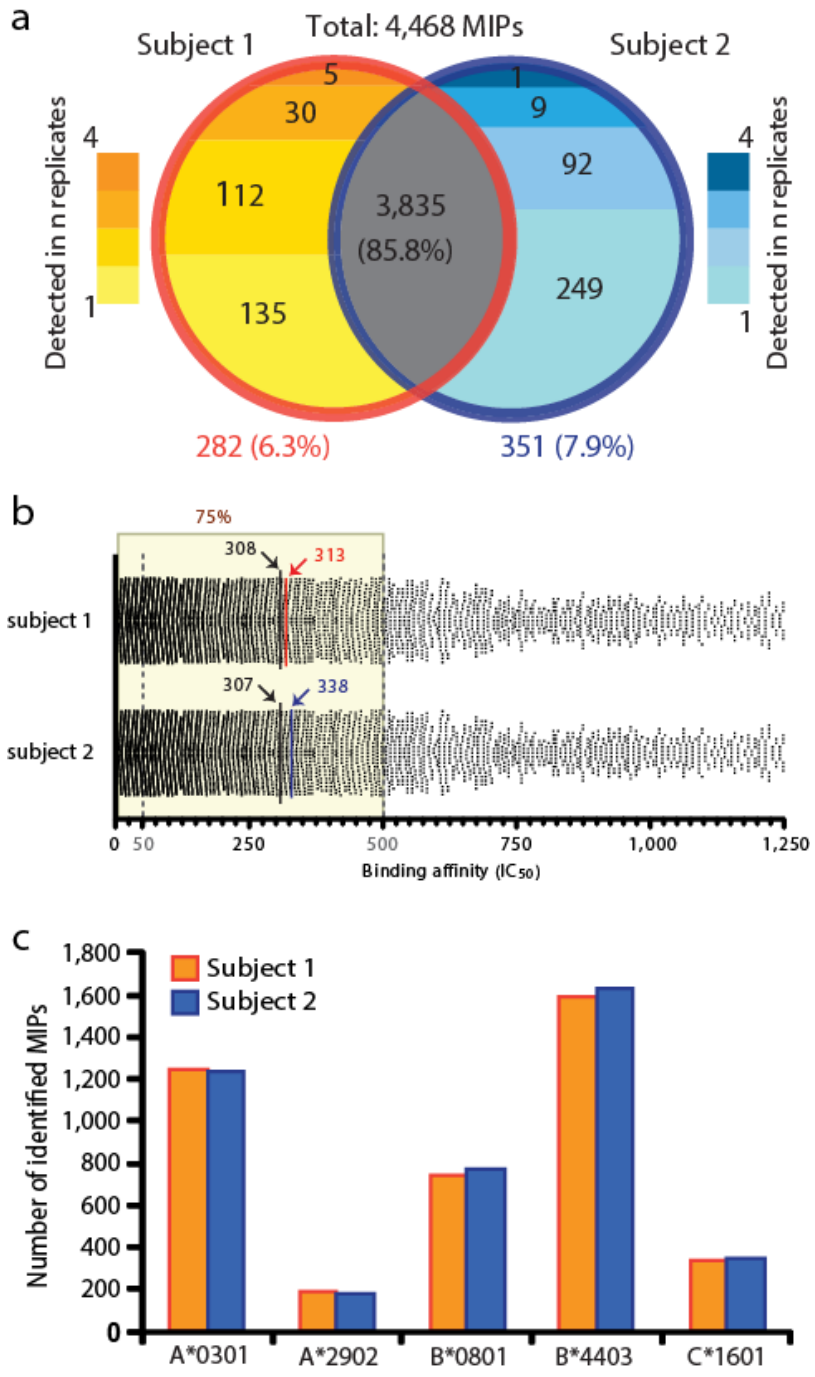


Figure 3. HLA-identical siblings present similar but not identical MIP repertoires. (**a**) Venn diagram showing that 86% of MIPs from HLA-identical siblings were detected in both subjects. A total of 4,468 MIPs were identified in the siblings after analysis of 8 biological samples (4 biological replicates per sibling). MIPs were detected in variable number of biological replicates. For peptides exclusively detected in one subject, the number of replicates in which the peptide was found is shown. The total

numbers of MIPs exclusively detected in subject 1 or 2 are shown in red and blue, respectively. (**b**) Scatter plot showing that 75% of identified MIPs are predicted to bind their respective HLA molecules with an IC_{50} < 500 nM. The IC_{50} for 5 HLA alleles was calculated with the NetMHCcons algorithm. For each peptide (represented by dots), the best binding score for a specific allele was kept. The yellow box highlights 75% of all peptides. The black lines and values indicate the average binding affinity of all peptides identified in each sibling. Red and blue lines and numbers represent the average binding affinity of 282 and 351 unshared peptides exclusively detected in subject 1 or 2, respectively. The predicted binding affinity of the two sets of unshared MIPs was statistically indistinguishable ($P = 8.5 \times 10^6$ by 2-tailed Mann-Whitney test). (**c**) The number of peptides associated to each HLA molecule was similar between the 2 subjects.

2.7.3 MiHAs among MIPs detected exclusively in one subject

MiHAs are typically encoded by bi-allelic loci[22, 23]. For each locus where two alleles are present in our subjects, three genotypes are possible: AA, AB and BB. At the peptidomic level, each allele can be dominant (generate a MIP) or recessive (a null allele that generates no MIP). Moreover, by comparing MIPs eluted from two HLA-identical individuals, dominant MiHAs can be separated into two groups based on their MS detection: shared MIPs and MIPs detected exclusively in one subject. MIPs detected in only one subject derive from different genotypes (e.g. AA vs. <u>BB</u> and AA vs. A<u>B</u> if only <u>B</u> is a dominant allele), while shared MIPs can originate from identical genotypes (AB vs. AB) or from different genotypes (e.g. <u>BB</u> vs. A<u>B</u> if only B is a dominant allele). Thus, subjects can be similar at the peptidomic level (display shared MIPs) even though they have different genotypes (Fig. 4).

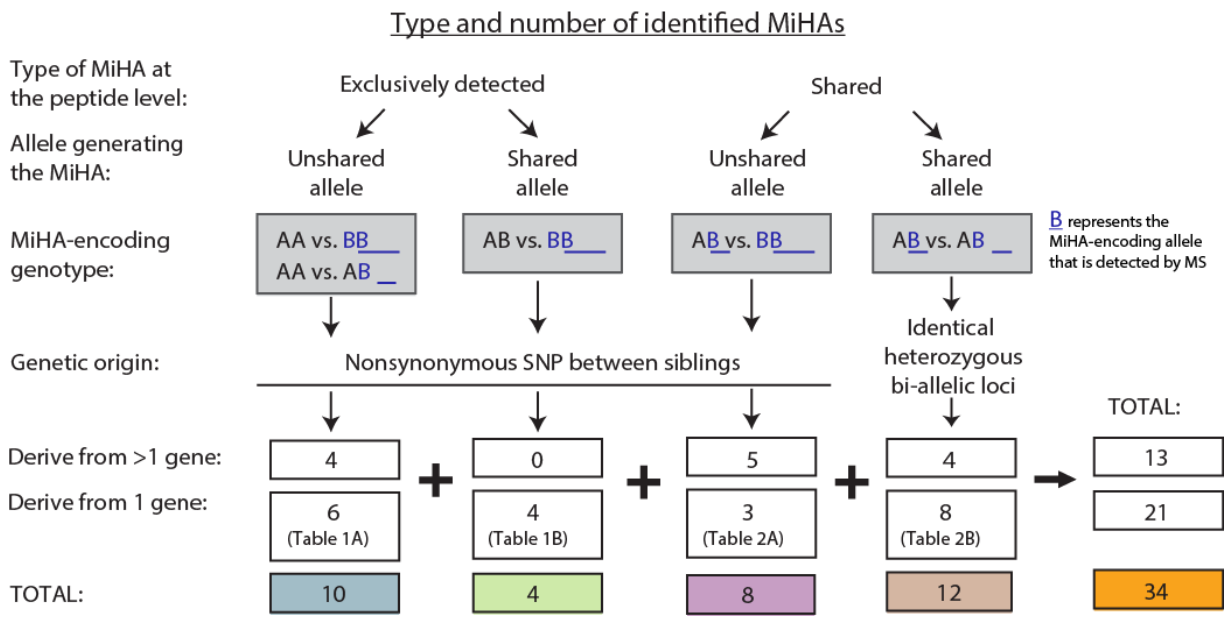


Figure 4. Overview of MiHAs identified following analysis of genomic and peptidomic data from our two subjects. See Tables 1 and 2 for more details.

In our search for MiHAs, we first performed in-depth analyses of MIPs detected in only one subject (Fig. 3a). Here the key finding was that out of 633 MIPs exclusively detected in one subject, only 14 (2%) were encoded by genomic regions harboring ns-SNPs between the two subjects (Fig. 4, n=10+4 and Supplementary Data 3). The origin of 4 of these 14 MIPs was ambiguous (they could derive from several genes), whereas the other 10 MiHAs were assigned to a single gene (Fig. 4, n=6+4 and Table 1). The genetic polymorphisms responsible for almost all MiHAs corresponded to reported SNPs (Table 1). Consistent with previous findings on human MiHAs[23], only one of the two possible variants was detected by MS for each MiHA locus (Table 1). In other words, at the peptide level, one allele was dominant (generated a MIP) and one was recessive (generated no MIP) (Table 1). In five out of 10 cases, absence of the variant MiHA at the cell surface could be explained by a decreased binding affinity of the variant

Table 1. MiHAs detected by MS in only one of the two subjects and resulting from ns-SNPs in MIP-coding regions. All MiHAs have one single genetic origin and are coded by an unshared (A) or shared allele (B) between subjects.

	MiHA name	Detected MiHA sequence	S	Gene symbol	HLA allele	IC_{50} (nM)	aa sub. 1	aa sub.2	Alternative MiHA variant	IC_{50} nM	IC_{50} ratio	dbSNP
A	ITGAL-1^{T}*	S**T**ALRLTAF	1	*ITGAL*	C*16:01	306	TR	RR	S**R**ALRLTAF	*2,969*	*9.7*	rs2230433
	IGHV5-51-1^{V}	**V**IYPGDSDTRY	1	*IGHV5-51*	A*29:02	27	VI	SS	**I/S**IYPGDSDTRY	19/31	0.7/1.1	rs199610746
	NQO1-1^{R}*	AMYDKGPF**R**SK	2	*NQO1*	A*03:01	12	WW	RW	AMYDKGPF**W**SK	11	0.9	rs1131341
	GRP-1^{R}	**R**ELPLVLL	2	*GRP*	B*44:03	285	SS	RR	**S**ELPLVLL	159	0.6	rs1062557
	C13orf18-1^{R}*	RVSLPTSP**R**	2	*C13orf18*	A*03:01	235	GG	RR	RVSLPTSP**G**	*11,858*	*50.5*	rs1408184
	IGLV2-11-1^{HH}*	SDVGG**HH**Y	2	*IGLV2-11*	A*29:02	660	YY-NN	HH-HH	SDVGG**YN**Y	412	0.6	---
B	R3HCC1-1^{H}	AENDFV**H**RI	1	*R3HCC1*	B*44:03	61	HH	HR	AENDFV**R**RI	*123*	*2*	rs11546682
	NADK-1^{K}	AVHNGLGE**K**	2	*NADK*	A*03:01	229	KN	KK	AVHNGLGE**N**	*24,349*	*106.3*	rs4751
	ACC-2^{G}	KEFED**G**IINW	2	*BCL2A1*	B*44:03	49	GD	GG	KEFED**D**IINW	59	1.2	rs3826007
	KIF20B-1^{I}	QELETS**I**KKI	2	*KIF20B*	B*44:03	288	IN	II	QELETS**N**KKI	*615*	*2.1*	rs12572012

Selected features of the MiHAs are shown: the detected amino acid sequence (polymorphic residues are highlighted in bold underlined), the subject (S) in which the MiHA was detected, the source gene, the HLA molecule for which the MiHA has the best predicted binding affinity (IC_{50}), the translated genotype of the polymorphic loci shown in amino acids (aa) for each subject, the alternative MiHA variant and its predicted HLA binding affinity (IC_{50}), the differential predicted HLA binding affinity of the variant relative to the detected peptide (IC_{50} ratio) and the dbSNP identification when the ns-SNP corresponds to a known SNP. MiHAs tested in cytotoxicity assays are indicated with * (see Figure 5a). IC_{50} values of the alternative MiHA variants and IC_{50} ratios are shown in italics when they show a fold difference ≥ 2 relative to the detected MiHAs. Further features can be found in Supplementary Data 2.

for the corresponding HLA molecule (IC_{50} difference ≥ 2x). Nine of our best characterized MiHAs are novel whereas one (KEFEDGIINW) corresponds to the allelic variant of a previously reported MiHA (KEFEDDIINW)[44] that has been recently identified[45]. Four MiHAs

were exclusively detected by MS in one of the subjects although they derived from a shared allele (Table 1B and Fig. 4). In all cases the MiHA was detected in the subject homozygous for the corresponding allele but not in the heterozygous subject (Table 1). This suggests that zygosity influences MiHA abundance and that low abundance MiHAs may fall below the MS detection threshold in heterozygous subjects. Consistent with this, the MS intensity for these four MiHAs was low in the homozygous subject (Supplementary Data 3). Six MiHAs were coded by an allele present only in one subject (Table 1A), and were thus potentially immunogenic for the other sibling. We further validated the peptide sequence using MS/MS from their respective synthetic peptide (Supplementary Fig. 5a). Furthermore, we confirmed the presence of the ns-SNP in the corresponding DNA and/or cDNA regions of these six MiHAs in both subjects by Sanger sequencing (Supplementary Fig. 6). Then, we determined the immunogenicity of four of these MiHAs by cytotoxicity assays. Peripheral blood mononuclear cells (PBMCs) from the MIP-negative subject were stimulated with autologous dendritic cells pulsed with an unshared MIP detected in the other subject. Primed cells were restimulated with autologous B-LCLs pulsed with the same peptide, then tested for *in vitro* cytotoxicity activity against autologous B-LCLs (MIP-negative) and allogeneic B-LCLs (MIP-positive). In all cases, *in vitro* generated MiHA-specific cytotoxic T lymphocytes selectively killed allogeneic MiHA-positive B-LCLs but not autologous MiHA-negative B-LCLs (Fig. 5a).

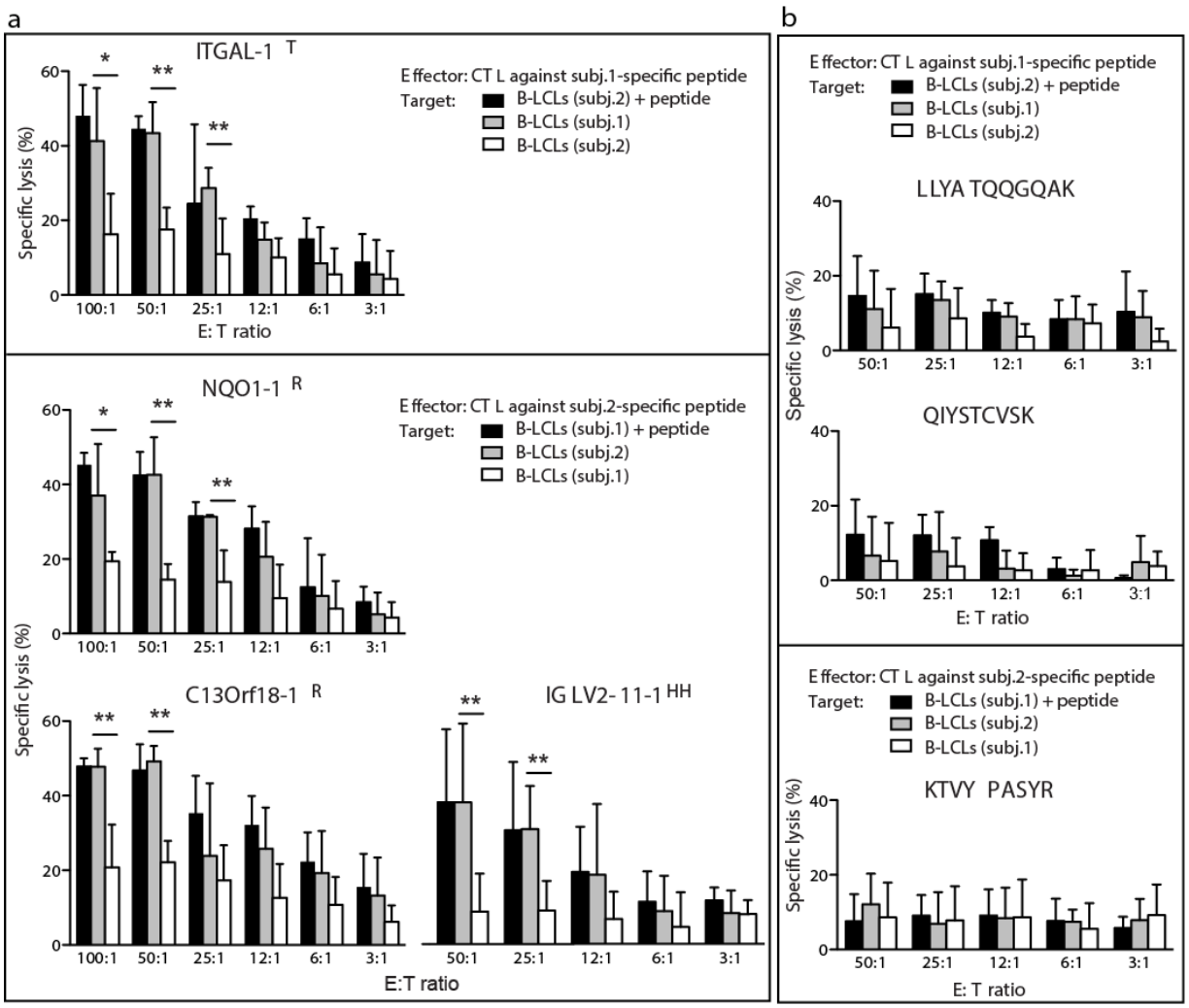


Figure 5. Only polymorphic MIPs are immunogenic. Frozen PBMCs from the MIP-negative subject were thawed and stimulated with autologous dendritic cells pulsed with an unshared MIP detected in the other individual. Primed cells were restimulated with irradiated autologous B-LCLs pulsed with the same peptide for another 7 days. Restimulated cells were tested for *in vitro* cytotoxicity activity against autologous B-LCLs pulsed with the relevant peptide (positive control, black), unpulsed autologous B-LCLs (negative control, white), or MIP-positive allogeneic B-LCLs (test, grey) at various effector-to-target (E:T) ratios. The minimal cytotoxic activity against unpulsed autologous B-LCLs is most likely due to recognition of EBV epitopes. Average and s.d. of three or four independent experiments are shown. Significant differences are indicated by * $P < 0.05$ or ** $P < 0.01$, 2-tailed Student *t*-test. (**a**) MIPs encoded by polymorphic loci and detected exclusively in one subject. (**b**) MIPs encoded by non-polymorphic loci but detected exclusively in one subject.

We next sought to determine why some MIPs derived from non-polymorphic regions were detected by MS in only one subject (n = 633-14 = 619; Fig. 3a**)**. Could they be MiHAs whose presence is regulated by *cis*- or *trans*- acting polymorphisms (outside of the MIP-coding genomic sequence) that would affect MIP processing[22, 24]? The MS/MS spectra of each of these MIPs were manually validated and, to further confirm the absence of the MIPs in one of the two subjects, we searched these MIPs in two additional biological replicates from each cell line. Most non-polymorphic MIPs found in only one subject were detected in only one or two replicates (Fig. 3a). This suggests that the presence of these MIPs was inconsistent, perhaps reflecting in part the limited sensitivity of MS. However, 41 unshared MIPs could not be discarded so easily because they were detected in 3-6 replicates of one sibling and absent in 6 replicates of the other sibling. With the exception of two cases, exclusive detection of these MIPs in one of the siblings was not caused by interindividual differences in abundance of the MIP-source transcript (Supplementary Fig. 7a) or in the expression of the MIP-coding exon (Supplementary Fig. 7b), nor by differences in the expression of genes involved in the antigen processing and presentation pathway (Supplementary Data 4). We therefore selected for further analyses the three most enticing MIPs coded by non-polymorphic regions but detected by MS in only one subject: MIPs showing the best values for the predicted binding affinity, MS intensity, reproducibility and Mascot score (Supplementary Data 3). We further confirmed their absence in one of the subjects by comparing the corresponding extracted ion chromatograms (Supplementary Fig. 5b) and validated their MS/MS spectra with synthetic peptides (Supplementary Fig. 5c). We reasoned that if these MIPs were MiHAs, they should be immunogenic, even if their presence was dictated by unidentified polymorphisms outside of the MIP-coding genomic sequence. None of the tested MIPs could elicit the generation of cytotoxic

T cells in the MIP-negative sibling (Fig. 5b). We therefore failed to discover a single MiHA among MIPs coded by non-polymorphic regions. The most parsimonious explanation is that these MIPs were simply differentially expressed peptides whose abundance was below the MS detection threshold in B-LCLs from one subject. A plausible explanation would be that exclusive detection of these MIPs in one subject reflects cellular differences caused by EBV infection during the establishment of the B-LCLs and/or clonal variation[46]. Accordingly, we conclude that identification of MiHAs absolutely requires a combination of MS and genomic data. Reliance solely on MS detection would overestimate the number of MiHAs. In contrast, the use of personalized databases based on whole exome and transcriptome sequencing allows to rapidly identifying genuine MiHAs coded by polymorphic loci.

2.7.4 The global imprint of ns-SNPs on the MIP repertoire

In order to assess the global imprint of ns-SNPs on the MIP repertoire, we asked the question: what proportion of ns-SNPs between our two subjects were located in MIP-coding exomic sequences? By comparing the combination of whole exome and RNA-seq data from our two subjects, we found a total of 4,833 ns-SNPs, 87% of which are reported as "validated" in dbSNP (Fig. 2**,** track 4 and Supplementary Data 2). Overall, 26 of these ns-SNPs were located in regions coding for 22 MiHAs identified by MS, of which 13 originated from a single gene (Fig. 4, n=6+4+3) and are depicted in the Circos plot (Fig. 2 track 4 vs. 7 blue, orange and pink**)** and 9 have an ambiguous origin (Fig. 4**,** n=4+0+5 and Supplementary Data 3). The 13 unambiguously assigned MiHAs were exclusively detected (Table 1) or shared (Table 2A) at

the peptidomic level. Thus, from a genomic perspective, only 0.5% of all ns-SNPs (26/4,833) found between our subjects were represented in their MIP repertoire.

Table 2. MiHAs detected in both subjects and coded by loci harboring ns-SNPs. For some MiHAs, one subject was homozygous and one subject heterozygous at the MiHA locus (A), whereas for other MiHAs both subjects were heterozygous at the MiHA locus (B).

	MiHA name	Detected MiHA sequence	Gene symbol	HLA allele	IC_{50} (nM)	aa sub. 1	aa sub. 2	Alternative MiHA variant	IC_{50} nM	IC_{50} ratio	dbSNP
A	MCPH1-1^{R}	EEINLQ**R**NI	*MCPH1*	B*44:03	503	RR	RI	EEINLQ**I**NI	212	0.4	rs2083914
	MDM1-1^{I}	V**I**QERVHSL	*MDM1*	B*08:01	61	IT	II	V**T**QERVHSL	*401*	*6.6*	rs962976
	FAM82B-1^{K}	VMGNPGTF**K**	*FAM82B*	A*03:01	23	KN	KK	VMGNPGTF**N**	*15,374*	*668*	rs6980476
B	TMEM132A-1^{A}	AAADRVGP**A**A	*TMEM132A*	C*16:01	1,236	AP	AP	AAADRVGP**P**A	1,203	1	---
	MAGEF1-1^{A}	ALAAK**A**LAR	*MAGEF1*	A*03:01	136	AS	AS	ALAAK**S**LAR	109	0.8	rs10937187
	TRIP11-1^{K}	DVQ**K**KLMSL	*TRIP11*	B*08:01	216	KN	KN	DVQ**N**KLMSL	*534*	*2.5*	rs145868557
	IMMT-1^{S}	KQ**S**ASQLQK	*IMMT*	A*03:01	65	SP	SP	KQ**P**ASQLQK	421	*6.5*	rs1050301
	DLGAP5-1^{H}	KTY**H**VTPMTPR	*DLGAP5*	A*03:01	27	HQ	HQ	KTY**Q**VTPMTPR	48	1.8	rs8010791
	ZWINT-1^{R}	QELD**G**VFQKL	*ZWINT*	B*44:03	366	RG	RG	QELD**R**VFQKL**	197	0.5	rs2241666
	MIIP-1^{K}	SEESAVP**K**RSW	*MIIP*	B*44:03	235	KE	KE	SEESAVP**E**RSW	245	1.0	rs2295283

Selected features of the MiHAs are shown: the detected amino acid sequence (polymorphic residues are highlighted in bold underlined), the source gene, the HLA molecule for which the MiHA has the best predicted binding affinity (IC_{50}), the translated genotype of the polymorphic loci shown in amino acids (aa) for each subject, the alternative MiHA variant and its predicted HLA binding affinity (IC_{50}), the differential predicted HLA binding affinity of the variant relative to the detected sequence (IC_{50} ratio) and the dbSNP identification when the ns-SNP corresponds to a known SNP. Note that in one case (marked with **) the alternative MiHA variant was detected by MS. IC_{50} values of the alternative MiHA variants and IC_{50} ratios are shown in italics when they show a fold difference ≥ 2 relative to the detected MiHAs. Further features can be found in Supplementary Data 2.

2.7.5 Identification of MiHAs among shared MIPs

Among 3,835 shared MIPs (Fig. 3a), 20 were encoded by bi-allelic loci and therefore represent MiHAs (Fig. 4**,** n=8+12). These shared MIPs would not be immunogenic for our subjects but would be immunogenic for subjects homozygous for the alternative allele. In eight cases, one subject was homozygous for a dominant MiHA allele (AA) and the other subject was heterozygous for the dominant and a recessive allele (AB) (Fig. 4**)**. The origin of five of these eight MiHAs was ambiguous (they could derive from several genes), whereas the other three MiHAs were assigned to a single gene (Table 2 and Fig. 4). The exome of our subjects shared 3,774 heterozygous loci (Fig. 2, track 5). Twelve MiHAs derived from such bi-allelic loci for which our subjects shared the same heterozygous genotype (AB). Eight of these 12 MiHAs could be unambiguously assigned to a single gene (Fig. 2 track 7 in green, Fig. 4 and Table 2B). The two alleles were co-dominant in one case, whereas only one allele was dominant (identified by MS) in the other cases. Notably, in four of the shared MiHAs, the product of the recessive allele was predicted to have a lower MHC binding affinity than the product of the dominant allele (Table 2**)**.

2.7.6 Differences in the MIP repertoire of HLA-identical siblings

Comparison of genomic and proteomic data from our subjects led to the discovery of 34 MiHAs (Fig. 4), of which 21 were unambiguously assigned to a specific gene (Fig. 2, track 7 and Tables 1 and 2). Out of 34 MiHAs, 14 were found in only one of the two subjects whereas 20 MiHAs were shared MIPs (Fig. 4). Without considering the 4 MiHAs that were exclusively

detected in one subject but that derived from a shared allele (Table 1B), this means that out of 4,468 MIPs only 10 (0.22%) would be immunogenic for one of our subjects. Assuming that unshared non-polymorphic MIPs are not immunogenic (Fig. 5b), this means that each subject would be tolerant to about 99.8% of the MIPs found on the B-LCLs of this sibling. The use of personalized databases for tandem MS sequencing was instrumental in the discovery of many MiHAs. Eleven of the 21 MiHAs listed in Tables 1 and 2 would have been missed in the absence of personalized databases, because these 11 peptides were absent in the Uniprot database.

2.7.7 Polymorphic MIP coding regions at the population level

We searched in the dbSNP database for validated ns-SNPs in the genomic sequences coding our 4,468 MIPs. We found that at the population level, 88% of our MIP coding sequences were invariant whereas 12% contained at least one ns-SNP: 670 ns-SNPs were found in the genomic region coding for 536 MIPs (Fig. 6a,b and Supplementary Data 5). Hence, at the population level, 536 MiHAs can be presented by the five HLA class I molecules studied herein: HLA-A*03:01, -A*29:02, -B*08:01, -B*44:03 and –C*16:01. Further studies will be required to determine the number of dominant and recessive peptide variants encoded by these 536 MiHA loci.

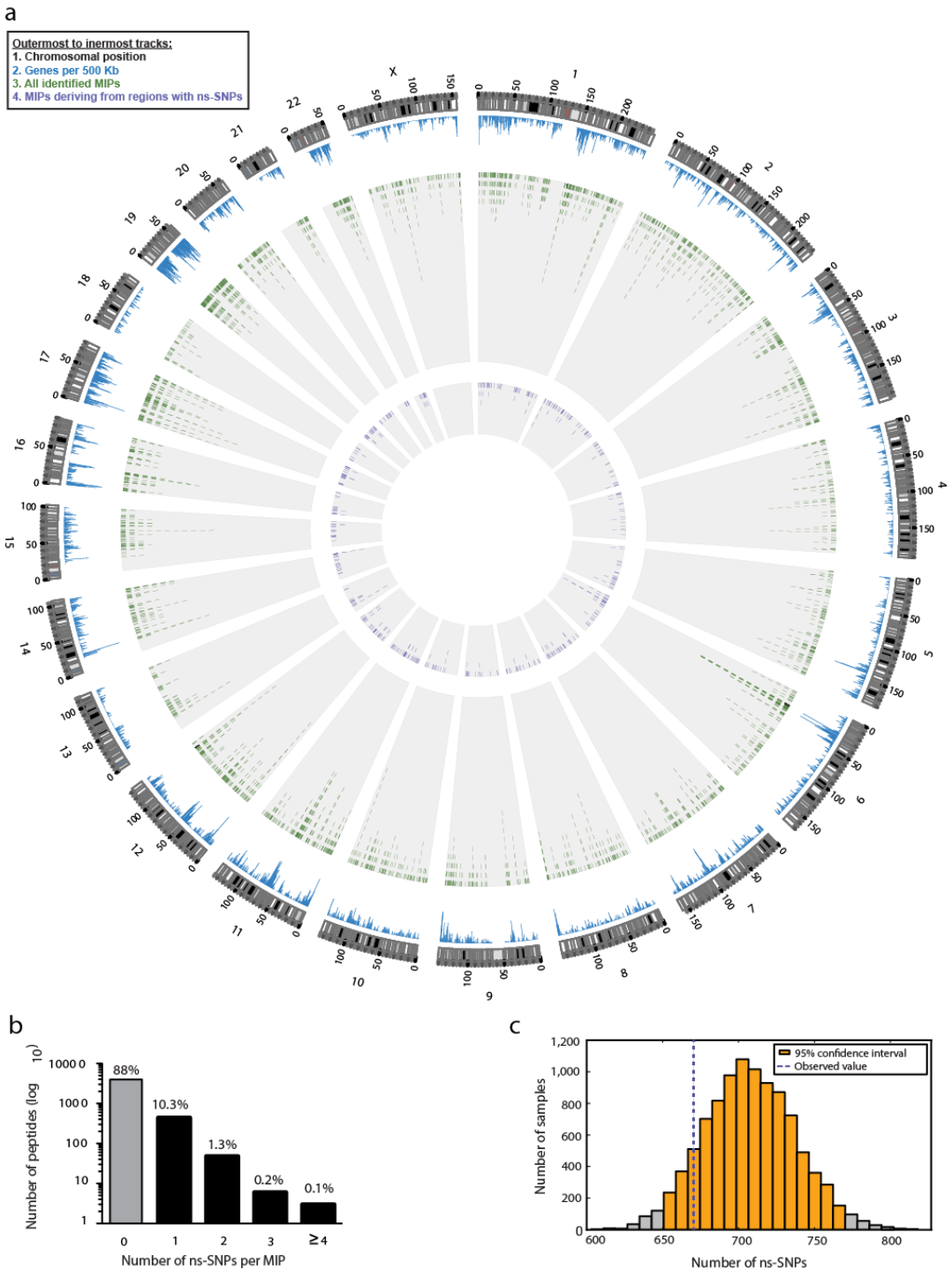


Figure 6. Frequency of ns-SNPs in the MIP coding exome. (**a**) Circos plot illustrates the relative proportion of polymorphic MIPs (n = 536) in the immunopeptidome and the genomic location of their coding loci. (**b**) Histogram showing the number and percentages of MIP coding regions containing ns-SNPs in the global population. We used dbSNP to find validated ns-SNPs in the exomic sequences

encoding the 4,468 MIPs identified in our subjects. In the case of MIPs deriving from multiple source regions, the average number of ns-SNPs of all possible MIP source regions was calculated. (**c**) The 4,468 MIPs of our subjects were encoded by 13,404 nucleotides. We performed 10,000 random samplings of 4,468 exomic sequences (containing a total of 13,404 nucleotides) from the human reference exome (Ensemble GRCh37.65). In all samplings, the frequency of exomic sequences coding for 8-,9-,10- and 11-mers was identical to the frequency found in the 4,468 MIP coding sequences from our subjects. The histogram depicts the distribution of validated ns-SNPs (dbSNP) in exomic sequences from the global population found in 10,000 random samplings of the whole exome. The average number of ns-SNPs of all random samplings was 708 (s.d. 30.4, 95% confidence interval: 650-768 shown in orange). The blue dotted line shows the number of ns-SNPs (n = 670) in the exomic sequences coding for the MIPs detected in our subjects.

2.7.8 Bias in favor or against ns-SNPs in MIP coding regions

We next wished to compare, in the global population, the frequency of ns-SNPs in the whole exome vs. the frequency in the 4,468 exomic sequences coding for the MIPs identified herein. To this end, we designed a bootstrap procedure (10,000 iterations) based on random samplings of 4,468 peptide-coding regions (13,404 base pairs/sampling) from the human reference exome (Ensemble GRCh37.65). For each sampling, we then calculated the number of validated ns-SNPs reported in dbSNP (Fig. 6c). Each sampling contained the same proportion of exomic sequences coding for 8-,9-,10- and 11-mers as the MIP coding sequences from our subjects. We found that the number of ns-SNPs in the MIP coding exome (n = 670) fell in the range of ns-SNPs found in 10,000 random samplings of the whole exome (average = 708; 95% confidence interval 650-768). We therefore conclude that the MIP coding exome reflects the frequency of ns-SNPs in the whole human exome.

2.8 Discussion

MS is the sole method that enables direct identification of MIPs and large scale analyses of the MIP repertoire[1, 15]. Indirect predictions based on reverse immunology approaches are fraught with false discovery rates that may reach 95%[47, 48]. Currently, MS sequencing has been largely limited to peptides represented in the reference UniProt database. Our work demonstrates that the universe of peptides identified by MS can be expanded and refined by using personalized databases that include whole exome and transcriptome sequencing data.

As well stated by J. Yewdell *et al.* "Despite the fact that quantitative aspects of systems are critical to their understanding, they are frequently ignored" [49]. In line with this concept, our data provide the answer to a longstanding question: what is the proportion of invariant vs. polymorphic MIPs presented by MHC molecules? In other words, to what extent do non-MHC genomic polymorphisms enhance the interindividual variability of the immunopeptidome? We found that, at the population level, at least one ns-SNP is found in 12% of exomic sequences coding the MIPs presented by five common HLA class I allotypes. That about 88% of the genomic landscape of the MHC class I immunopeptidome is invariant in the global population illustrates the overwhelming importance of the HLA genotype in defining the content of the MHC class I immunopeptidome.

In depth analyses of genomic and proteomic data revealed that about 0.5% of ns-SNPs between the exome of our subjects were represented in their MIP repertoire. Consequently, 10 MIPs coded by an unshared allele were unique to one subject and might elicit allogeneic T-cell responses from his sibling, as demonstrated for four of them. Integration of personalized genomic and proteomic data was essential for identification of these rare polymorphic MIPs

interspersed among thousands of non-polymorphic MIPs. Since the MIP repertoire is molded by the transcriptome, some MIPs are ubiquitous and others are cell lineage-specific[8, 50]. Accordingly, various cell types present non-identical MIP repertoires. MIPs derive mostly from transcripts expressed at medium to high levels (as opposed to very low or low levels), and about 8,500 transcripts are expressed at medium to high levels in B-LCLs[19]. We therefore posit that, at the organismal level, the total number of MiHAs derived from unshared ns-SNPs between two HLA-identical siblings would be about 2.5-fold the number found in B cells, assuming a total number of 21,000 human transcripts (i.e., 10 x (21,000/8,500) = 25). Unrelated individuals share fewer gene sequences than siblings. As a consequence, it has been calculated that the frequency of unshared MiHAs is increased by about 1.8-fold in unrelated (HLA-matched) subjects relative to siblings[23]. Thus, two unrelated HLA-identical subjects would display about 45 unshared MHC class I-restricted MiHAs. Of note, these numbers might increase with better sequencing coverage of difficult regions (e.g. GC-rich) and more sensitive MS instruments. As illustrated here, four low abundance MiHAs could only be detected in the homozygous but not in the heterozygous individual. Furthermore, our estimate could vary depending on the cell type and it does not take into account MiHAs presented by MHC class II proteins. Though only six MHC class II-restricted MiHAs have been discovered in humans[24, 51], a fair estimate of their repertoire will require systems-level studies using methods such as the one described herein.

All MHC antigens are dominant. Our data show that this is not the case for MiHAs. With a single exception, all MiHA loci had one dominant (MIP generating) and one recessive (no MIP generated) allele (Tables 1 and 2). This observation is clearly consistent with population analyses of 10 well characterized autosomal MiHA loci: only one locus has two dominant alleles[23]. For slightly less than 50% of our recessive alleles, the absence of MIP could be

explained by a decreased MHC binding affinity of peptides. For the other recessive alleles, the absence of MIP must be due to interference of the polymorphism with some step in MIP processing that precedes MHC binding (e.g., cleavage by the proteasome or other proteases)[3,38]. With tens of thousands of proteins, mammalian cells are the most complex entity in the antigenic universe faced by our immune system[52]. Theoretical estimates suggest that the immunopeptidome contains 0.1% of the 9-mer sequences present in the proteome[1]. Few peptides win the fierce competition for inclusion in the immunopeptidome. Thus, if we consider MiHAs coded by dominant alleles as winners, it follows that in most cases a single ns-SNP is sufficient to transform winners into losers (the recessive alleles). This is an eloquent reminder that we cannot predict the molecular composition of the immunopeptidome based on our limited understanding of the complexity of the MIP processing pathway.

Allogeneic hematopoietic cell transplantation has led to the discovery of the allogeneic graft-versus-leukemia (GVL) effect, which remains the most widely effective strategy for cancer immunotherapy in humans. GVL is mediated mainly, if not exclusively, by donor T cells that recognize host MiHAs. In line with recent progress in the field of cell therapy, MiHAs are therefore attractive targets for adoptive T-cell immunotherapy of cancer, particularly hematologic cancers[31-36]. However, because of the low number of molecularly defined human MiHAs, less than 30% of patients would currently be eligible for immunotherapy targeted to specific MiHAs[53]. Our report reveals a strategy for high-throughput MiHA discovery that could greatly accelerate the development of MiHA-targeted immunotherapy.

Our genoproteomic method combining next generation sequencing and MS shows how it is possible to accurately identify by MS any MIP, provided that its source DNA or RNA has been sequenced. The personalized protein databases could be further refined by including other

types of polymorphisms such as indels and using linkage disequilibrium information to diminish the number of possible proteins that will be expressed in an individual given his SNPs. This approach opens new avenues in systems immunology and should be invaluable for exploration of several "black holes" in the immunopeptidome. One particularly important black hole is the "cancer immunome" [54]. Compelling evidence suggests that the most immunogenic antigens present on cancer cells are mutant peptides derived from the numerous mutations found in neoplastic cells[55-57]. However, tumor-specific mutant peptides (alike MiHAs) are not detected by standard large scale MS approaches. We posit that our method should enable discovery of tumor-specific peptides (the product of somatic mutations) with the same accuracy as MiHAs (the product of germline genetic polymorphisms). Accordingly, our next priority will be to use this method to explore the impact of the cancer mutations on the immunopeptidome of cancer cells.

2.9 Methods

Cell culture and HLA typing

This study was approved by the Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont and all subjects provided written informed consent. Because fresh blood samples were required for cytotoxicity assays we elected to generate new B-LCLs from available donors, instead of studying the highly characterized B-LCLs from the Centre d'Etude du Polymorphisme Humain. PBMCs were isolated from blood samples of 2 non-twin HLA-identical Caucasian female siblings. B-LCLs were derived from PBMCs with Ficoll-Paque Plus (Amersham) followed by EBV infection as described[58]. High-resolution HLA genotyping was performed at the Maisonneuve-Rosemont Hospital. The two siblings are HLA-A*03:01,*29:02; B*08:01,*44:03; C*07:01,*16:01; DRB1*03:01,*07:01.

RNA extraction and preparation of transcriptome libraries

Total RNA was isolated from 5 million B-LCLs using RNeasy mini kit including DNase I treatment (Qiagen) according to the manufacturer's instructions. Total RNA was quantified using the NanoDrop 2000 (Thermo Scientific) and RNA quality was assessed with the 2100 Bioanalyzer (Agilent Technologies). Transcriptome libraries were generated from 1 μg of total RNA using the TruSeq RNA Sample Prep Kit (v2) (Illumina) following the manufacturer's protocol. Briefly, poly-A mRNA was purified using poly-T oligo-attached magnetic beads using two rounds of purification. During the second elution of the poly-A RNA, the RNA was fragmented and primed for cDNA synthesis. Reverse transcription of the first strand was

performed using random primers and SuperScript II (InvitroGene). A second round of reverse transcription was done to generate a double-stranded cDNA, which was then purified using Agencourt AMpure XP PCR purification system (Beckman Coulter). End repair of fragmented cDNA, adenylation of the 3' ends and ligation of adaptors were completed following the manufacturer's protocol. Enrichment of DNA fragments containing adapter molecules on both ends was done using 15 cycles of PCR amplification and the Illumina PCR mix and primers cocktail.

DNA extraction and exome capture

Genomic DNA was extracted from 5 million B-LCLs using the PureLink Genomic DNA Mini Kit (Invitrogen) according to the manufacturer's instructions. DNA was quantified and quality-assessed using the NanoDrop 2000 (Thermo Scientific). Genomic libraries were constructed from 1µg of genomic DNA using the TruSeq DNA Sample Preparation Kit (v2) (Illumina) following the manufacturer's protocol. We used 500 ng of DNA-Seq libraries for hybrid selection-based exome enrichment with the TruSeq exome enrichment kit (Illumina) according to the manufacturer's instructions.

Whole transcriptome sequencing, exome sequencing and mapping

Paired-end (2 x 100 bp) sequencing was performed using the Illumina HiSeq2000 machine running TruSeq v3 chemistry. Two RNA-Seq or four exomic libraries were sequenced per lane (8 lanes per slide). Cluster density was targeted at around 600-800k clusters/mm^2. The Illumina chastity quality filter was used to remove the low-quality reads. The chastity of a base

call is the ratio of the intensity of the greatest signal divided by the sum of the two greatest signals. Reads passed this filter if no more than one base call in the first 25 cycles had a chastity <0.6. More than 96% of the reads passed this filter **(**Supplementary Data 1**)**. Sequence data were mapped to the human reference genome (hg19) using the Casava 1.8.1 and the Eland v2e mapping softwares (Illumina). First, the *.bcl files were converted into compressed FASTQ files, following by demultiplexing of separate multiplexed sequence runs by index. Single reads were aligned to the human reference genome using the multiseed and gapped alignment method. Multiseed alignment works by aligning the first seed of 32 bases and consecutive seeds separately. Gapped alignment extends each candidate alignment to the full length of the read and allows for gaps up to 10 bases. The following criteria were applied: i) a read contains at least one seed that matches with at most 2 mismatches without gaps and ii) gaps were allowed for the whole read, as long as they correct at least five mismatches downstream. For each candidate alignment a probability score, which is based on the sequencing base quality values and the positions of the mismatches, was calculated. The alignment score of a read, which is expressed on the Phred scale, was computed from the probability scores of the candidate alignments. The best alignment for a given read corresponded to the candidate alignment with the highest probability score and was kept if the alignment score exceeded a threshold. Read alignments were further filtered out if they contained adjacent insertion/deletion events or if paired-end anomalies were present. Reads that mapped at 2 or more locations were not included in further analyses. For the exome paired-end libraries, the best scoring alignments for each half of the pair were computed and compared to find the best paired-read alignments according to the estimated insert size distribution. In the case of RNAseq libraries, an additional alignment was performed against splice junctions and contaminants (mitochondrial and ribosomal RNA).

Sequences mapping to contaminants were discarded whereas reads uniquely mapping to splice junctions were kept and converted back to genome coordinates.

Quantification of transcript expression

We used two methods to estimate and compare transcript expression between subjects. In the first method, the Casava 1.8.1 software (Illumina) was used to estimate gene or exon expression levels (RNA-seq) measured as RPKM (i.e. Read Per Kilobases of exon model per Million mapped reads) using the following formula: Gene or exon RPKM = 10^9 x Cb/Nb L, where Cb is the number of bases that fall on the feature, Nb is the total number of mapped bases and L is the length of the feature in base pairs. We also used the DESeq package[59], which is based on raw counts, to compare transcript expression. Transcript expression level was not considered in SNP calling.

Identification of SNPs and read counting

Variant call, indel detection and read counting were done using the Casava 1.8.1 software (Illumina). Reads were re-aligned around candidate indels to improve the quality of variant calls and site coverage summaries. Individual base calls were further filtered based on mismatch density or ambiguity and the remaining base calls were used to predict site genotypes. Casava was also used to retrieve all SNPs observed between the reference genome (GRCh37.p2, NCBI) and the sequenced transcriptome and exome of our subjects. SNPs and indel calls near centromeres and within high-copy number regions were removed. For each called SNP, Casava calculates the most probable genotype (max_gt), and a Q-value expressing the probability of the

most probable genotype (Qmax_gt). The Q-value is a quality score that measures the probability that a base is called incorrectly and was used to filter out low-quality SNPs (see "*In silico* generated proteomes" section). SNPs sequenced with at least 5x coverage were kept. This information (.txt files) was loaded into an in-house python module, pyGeno[19], for further processing.

In silico generated proteomes (personalized databases)

We used various in-house scripts that rely on pyGeno for data retrieval, parsing and processing. We integrated the exome sequencing data to the transcriptome sequencing data. For every SNP found by transcriptome sequencing, we retained the most probable genotype if the Q-value (Qmax_gt) was ≥20, which corresponds to a 1% error rate (a higher quality score indicates a smaller probability of error). If the SNP was also covered by the exome sequencing, we included not only the most probable genotype found by RNA-seq but also all bases in common with the exome sequencing. Because exome sequencing is not affected by mRNA expression, we also included the genotypes of SNPs that were only found by exome sequencing and that had a Q-value ≥20. Lastly, we included all bases of SNPs called by both the transcriptome and exome sequencing, regardless of the Q-Value. The retained genotypes of all SNPs were then integrated in the reference genome (GRCh37.p2, fasta file) at their right position to construct a "personalized genome" for each subject. These personalized genomes were used to extract all transcripts reported in the Ensembl gene set (GRCh37.65, gtf file) for all chromosomes except for the Y chromosome and mitochondrial DNA. These transcripts were then *in silico* translated into proteins using the reading frame specified in the Ensembl gene set.

Considering that the vast majority of MIPs have a maximum length of 11 amino acids, we established a window of 21 amino acids centered at each heterozygous ns-SNP. When a window contained more than one SNP, we translated *in silico* all possible combinations and included them in the personalized databases (Fig. 1b). Finally, we compiled all translation products into two fasta file databases (one for each subject) that were used for the identification of MIPs (see "MS/MS sequencing and peptide clustering" section). Both resulting databases had a similar size, in terms of number of residues (36,007,210 in subject 1 and 36,010,026 in subject 2) and number of entries (95,806 in subject 1 and 95,687 in subject 2). Moreover their size is comparable to the size of the reference UniProt human database used (43,384,120 residues and 75,530 entries).

MS/MS sequencing and peptide clustering

Based on our previous studies on MS data reproducibility across technical and biological replicates[8], we prepared four biological replicates of 5 x 10^8 exponentially growing B-LCLs from each subject. MIPs were released by mild acid treatment, desalted on an HLB cartridge 30cc, filtered with a 3000Da cut-off membrane and separated into seven fractions by cation exchange chromatography using an off-line 1100 series binary LC system (Agilent Technologies) as previously described[8, 9]. Fractions containing MIPs were resuspended in 0.2% formic acid and analyzed by LC-MS/MS using an Eksigent LC system coupled to a LTQ-Orbitrap ELITE mass spectrometer (Thermo Electron). Peptides were separated on a custom C_{18} reversed phase column (150 μm i.d. X 100 mm, Jupiter Proteo 4 μm, Phenomenex) using a flow rate of 600 nL/min and a linear gradient of 3-60% aqueous ACN (0.2% formic acid) in 120

mins. Full mass spectra were acquired with the Orbitrap analyzer operated at a resolving power of 30 000 (at m/z 400). Mass calibration used an internal lock mass (protonated $(Si(CH_3)_2O))_6$; *m/z* 445.120029) and mass accuracy of peptide measurements was within 5 ppm. MS/MS spectra were acquired at higher energy collisional dissociation with a normalized collision energy of 35%. Up to 6 precursor ions were accumulated to a target value of 50000 with a maximum injection time of 300 ms and fragment ions were transferred to the Orbitrap analyzer operating at a resolution of 15000 at *m/z* 400.

Mass spectra were analyzed using Xcalibur software and peak lists were generated using Mascot distiller Version 2.3.2 (http://www.matrixscience.com). Database searches were performed against UniProt Human database (43,384,120 residues, released on April 2, 2013), databases specific to subjects 1 and 2 (34,976,580 and 34,990,381 residues, respectively, see "*in silico* generated proteome" section) and EBV_B95.8 database (40,946 residues), using Mascot (Version 2.3.2, Matrix Science). To calculate the FDR, we performed a Mascot search against a concatenated target/decoy database using the human UniProt or subject-specific databases. The target represents the forward sequences and the decoy its reverse counterparts. Mass tolerances for precursor and fragment ions were set to 5 ppm and 0.02 Da, respectively. Searches were performed without enzyme specificity with variable modifications for cysteinylation (Cys), phosphorylation (Ser, Thr and Tyr), oxidation (Met) and deamidation (Asn, Gln). Raw data files were converted to peptide maps comprising *m/z* values, charge state, retention time and intensity for all detected ions above a threshold of 8,000 counts using in-house software (Proteoprofile)[9]. Peptide maps corresponding to all identified peptide ions were aligned together to correlate their abundances across sample sets and replicates. The MS/MS spectra of unshared MIPs were validated manually.

Identification of MIPs

MIP identification was based on four criteria: i) the canonical MIP length of 8 to 11 amino acids, (ii) the predicted MHC binding affinity given by the NetMHCcons algorithm[43], (iii) the Mascot score, which reflects the quality of peptide assignation, and (iv) the FDR, which indicates the proportion of decoy (false) vs. target (true) identifications. First, we evaluated the correlation between these parameters. We found a strong correlation (0.88) between FDR values <60% and MHC binding affinity values ≤ 1,750 nM for all 8-11mers (Supplementary Fig. 1). Indeed, the proportion of peptides with an MHC binding affinity ≤ 1,750 nM increases as the FDR decreases (Supplementary Fig. 2a**).** This correlation was specific to MIPs, since no correlation was found for random peptides (Supplementary Fig. 1 and 2b). These results show that low FDR values allow enrichment of high affinity peptides (MHC binding affinity ≤ 1,750 nM) and thus of MIPs. However, the drawback of using a stringent low FDR as the main filter is that the total number of identifications considerably decreases (Supplementary Fig. 2a**)** as well as the proportion of small peptides (8-9 mers) identified (Supplementary Fig. 2c**)**. Accordingly, the relative proportion of peptides found in target vs. decoy decreased with increasing peptide length[60], in accordance with the notion that short peptides such as MIPs generally require higher Mascot scores to achieve a low FDR. Moreover, the tandem MS fragment ions of MIPs are less predictable and evenly distributed than those of tryptic peptides which further complicate their assignment by database search engines such as Mascot. To set a more suitable Mascot score threshold for high-throughput MIP detection, we evaluated the

relation between the Mascot score and the predicted binding affinity for all 8-11 mer peptides identified with an FDR ≤ 5% (Fig. 1c). Then, we calculated the number of MIPs identified with all combinations of Mascot score and predicted binding affinity. We found that the highest number of MIP identifications was obtained by combining a Mascot score ≥ 21 and an MHC binding affinity ≤1,250 nM at a 5% FDR (Fig. 1c).

MS/MS validation of a subset of MIPs

Polymorphic and non polymorphic MIPs exclusively detected in one of the two subjects (Table 1 and Supplementary Data 3) were synthesized by Bio Basic Inc. and JPT peptide technologies. Subsequently, 500 fmols of each peptide were injected in the LTQ-Orbitrap ELITE mass spectrometer using the same parameters as those used to analyse the biological samples.

Ns-SNPs found in MIP coding regions in the population

For each MIP, we retrieved the coordinates of the peptide-coding DNA region. These coordinates were then used to extract both the corresponding reference sequence and all non-synonymous validated SNPs reported by dbSNP (Build 137) for that region. For MIPs deriving from multiple source regions, the number of ns-SNPs reported, corresponds to that of the MIP source region possessing the maximal number of ns-SNPs.

Random peptide sampling

We constructed a genome-wide index. To do so, we indexed every coding sequences reported in the Ensembl gene set (GRCh37.65), except for those located in the Y chromosome or the mitochondrial DNA, into a segment tree. Next, we kept only the first layer of the tree and removed the gaps between the indexed regions, effectively transforming the tree into a coding DNA sequence list, which was used for the random peptide sampling. For each of the 4,468 identified peptides, a random peptide of the same length and that fell entirely into a single coding DNA sequence, was chosen. Next, for each randomly selected peptide, we counted the number of ns-SNPs reported in dbSNP137 (validated and missense). The distribution was obtained after repeating the sampling of 4,468 random peptides 10,000 times.

PCR and Sanger sequencing

PCR amplification of the MiHA-encoding DNA and cDNA regions was performed with the Phusion ® High-Fidelity PCR kit (New England BioLabs). For each candidate, 1-2 pairs of sequencing primers were designed manually and with the PrimerQuest software (Integrated DNA Technologies, Supplementary Table 1), and were synthesized by Sigma. PCR products were purified with the PureLink Quick Gel Extraction Kit (Invitrogen). Sanger sequencing was performed on candidate DNA and cDNA at the IRIC's Genomics Platform. Sequencing results were visualized with the Sequencher software v4.7 (Gene Codes Corporation).

Cytotoxicity assays

Dendritic cells (DCs) were generated from frozen PBMCs, as previously described[61]. To generate cytotoxic T cells, autologous DCs were irradiated (4,000 cGy), loaded with 2 μM of

peptide and cultured for 7 days with freshly thawed autologous PBMCs at a DC:T cell ratio of 1:10. From day 7, responder T cells were restimulated for 7 additional days with irradiated autologous B-LCLs pulsed with the same peptide (B-LCL:T cell ratio 1:5). Expanding T cells were cultured in RPMI 1640 (Invitrogen) containing 10% human serum (Sigma-Aldrich) and L-glutamine. IL-2 (50 U/ml) was added for the last 5 days of the culture. Cytotoxicity assays were performed as described[9], with minor modifications. Briefly, B-LCLs were labeled with carboxyfluorescein succinimidyl ester (CFSE) (Invitrogen), extensively washed, irradiated (4,000 cGy) and then used as targets in cytotoxicity assays. Target cells were plated in 96-well U-bottom plates at 5,000 cells/well. Effector cells were added at different effector-to-target ratios in a final volume of 200 µl/well. Plates were centrifuged and incubated for 18h-20h at 37°C. Flow cytometry analysis was performed using a LSRII cytometer with a high throughput sampler device (BD Biosciences). The percentage of specific lysis was calculated as follows: [(number of $CFSE^+$ cells remaining after incubation with unpulsed target cells – number of $CFSE^+$ cells remaining after incubation with peptide-pulsed target cells) / number of $CFSE^+$ cells remaining after incubation with unpulsed target cells] x100.

Statistical analysis and data visualization

The 2-tailed Student *t*-test was used to identify differentially expressed MIPs and MiHAs that induced cytotoxicity. The 2-tailed Mann-Whitney test was used to compare the MHC binding affinity of unshared MIPs. Differentially expressed transcripts were identified with the DESeq package that uses a model based on the negative binomial distribution[59]. The Spearman correlation was used to evaluate the relation between differences in MIP abundance and

differences in MIP-coding gene or exon expression. The genomic location of identified MIPs including MiHAs and the RNA-seq and exome sequencing coverage were visualized with the Circos software[62]. The Integrative Genomics Viewer v2.0[63] was used to visualize and inspect regions coding MIPs including MiHAs.

2.10 References

1. de Verteuil,D., Granados,D.P., Thibault,P., & Perreault,C. Origin and plasticity of MHC I-associated self peptides. *Autoimmun. Rev.* **11**, 627-635 (2012).
2. Yewdell,J.W. DRiPs solidify: progress in understanding endogenous MHC class I antigen processing. *Trends Immunol.* **32**, 548-558 (2011).
3. Neefjes,J., Jongsma,M.L.M., 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).
4. Davis,M.M. *et al.* T cells as a self-referential, sensory organ. *Annu. Rev. Immunol.* **25**, 681-695 (2007).
5. Gilchuk,P. *et al.* Discovering naturally processed antigenic determinants that confer protective T cell immunity. *J Clin. Invest* **123**, 1976-1987 (2013).
6. 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).
7. Lemmel,C. *et al.* Differential quantitative analysis of MHC ligands by mass spectrometry using stable isotope labeling. *Nat Biotechnol.* **22**, 450-454 (2004).
8. Fortier,M.H. *et al.* The MHC class I peptide repertoire is molded by the transcriptome. *J. Exp. Med.* **205**, 595-610 (2008).
9. 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).
10. Illing,P.T. *et al.* Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature* **486**, 554-558 (2012).
11. Croft,N.P. *et al.* Kinetics of antigen expression and epitope presentation during virus infection. *PLoS Pathog.* **9**, e1003129 (2013).

12. 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).

13. Bensimon,A., Heck,A.J., & Aebersold,R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **81**, 379-405 (2012).

14. Milner,E., Barnea,E., Beer,I., & Admon,A. The turnover kinetics of MHC peptides of human cancer cells. *Mol. Cell. Proteomics* **5**, 357-365 (2006).

15. Adamopoulou,E. *et al.* Exploring the MHC-peptide matrix of central tolerance in the human thymus. *Nat Commun.* **4**, 2039 (2013).

16. Weinzierl,A.O. *et al.* Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface. *Mol. Cell. Proteomics* **6**, 102-113 (2007).

17. Petersdorf,E.W. & Hansen,J.A. New advances in hematopoietic cell transplantation. *Curr. Opin. Hematol.* **15**, 549-554 (2008).

18. The 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 55-65 (2012).

19. Granados,D.P. *et al.* MHC I-associated peptides preferentially derive from transcripts bearing miRNA recognition elements. *Blood* **119**, e181-e191 (2012).

20. Wallny,H.J. & Rammensee,H.G. Identification of classical minor histocompatibility antigen as cell-derived peptide. *Nature* **343**, 275-278 (1990).

21. Simpson,E., Roopenian,D., & Goulmy,E. Much ado about minor histocompatibility antigens. *Immunol Today* **19**, 108-112 (1998).

22. Roopenian,D., Choi,E.Y., & Brown,A. The immunogenomics of minor histocompatibility antigens. *Immunol. Rev.* **190**, 86-94 (2002).

23. Spierings,E. *et al.* Phenotype frequencies of autosomal minor histocompatibility antigens display significant differences among populations. *PLoS. Genet.* **3**, e103 (2007).

24. Warren,E.H. *et al.* Effect of MHC and non-MHC donor/recipient genetic disparity on the outcome of allogeneic HCT. *Blood* **120**, 2796-2806 (2012).

25. Morse,M.C. *et al.* The COI mitochondrial gene encodes a minor histocompatibility antigen presented by H2-M3. *J. Immunol.* **156**, 3301-3307 (1996).

26. Wang,W. *et al.* Human H-Y: a male-specific histocompatibility antigen derived from the SMCY protein. *Science* **269**, 1588-1590 (1995).

27. den Haan,J.M. *et al.* The minor histocompatibility antigen HA-1: a diallelic gene with a single amino acid polymorphism. *Science* **279**, 1054-1057 (1998).

28. Simpson,E. & Roopenian,D. Minor histocompatibility antigens. *Curr. Opin. Immunol.* **9**, 655-661 (1997).

29. Zuberi,A.R., Christianson,G.J., Mendoza,L.M., Shastri,N., & Roopenian,D.C. Positional cloning and molecular characterization of an immunodominant cytotoxic determinant of the mouse *H3* minor histocompatibility complex. *Immunity* **9**, 687-698 (1998).

30. Klein,C.A. *et al.* The hematopoietic system-specific minor histocompatibility antigen HA-1 shows aberrant expression in epithelial cancer cells. *J. Exp. Med.* **196**, 359-368 (2002).

31. Fontaine,P. *et al.* Adoptive transfer of T lymphocytes targeted to a single immunodominant minor histocompatibility antigen eradicates leukemia cells without causing graft-versus-host disease. *Nat. Med.* **7**, 789-794 (2001).

32. Spierings,E., Wieles,B., & Goulmy,E. Minor histocompatibility antigens - big in tumour therapy. *Trends Immunol.* **25**, 56-60 (2004).

33. Bleakley,M. & Riddell,S.R. Molecules and mechanisms of the graft-versus-leukaemia effect. *Nat. Rev. Cancer* **4**, 371-380 (2004).

34. Meunier,M.C. *et al.* T cells targeted against a single minor histocompatibility antigen can cure solid tumors. *Nat. Med.* **11**, 1222-1229 (2005).

35. Vincent,K., Roy,D.C., & Perreault,C. Next-generation leukemia immunotherapy. *Blood* **118**, 2951-2959 (2011).

36. Warren,E.H. *et al.* Therapy of relapsed leukemia after allogeneic hematopoietic cell transplant with T cells specific for minor histocompatibility antigens. *Blood* **115**, 3869-3878 (2010).

37. Hombrink,P. *et al.* Discovery of T cell epitopes implementing HLA-peptidomics into a reverse immunology approach. *J. Immunol.* **190**, 3869-3877 (2013).

38. Yewdell,J.W., Reits,E., & Neefjes,J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nature Rev. Immunol.* **3**, 952-961 (2003).

39. Perreault,C. The origin and role of MHC class I-associated self-peptides. *Prog. Mol Biol. Transl. Sci.* **92**, 41-60 (2010).

40. Wang,Z., Gerstein,M., & Snyder,M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57-63 (2009).

41. Sherry,S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).

42. Gebreselassie,D., Spiegel,H., & Vukmanovic,S. Sampling of major histocompatibility complex class I-associated peptidome suggests relatively looser global association of HLA-B*5101 with peptides. *Hum. Immunol.* **67**, 894-906 (2006).

43. Karosiene,E., Lundegaard,C., Lund,O., & Nielsen,M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64**, 177-186 (2012).

44. Akatsuka,Y. *et al.* Identification of a polymorphic gene, BCL2A1, encoding two novel hematopoietic lineage-specific minor histocompatibility antigens. *J. Exp. Med.* **197**, 1489-1500 (2003).

45. Hassan,C. *et al.* The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell Proteomics.* **12**, 1829-1843 (2013).

46. Choy,E. *et al.* Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS. Genet.* **4**, e1000287 (2008).

47. Popovic,J. *et al.* The only proposed T-cell epitope derived from the TEL-AML1 translocation is not naturally processed. *Blood* **118**, 946-954 (2011).

48. Robbins,P.F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* **19**, 747-752 (2013).

49. Princiotta,M.F. *et al.* Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity* **18**, 343-354 (2003).

50. de Verteuil,D. *et al.* Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules. *Mol Cell Proteomics* **9**, 2034-2047 (2010).

51. Spaapen,R.M. *et al.* Toward targeting B cell cancers with CD4+ CTLs: identification of a CD19-encoded minor histocompatibility antigen using a novel genome-wide analysis. *J Exp. Med* **205**, 2863-2872 (2008).

52. Malarkannan,S. *et al.* Differences that matter: major cytotoxic T cell–stimulating minor histocompatibility antigens. *Immunity* **13**, 333-344 (2000).

53. Bleakley,M. *et al.* Leukemia-associated minor histocompatibility antigen discovery using T-cell clones isolated by in vitro stimulation of naive $CD8^+$ T cells. *Blood* **115**, 4923-4933 (2010).

54. Kroemer,G. & Zitvogel,L. Can the exome and the immunome converge on the design of efficient cancer vaccines? *Oncoimmunology.* **1**, 579-580 (2012).

55. Lennerz,V. *et al.* The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc. Natl. Acad. Sci. U. S. A* **102**, 16013-16018 (2005).

56. Heemskerk,B., Kvistborg,P., & Schumacher,T.N. The cancer antigenome. *EMBO J.* **32**, 194-203 (2013).

57. Zitvogel,L., Galluzzi,L., Smyth,M.J., & Kroemer,G. Mechanism of action of conventional and targeted anticancer therapies: reinstating immunosurveillance. *Immunity* **39**, 74-88 (2013).

58. Tosato,G. & Cohen,J.I. Generation of Epstein-Barr Virus (EBV)-immortalized B cell lines. *Curr. Protoc. Immunol.* **Chapter 7**, Unit 7.22 (2007).

59. Anders,S. & Huber,W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

60. Elias,J.E. & Gygi,S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207-214 (2007).

61. Bollard,C.M. *et al.* Complete responses of relapsed lymphoma following genetic modification of tumor-antigen presenting cells and T-lymphocyte transfer. *Blood* **110**, 2838-2845 (2007).

62. Krzywinski,M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).

63. Robinson,J.T. *et al.* Integrative genomics viewer. *Nat Biotechnol.* **29**, 24-26 (2011).

2.11 Acknowledgments

We thank Wafaa Yahyaoui for technical assistance, and the personnel of the following IRIC core facilities: genomics, proteomics, bioinformatics and flow cytometry. We are thankful to Pierre Chagnon and Brian Wilhem for advice and thoughtful comments on exome and transcriptome sequencing. We also thank our blood donors. This work was supported by the Canadian Cancer Society (grant # 701564). D.P.G. is supported by a studentship from the Canadian Institutes of Health Research. C.P. and P.T. hold Canada Research Chairs in Immunobiology, and Proteomics and Bioanalytical Spectrometry, respectively. IRIC is supported in part by the Canada Foundation for Innovation, and the Fonds de Recherche Santé Québec.

2.12 Author contributions

D.P.G. and D.S. designed the study, performed experiments, analyzed data, prepared the figures and wrote the first draft of the manuscript. T.D. designed the study, developed pyGeno, performed bioinformatic analyses and contributed to the writing. O.C. and A.Z. developed bioinformatics tools for the analysis of MS data and prepared figures. C.L. performed analyses and prepared a figure. C.C. and M.P.H. performed experiments. G.B. prepared Circos figures and bioinformatics analyses. P.G. performed sequencing mapping and analysis. S.L. designed the study and discussed statistical analyses and results. P.T. and C.P. designed the study, analyzed data, discussed results, wrote the manuscript and contributed equally as senior authors. All authors edited and approved the final manuscript.

2.13 Additional information

MS/MS spectra: PASS00270 (PeptideAtlas, http://www.peptideatlas.org/), MHC-I peptide sequences: under revision in Immune Epitope Database (http://www.iedb.org/), RNA-seq data: GSE48918 (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/), exome data PRJNA210790 (NCBI Sequence Read Archive, http://www.ncbi.nlm.nih.gov/sra).

2.14 Competing Financial Interests

Université de Montréal has filed a patent related to the research presented in this manuscript.

2.15 Supplementary Informations

Les informations supplémentaires peuvent être retrouvées en annexe 2.

Chapitre 3

pyGeno: A Python package for precision medicine and proteogenomics

3.1 Résumé

pyGeno est une librarie Python destinée aux applications de médecine personnalisée impliquant la génomique et la protéomique. pyGeno intègre les séquences de référence ainsi que les annotations d'Ensembl, les polymorphismes génomiques provenant de la base de données dbSNP et les données de séquençage de nouvelle génération dans un cadre facile d'utilisation, rapide, avec usage de la mémoire optimisé, permettant ainsi à l'utilisateur de facilement explorer les génomes et protéomes personnalisés de leurs sujets d'étude. Comparativement à un programme autonome, pyGeno fournit un accès à l'expressivité complète de Python, un langage de programmation général. De ce fait, l'étendue des applications rendues possibles par pyGeno inclut à la fois des scripts courts, ainsi que des études à grande échelle sur l'ensemble du génome.

3.2 Contributions des auteurs

T.D : Écriture de pyGeno et de la première ébauche du manuscrit

C.P et S.L : Rédaction et révisions du manuscrit

3.3 Référence de publication

Daouda, T., Perreault, C., & Lemieux, S. (2016). pyGeno: A Python package for precision medicine and proteogenomics. F1000Research, 5.

3.4 Article

pyGeno: A Python package for precision medicine and proteogenomics

Tariq Daouda[1,2], Claude Perreault[1,3,4], Sébastien Lemieux[1,5]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, Montreal, Canada

[2]Department of Biochemistry, Faculty of Medicine, Université de Montréal, Montreal, Canada

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

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

[5]Department of Computer Science and Operations Research, Faculty of Arts and Sciences, Université de Montréal, Montreal, Canada

Corresponding authors: Tariq Daouda (tariq.daouda@umontreal.ca), Sébastien Lemieux (s.lemieux@umontreal.ca)

3.5 Abstract

pyGeno is a Python package mainly intended for precision medicine applications that revolve around genomics and proteomics. It integrates reference sequences and annotations from Ensembl, genomic polymorphisms from the dbSNP database and data from next-gen sequencing into an easy to use, memory-efficient and fast framework, therefore allowing the user to easily explore subject-specific genomes and proteomes. Compared to a standalone program, pyGeno gives the user access to the complete expressivity of Python, a general programming language. Its range of application therefore encompasses both short scripts and large scale genome-wide studies.

3.6 Introduction

High-throughput systems biology and precision medicine applications require the integration of data from many different sources. For instance, a significant part of precision medicine research revolves around the identification of relevant single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELS) and the study of their context[1]. Furthermore, recent studies in proteogenomics show that replacing traditional reference databases such as Uniprot[2] by customized databases that integrate the subject's genomic polymorphisms, can significantly improve the identification of peptides or proteins using mass spectrometry[3–6]. These applications usually require the integration of reference sequences, reference genome annotations, specific SNPs and INDELs along with an external SNP database such as dbSNP[7] for validation. The sheer amount of data generated by these studies rules out most spreadsheet analyses and requires tools that are both fast and memory efficient. Furthermore, these studies often require the collaboration of people with different sets of skills. Thus, it was important to us to develop a tool that is powerful enough to be integrated in complex high-throughput pipelines, while still being understandable by users with limited technical abilities. In contrast to other projects such as BioPython[8] and PyCogent[9] whose objective is to provide a general set of tools for bioinformatics, the primarily ambition behind pyGeno is to provide the community with a powerful genome and proteome exploration tool that can be easily integrated into scripts. The current version integrates gene set annotations and reference sequences from Ensembl[10] along with polymorphisms (both SNPs and INDELs) derived from dbSNP[7], and experimentally detected patient-specific polymorphisms.

To our knowledge pyGeno is the only available tool that provides this kind of integration in an easy-to-use and programming-friendly environment. Furthermore, more advanced users can rely on object-oriented inheritance to extend the functionalities of pyGeno to implement support for polymorphisms from other sources. pyGeno has been used with human and mouse genomes and should readily work with any diploid organism whose annotations are made available by Ensembl.

3.7 Methods

3.7.1 Design and implementation

pyGeno is written in Python, a language that enjoys a large set of well established and mature scientific libraries that are used in research fields such as physics, mathematics and bioinformatics[8, 11–13]. pyGeno gives users access to the full expressivity of Python to explore reference and patient-specific genomes and proteomes, by manipulating familiar objects such as genomes, chromosomes, genes, transcripts, proteins and exons. In order to make pyGeno as easy to use and learn as possible, we have created an interface where only one function, *get()*, can be used for almost any query. An example of usage can be seen in Figure 1. An integrated documentation is also available through the *help()* function.

The current version of pyGeno does not require any access to remote REST APIs. This results in more robust and faster processing since the application is not affected by connection speed or sudden changes to the server API. On the other hand it also implies that extra care must be taken regarding the optimization of the application.

Memory efficiency and speed are mainly achieved through the use of a custom lazy object-oriented database system that we have specifically written for pyGeno (https://github.com/tariqdaouda/rabaDB). When an object is loaded through the *get()* function, only a minimal version of it is served. The object fully develops only once the user accesses a field that is not present in the minimal version (Figure 1). The transformation is entirely transparent and does not require more memory than necessary to store the fully developed object. This is especially important, since most of the time users are only interested in specific regions of the genome, and do not require that the full genome be loaded into memory. Every loaded object is also a singleton, if the user asks for a previously loaded object, pyGeno will serve the object in memory.

Furthermore, this database system is built on top of SQLite version 3 (http://www.sqlite.org/), a serverless relational database. Because SQLite3 uses single files to store data, pyGeno's database can be easily backed up and shared by a simple copy/paste. Moreover, the files can be directly read, modified and analyzed through any SQLite3 client.

As with any other database system, indexes play a crucial role in determining the general performance. Within pyGeno's database, several reference genomes along with patient-specific data and versions of dbSNP can coexist. Therefore, building indexes for all the stored information would result in unnecessarily large databases. We therefore have taken the approach of giving the end user full control over indexation through the *ensureGlobalIndex()* and *dropGlobalIndex()* functions. Users can, for example, decide to index the field *'id'* of transcripts by using *Transcript.ensureGlobalIndex('id')* and dramatically improve queries based on transcript ids.

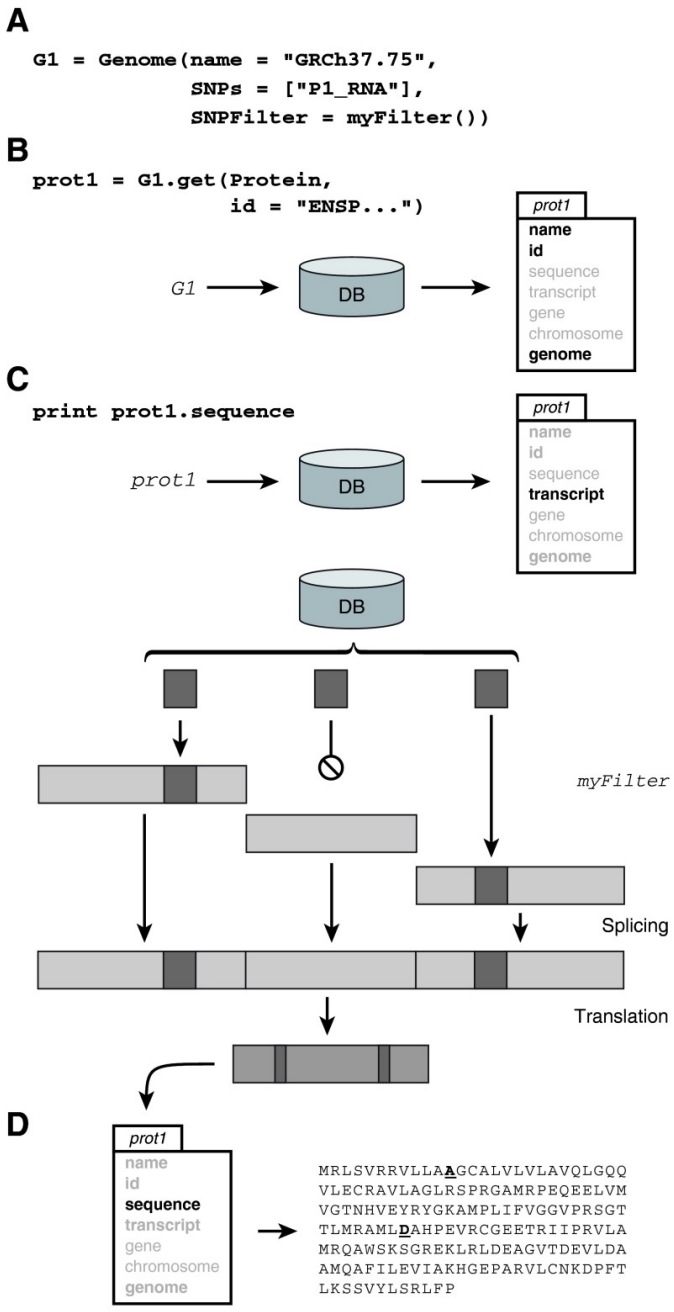


Figure 1. Extracting the subject-specific sequence of a protein. (A) Here we instantiate a personalized genome G1 by providing the Genome constructor with the name of a reference genome, a set of polymorphisms and a user defined SNP filter (for example a quality filter). (B) We then ask the get function of G1 to return a protein by id. The result is an object where only the fields in bold are fully loaded, other fields will be automatically loaded when and if accessed. (C) Asking for the currently unloaded sequence of the protein triggers the following sequence of events. The transcript, as well as the exons that encode for it, and any polymorphisms in their regions are loaded. The polymorphisms are filtered according to the filter provided to the genome constructor (for example, according to sequencing quality) and inserted at their corresponding locations. The exons are then assembled into the transcript sequence and the sequence is translated. (D) The sequence as well as the transcript are now fully loaded and the sequence of the precision protein is printed.

pyGeno's database is populated through imports of datawraps using *importSNPs* and *importGenome* functions. Datawraps are compressed archives that can be shared among co-workers, and are designed to solve the version and update problems. A datawrap contains at

least one file named manifest.ini that contains basic information about the package such as a description, a version and a maintainer, as well a list of files from which data must be extracted. It is possible to either compress these files within the archive, or to specify URLs from which the files can be downloaded.

In an effort to make pyGeno as easy to install as possible we have made it as dependency-free as possible. This approach has motivated our choice for SQLite3, since it is natively supported by Python 2.5 and above, and it also lead us to develop many tools that were subsequently integrated into pyGeno. Among theses tools are various functions for translating sequences, parsers for GTF/GFF, VCF, FASTA, FASTQ and CSV files, a progress bar, and an efficient way of annotating the genome called segment trees.

3.8 Personalized genomes

One of the biggest strengths of pyGeno is to allow the user to define personalized genomes. These genomes are built by combining a reference genome with sets of polymorphisms and a filtering function that returns the alleles to be inserted at the appropriate locus (Figure 1). Personalized genomes are a powerful tool that can go beyond the definition of patient-specific genomes. For instance, we recently used this tool to combine the results of both RNA- and DNA-seq data and create more robust personalized genomes that were used to identify protein-derived peptides by mass spectrometry[3]. Furthermore, because pyGeno loads the necessary parts of a given reference genome only once, a pyGeno application can handle several personalized genomes without significantly increasing its memory consumption.

3.8.1 Operation

pyGeno's only requirement is Python2 and we highly recommend version 2.7.6 or later. pyGeno can be easily installed using the *pip* package manager (https://pip.pypa.io/) by typing *pip install pyGeno* into command line interface. Alternatively, the latest developments can be obtained from the github repository. Once pyGeno's installation has been completed, the first action that users must perform is the importation of a reference genome datawrap. In order to simplify the process pyGeno comes with several datawraps that can be directly listed and installed using its *bootstrap* module. If the desired reference genome is not among the ones provided, users also have the possibility to create their own from scratch by following the steps described in the documentation. After the first reference genome importation, pyGeno is fully functional and users can further expand its database by importing other reference genomes or SNP sets.

3.9 Summary

We have developed pyGeno because, in an age where both precision medicine and DNA/RNA sequencing are becoming more and more important, we needed a tool that would allow us to easily work on personalized genomes that include subject-specific genomic features. Nowadays research teams are increasingly multidisciplinary and are composed of people with very different backgrounds. Since we wanted pyGeno to serve as a common language between users, we therefore took great care in making pyGeno easy to install, easy to use and optimized it so it can run on computers with limited resources (eg. laptops). The fact that pyGeno has been downloaded more than 12,000 times over its first year of existence suggests that there is indeed

a need for powerful user-friendly precision medicine tools. With pyGeno we have taken a rather unusual approach to user-friendliness. Instead of writing a program with a graphical user interface (GUI), we have decided to create a Python module that fully integrates within the Python environment. This ensures that users can leverage the full expressiveness of Python as well as the functionalities of other python modules such as SciPy and numpy[11], pandas (http://pandas.pydata.org/) and matplotlib[13], to meet their specific needs. Furthermore, it led us to think of the functions and objects the user manipulates as pyGeno's interface and we strived to make it as simple and easy to learn as possible.

In the past few years great technologies have been developed. Scripting languages such as Python and JavaScript have taken programming to a whole new level of simplicity, and are now fast enough to serve as foundations to large-scale projects. Freely available libraries such as D3.js (http://d3js.org/) allow for the creation of stunning data representations, that once coupled with tools such as pyGeno, could be used to create powerful interactive representations of biological data. The NoSQL movement has produced several new database systems from which developers can choose, offering them the opportunity to store sheer amounts of data with a flexibility that was not present only a few years ago. These technologies and many others are only waiting to be put together into ground breaking tools for the treatment of biological data. In life saving research areas, we believe that great tools that dramatically improve workflow efficiency are not a luxury but a necessity.

3.10 Software Availability

1. pyGeno is available from the Python Package Index (PyPI; https://pypi.python.org) via: pip install pyGeno.
2. Latest source code: https://github.com/tariqdaouda/pyGeno.
3. Documentation: http://pyGeno.iric.ca
4. Link to archived source code as at time of publication: https://zenodo.org/record/50587#.VyIP0UErJB0 (doi: 10.5281/zenodo.50587)
5. License: Apache License Version 2.0

3.11 Acknowledgements

We would like to thank Jean-Philippe Laverdure, Céline Laumont and Hillary Pearson for being the first users (outside of the developer) and the first testers of pyGeno.

3.12 Funding statement

This work was supported by the Canadian Cancer Society (Grant number 701564), assigned to Claude Perreault.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

3.13 References

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793–795.
2. Uniprot Consortium: Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013;41(Database issue):D43–47.
3. Granados DP, Sriranganadane D, Daouda T, *et al.* Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. Nat Commun. 2014;5: 3600.
4. Kim MS, Pinto SM, Getnet D, *et al.* A draft map of the human proteome. Nature. 2014;509(7502):575–581.
5. Wilhelm M, Schlegl J, Hahne H, *et al.* Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582–587.
6. Laumont CM, Daouda T, Laverdure JP, *et al.* Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun. 2016;7: 10238.
7. Sherry ST, Ward MH, Kholodov M, *et al.* dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–311.
8. Cock PJ, Antao T, Chang JT, *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423.
9. Knight R, Maxwell P, Birmingham A, *et al.* PyCogent: a toolkit for making sense from sequence. Genome Biol. 2007;8(8):R171.
10. Flicek P, Amode MR, Barrell D, *et al.* Ensembl 2014. Nucleic Acids Res. 2014;42(Database issue):D749–D755.
11. Jones E, Oliphant T, Peterson P, *et al.* SciPy: Open source scientific tools for Python.2001; [Online; accessed 2016-02-22].

12. SymPy Development Team: SymPy: Python library for symbolic mathematics.2014.

13. Hunter JD: Matplotlib: A 2d graphics environment. Comput Sci Eng. 2007;9(3):90–95.

Chapitre 4

Codon usage regulates biogenesis of human MHC class I-associated peptides

4.1 Résumé

Les peptides associés aux molécules du MHC de classe I (MIP) forment collectivement l'immunopeptidome, qui définit le soi immunitaire pour les lymphocytes T CD8. Les MIP régulent le développement et le fonctionnement des cellules T CD8, et représentent les cibles principales de l'immunosurveillance tumorale. Les MIP proviennent de régions spécifiques du génome. En effet, alors que toutes les protéines contiennent des séquences peptidiques capables de se lier aux molécules du MHC, la plupart ne se retrouve jamais à la surface des cellules. Afin de comprendre pourquoi l'immunopeptidome ne comprend qu'une petite fraction du transcriptome codant pour des protéines, les études se sont concentrées jusqu'à ce jour sur les événements post-traductionnels. Cependant, il a été démontré que la plupart des MIP proviennent de protéines dégradées soit pendant leur synthèse ou dans les minutes qui suivent. En utilisant diverses méthodes bio-informatiques, incluant des réseaux de neurones artificiels, nous avons analysé un grand ensemble de données de transcrits codant ou non pour des MIP. Nos résultats montrent que la biogenèse des MIP est régulée par l'usage des codons dans les régions des mRNA entourant les séquences codant pour les MIP. Notamment, nous avons trouvé que les codons synonymes ont un effet non-redondant sur la biogenèse des MIP. Comme le biais

de codon régule spécifiquement le processus de traduction, cela implique que l'impact de l'usage des codons sur la biogenèse des MIP opère au niveau de la traduction. Nous concluons que la biogenèse des MIP est régulée par le contexte de mRNA environnant la séquence codant pour le MIP, et ce, indépendamment de la séquence codant pour le MIP. Notre étude suggère que des algorithmes intégrant à la fois les événements traductionnels et post-traductionnels pourraient rendre possible la modélisation prédictive de l'immunopeptidome et l'identification de néoantigènes tumoraux basée sur des données transcriptomiques.

4.2 Contributions des auteurs

T.D : Conception de l'étude et développement de la méthodologie. Rédaction de la première ébauche du manuscrit. Rédaction et révision du manuscrit. Préparation des figures. Analyse et interprétation des données.

M.D.L : Analyse des données, préparations des figures.

P.T, Y. B : Rédaction et révision du manuscrit.

S.L, C.P : Supervision de l'étude. Rédaction et révision du manuscrit.

Soumis à *Cancer Immunology Research*, Juillet 2017.

4.3 Article

Codon usage regulates biogenesis of human MHC class I-associated peptides

Tariq Daouda[1,2], Maude Dumont-Lagacé[1,3], Pierre Thibault[1,4], Yoshua Bengio[5], Sébastien Lemieux[1,5,*], Claude Perreault[1,3,6,*]

[1]Institute for Research in Immunology and Cancer, Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

[2]Department of Biochemistry, Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

[3]Department of Medicine, Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

[4]Department of Chemistry, Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

[5]Department of Informatics and Operational Research, Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

[6]Canadian National Transplant Research Program, Université de Montréal, Q P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada

Corresponding Author: Claude Perreault, IRIC-Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, CANADA. Phone: 514-343-6126; Fax 514 343-7780; E-mail: claude.perreault@umontreal.ca

4.4 Abstract

MHC class I-associated peptides (MIPs) collectively form the immunopeptidome which defines the immune self for CD8 T lymphocytes. MIPs regulate the development and function of CD8 T cells, and represent the main targets of cancer immunosurveillance. Importantly, MIPs originate from selective regions of the genome. Indeed, while all proteins contain peptide sequences that could potentially bind to MHC molecules, most of these peptides never become MIPs. In order to understand why the immunopeptidome covers only a small fraction of the protein-coding transcriptome, studies have heretofore focused on post-translational events. Nonetheless, it has been shown that most MIPs originate from proteins that undergo proteasomal degradation co-translationally or in the minutes that follow translation. Using various bioinformatics methods, including artificial neural networks, we analyzed large datasets of transcripts coding for MIPs or not. We report that MIP biogenesis is regulated by codon usage in the mRNA regions flanking MIP-coding codons. Notably, we found that synonymous codons had non-redundant effects on MIP biogenesis. Since codon bias specifically regulates the translation process, the impact of codon usage on MIP biogenesis must operate at the translational level. We conclude that MIP biogenesis is regulated by the mRNA context of MIP-coding codons, independently of MIP-coding codons per se. Our work suggests that algorithms integrating both translational and post-translational events may enable predictive modeling of the immunopeptidome and identification of tumor neoantigens based on transcriptomic data.

4.5 Introduction

In jawed vertebrates, all nucleated cells present at their surface major histocompatibility complex (MHC) class I-associated peptides (MIPs), which are collectively referred to as the immunopeptidome (1,2). Recognition of abnormal MIPs is essential to the elimination of infected and neoplastic cells (3). Furthermore, self MIPs play a central role in shaping the adaptive immune system: they orchestrate the development of CD8 T cells in the thymus, as well as their survival and activation threshold in peripheral organs (4). Given the pervasive role of the immunopeptidome, systems-level understanding of its genesis and molecular composition is a central issue in immunobiology (5,6). These questions are particularly relevant to the identification of immunogenic antigens that can be targeted for cancer immunotherapy (3,7-13).

Recent high-throughput mass spectrometry analyses have revealed that MIPs originate from selected regions of the genome and that the immunopeptidome is not a random excerpt of the transcriptome or the proteome (1). Indeed, proteogenomic analyses of 25,270 MIPs isolated from the B lymphocytes of 18 individuals showed that 41% of expressed protein-coding genes generated no MIPs, while 59% of genes generated up to 64 MIPs/gene (14). The notion that the MIP repertoire presents only a small fraction of the protein-coding genome for monitoring by the immune system begs the question: what are the rules governing the molecular composition of the immunopeptidome? Relatedly, is it possible to predict which parts of the proteome will be presented by MHC-I molecules?

Genesis of the immunopeptidome depends on two main events: (a) the biogenesis (or "processing") of MIPs and (b) their binding to MHC-I molecules (15,16). The rules that regulate the second event, binding of MIPs to MHC-I, have been well defined by artificial neural

networks (ANN) (17-19). However, current tools are unable to predict which peptides will ultimately reach MHC molecules following a multistep processing in the cytosol and endoplasmic reticulum. Considering preferential sites of proteasome cleavage is useful to enrich for MIP candidates, but remains insufficient for MIP prediction, mostly because of prohibitive false discovery rates (10,20-22).

Most efforts at modeling MIP processing have focused on post-translational events (e.g., cleavage by proteases) and their regulation by the amino acid sequence of MIPs and their adjacent residues (10-mers at the N- and C-termini). However, a large body of evidence suggests that MIPs are produced during translation or a few minutes afterward (23). Indeed, many MIPs derive from defective ribosomal products (DRiPs), that is, peptides that fail to achieve a stable conformation during translation and are therefore rapidly degraded by the proteasome and other proteases. The genetic code being redundant, many (synonymous) codons are translated into the same amino acids. However, synonymous codons are not used in equal frequencies. This phenomenon, is termed codon-usage bias. Notably, the efficiency of protein synthesis heavily depends on codon usage (i.e which codons are used at specific positions in the mRNA sequence) (24,25). We therefore analyzed codon usage by genes that code for MIPs or not. In our effort to decipher the rules of MIP biogenesis, we mainly used ANNs because they provide a powerful array of methods to model non-linear interactions in large datasets (26). Although historically ANNs have been used essentially for their ability to make predictions, the fact that they can be trained to answer specific questions allows them to be used as powerful exploratory tools. The present work demonstrates that codon usage plays a significant role in MIP biogenesis, and that this effect must take place during translation.

4.6 Materials and Methods

Sequence extraction

This study was approved by the Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont, and informed consent was obtained from all subjects. Proteogenomic analyses on our subjects have been previously reported (14,27). Sequences were extracted using the Python package pyGeno (28) (version 1.2.8) with the human reference genome GRCh37.75.

Synonymous codon shuffling

For the KL analysis, each sequence was re-encoded by replacing each of codon by itself or by a random synonym according to usage frequency calculated on the sequence dataset (positive or negative). This transformation ensures that codon usage biases specific to positive and negative datasets are conserved. For ANN analyses, the same transformation was applied to sequences of both datasets (positive or negative). In this case, codons were replaced according to the human transcriptome usage frequencies provided by pyGeno. These frequencies were calculated *in silico* on transcript coding sequences using the annotations provided by Ensembl for the human reference genome GRCh37.75. This transformation erases all codon specific features from each dataset, while retaining amino acid features.

Statistics

Correlations and Fisher exact test results were computed using the R software. The spike selection algorithm has been implemented in-house in the Python programming language.

AUCs where computed using the Python package Sklearn (29). For ANN predictions on MIP-flanking sequences, transcript lengths where extracted using pyGeno on annotations provided by Ensembl for the human reference genome GRCh37.75.

ANN Sequence encoding and training

ANNs where trained on sequences resulting from the concatenation of pre- and post-MCC regions. Before presenting sequences to our ANNs, we associated each codon to a unique number ranging from 1 to 65 (we reserved 0 to indicate a null value) and used this encoding to transform every sequence into a vector of integers representing codons. Neural networks where built using the Python package Mariana (http://bioinfo.iric.ca/~daoudat/Mariana/). The *Embedding* layer of Mariana was used to associate each label superior to 0 to a set of 2D trainable parameters; the 0 label represents a *null* (masking) embedding fixed at coordinates (0,0). As an output layer, we used a *Softmax* layer with two outputs (positive / negative). Because negative sequences are more numerous than positive ones, we used an oversampling strategy during training. At each epoch, ANNs where randomly presented with the same number of positive and negative sequences.

We trained ten ANNs for each combination of conditions (context size x codon-shuffling x context availability), each one using a different random split of train/validation/test sets. We used an early stopping strategy on the validation sets to prevent over-fitting and reported average performances computed on test sets. To mask sequences either before or after the MCC, we masked either half with *null* value. Ten ANNs where trained for each condition (without pre-MCC context, without post-MCC context, with full context). All ANNs where trained using the

same train/validation/test split. For each sequence in the test set we calculated the average prediction score given by ANNs in each condition, and calculated the Pearson correlation using the R software. Densities where calculated on all points and drawn using ggplot2. Only a random subset of the points is represented in the figures to limit their size. All ANNs in this work share the same architecture, number of parameters and hyper-parameter values: learning rate: 0.001; mini-batch size: 64; embedding dimensions: 2; linear output without offset on the embedding layer; *Softmax* non-linearity without offset on the output layer.

Codon preferences

Preferences where obtained by feeding the ANN embedding vectors where all codons values were set to *null* (coordinates (0,0)), except for a single position that received a non-null codon label.

Data visualization and availability

All figures where generated using R's package ggplot2 and illustrator. Source code, documentation and use case examples for pyGeno are freely available from: https://github.com/tariqdaouda/pyGeno. Source code, documentation and use case examples for Mariana are freely available from: and https://github.com/tariqdaouda/Mariana. RNA-Seq data can be accessed on the NCBI Bioproject database (http://www.ncbi.nlm.nih.gov/bioproject/; accession PRJNA286122) and the mass spectrometry data can be found on the ProteomeXchange Consortium via the PRIDE partner repository (PXD004023).

4.7 Results

4.7.1 Codon affinity in MIP-source transcripts

Our dataset was constructed with MIPs presented by 33 HLA class I alleles on B lymphocytes from 18 subjects (14,27). From the entire datasets, we extracted the 19,656 nonamers with a predicted MHC binding affinity < 1,250 nM for at least one of the subject's MHC-I allotypes, according to NetMHC3.4 (30). We then used pyGeno (28) to extract the sequence of transcripts coding these 19,656 transcripts which constituted our positive dataset. We next created a negative (or decoy) dataset by randomly selecting 98,290 non-MIP nonamers from transcripts that generated no MIPs, and extracted their coding sequences using pyGeno as well. We reasoned that a transcript should be considered as a genuine positive or negative (regarding MIP biogenesis) only if it was expressed in the cells that were being studied. We therefore excluded from the datasets all transcripts whose expression was barely detectable (below the 99^{th} percentile in terms of FPKM). The resulting positive and negative datasets therefore contained the canonical reading frame of non-redundant MIP-source transcripts (n = 19,656) and non-source transcripts (n = 98,290), respectively (Fig. 1).

Codon usage bias regulates translation dynamics, and thereby affects translation efficiency, accuracy, and protein folding (31-34). To evaluate whether codon-anticodon affinity might influence MIP biogenesis, we compared the global usage of high affinity codons, as defined by Frenkel-Morgenstern *et al.* (31), between the 19,656 MIP-source transcripts and the 98,290 non-source transcripts. Transcript sequences were separated along their lengths in 100 bins of equal size. For each bin, we then calculated the frequency of high affinity codons for

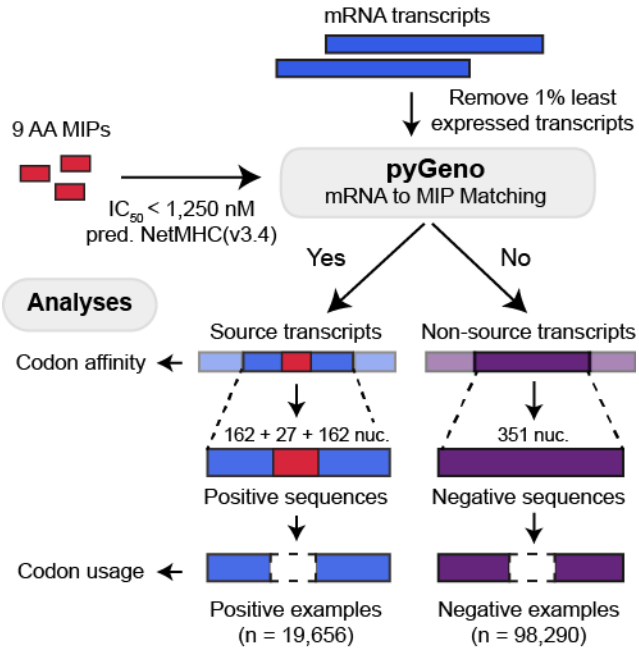


Figure 1. Construction of the dataset. Transcripts expressed in B cells from our subjects were considered as source or non-source transcripts depending on whether they were matched or not to at least one MIP. The entire length of source and non-source transcripts (from start to stop codon) was used for analyses of codon affinity (Fig. 2A). For other analyses of codon usage (Fig. 2B-D, 3 and 4), we focused our attention on the mRNA sequences flanking the nine MCCs (54 codons on each side of MCCs).

source and non-source transcripts (Fig. 2A). The two resulting distributions differed significantly at every position ($P < 10^{-16}$, Fisher exact test). The salient feature was that MIP-source transcripts contained a lower proportion of high affinity codons than non-source transcripts. The discrepancy between the two gene sets was particularly conspicuous on the 5'-side of the mRNAs, i.e., the initial 25% of the mRNA sequences. Usage of high affinity codons increased continuously when progressing from the 5'- to the 3'-end of MIP-source transcripts, but never reached the frequency found in non-source transcripts (Fig. 2A). The relatively low frequency of high affinity codons in MIP-source transcripts provides a plausible mechanistic

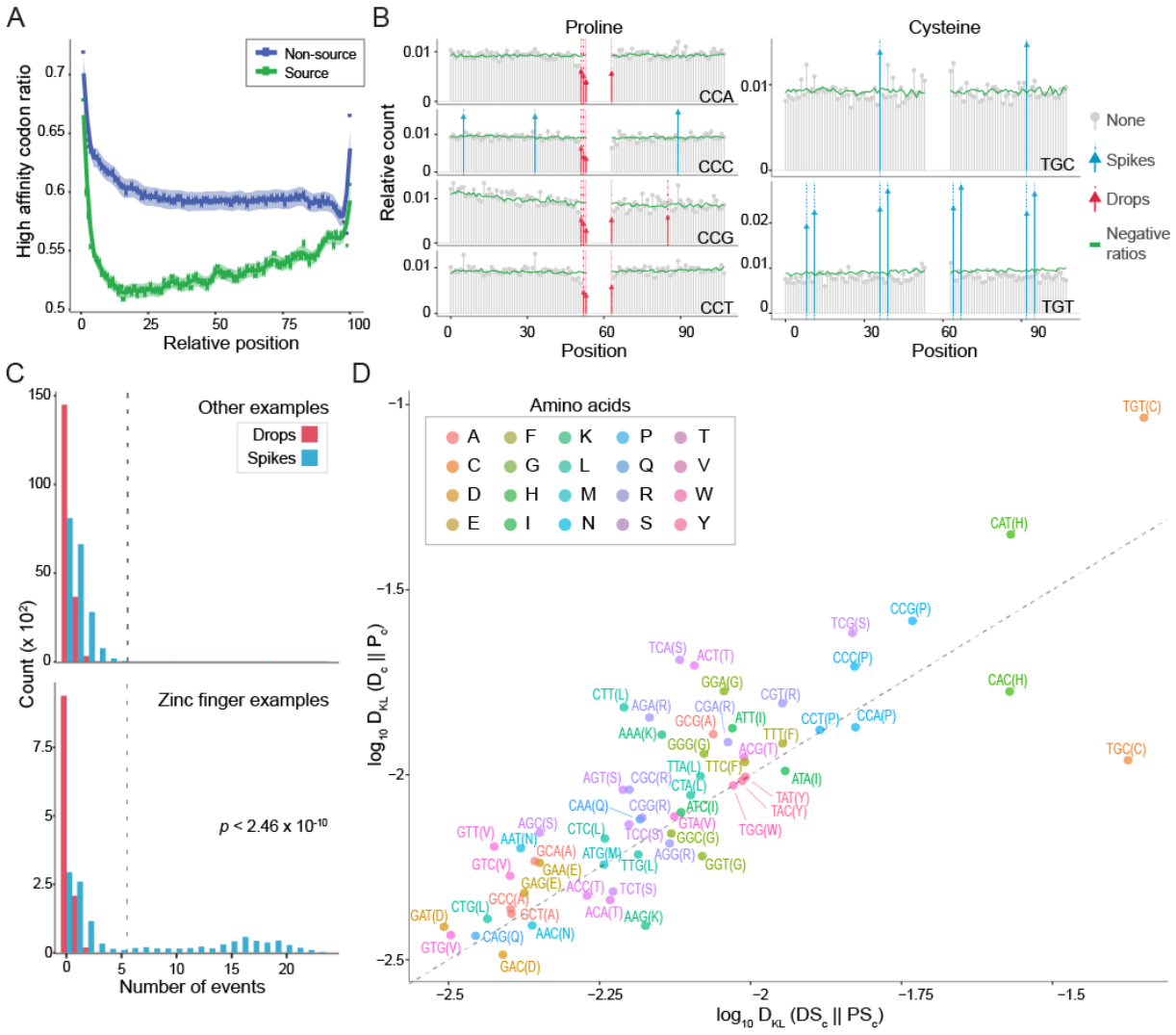


Figure 2. Codon usage in positive and negative datasets. (**A**) High-affinity codon usage with respect to normalized transcript length, areas around the curves represents 95% confidence intervals. (**B**) Distribution of proline and cysteine codons in positive (bars) and negative (green lines) datasets. The MCC is represented by the central blank space. (**C**) Distribution of spikes and drops in the positive dataset. Zinc finger protein transcripts showed the highest number of spikes. (**D**) KL divergences comparing positive and negative datasets. The $D_{KL}(D_c||P_c)$ axis shows the divergences between codon distributions in positive and negative datasets, the $D_{KL}(DS_c||PS_c)$ axis shows divergences after synonymous codon shuffling.

link between two seemingly unrelated reports: cell cycle-regulated genes prefer low affinity codons (31) and are a preferential source of MIPs (14).

4.7.2 Codon preferences in MIP-flanking region

For the next series of analyses, we focused our attention on the mRNA sequences flanking the nine MIP-coding codons (MCCs). We limited our analyses of flanking sequences to 162 nucleotides (54 codons) on each side of MCCs, because longer lengths would entail the exclusion of a significant proportion of transcripts (Fig. 3A). Because we were searching for features that might influence MIP generation rather than binding of MIP to MHC, we elected to analyze the MIP context rather than MCCs per se. We therefore removed the 9 central codons (i.e., the MCCs) from the positive and negative datasets and kept only the MCC-flanking sequences. We then compared codon frequency at each position of the MCC-flanking sequences from the positive and negative datasets. Results for proline and cysteine codons are shown in Fig. 2B, and all codons are presented in Supplementary Fig. S1 and S2. The specific question here was whether some codons were used at specific positions more (spike) or less (drop) frequently in the positive than in the negative dataset. We formally defined spikes and drops for each codon by fitting a Poisson distribution on 90% of the counts (codon usage), ignoring the 5% highest and lowest values (outliers). We excluded the outliers because they might have a disproportionate impact on the mean of the distributions. We then defined spikes and drops as positions with counts for a given codon belonging to the top or bottom 0.1% values of the Poisson distribution (corresponding to $P \leq 0.001$), respectively. Finally, to extract only the highest spikes and lowest drops, we only kept those corresponding to values respectively, 50%

higher or lower than the average count value of the codon distribution. We observed significant drops in the occurrence of the four proline codons immediately before and after the MCCs (Fig. 2B). Proline is the only amino acid that displayed such a strong decline around the MCCs across all its codons. It is also the only amino acid whose codons displayed drops both before and after the MCCs. Other codons with drops after the MCCs are ATA, ATC, ATT(I) and CTC, TTA, TTG(L) that showed one drop after the MCCs and TGG(W) that showed two (Supplementary Fig. S1 and S2). A different pattern emerged for cysteine codons which displayed periodic spikes that were both higher and more numerous for the TGT than the TGC codon. Spikes were also found in the codon distribution of 12 other amino acids (A, F, G, H, I, K, L, P, R, S, T, Y) (Supplementary Fig. S1 and S2). Interestingly, the only other amino acids for which spikes and drops were shared among all synonymous codons were phenylalanine (F) and tyrosine (Y). The fact that spikes and drops can be either shared (C, F, P, Y) or not (A, F, G, H, I, K, L, R, S) by synonymous codons suggests that both codon and amino acid choice influence MIP presentation.

To investigate genes contribution to spikes and drops, we extracted all the source sequences with at least one drop or one spike, and used the Hugo Gene Nomenclature Committee database (35) to search for potential overrepresentation of specific gene families in the positive dataset. We found that while transcripts coding for zinc finger proteins represented only 6% of our positive dataset, the zinc finger gene family was the only family with transcripts that contributed to at least 6 spikes (Fig. 2C), the maximum being 23. Of note, the distribution of drops was similar for zinc-finger transcripts and the rest of the positive dataset (Fig. 2C). Zinc finger-derived sequences represent 6% of the positive dataset (1,164 out of 19,656), and about 5% of the negative dataset (4,739 out of 98,290). Although small, the enrichment for zinc finger

transcripts in the positive dataset was very significant (odds ratio = 1.24, $P = 2.46 \times 10^{-10}$; Fisher exact test). Altogether, these results demonstrate that a specific subset of zinc finger proteins, defined by their use of specific codons at specific positions flanking the MCCs, constitutes a small but preferential source of MIPs.

4.7.3 Distribution of synonymous codons

To further validate the relative importance of codon vs. amino acid usage in MIP biogenesis, we asked the following question: how different are the codon and amino acid distributions in the positive and negative datasets? A higher divergence for codon distributions than for amino acid distributions would mean that codon variations are not entirely accounted for by amino acid variations. To address this question, we derived positive and negative datasets in which the original codons were replaced by synonymous codons according to their usage frequency in the datasets. We then defined the probability of having codon *c* at position *i* as a function of the number of occurrences of *c* at position *i*, divided by the total number of occurrences of that same codon:

$$Q_{(c,y,s)}(i) = \frac{N_{c,y,s}(i)}{\sum_j N_{c,y,s}\ (j)}$$

Here Q is a probability, N is a number of occurrences, c is a codon, y is a class (positive or negative), s indicates if codons have been randomized (true or false), i is a position in sequence. For the remainder of the text we will use the following abbreviations:

$$P_c(i) = Q_{c,y=positive,s=false}(i)$$

$$D_c(i) = Q_{c,y=negative,s=false}(i)$$

$$PS_c(i) = Q_{c,y=positive,s=true}(i)$$

$$DS_c(i) = Q_{c,y=negative,s=true}(i)$$

We then used the Kullback-Leibler (KL) divergence to compute how well P_c distributions approximate D_c distributions and PS_c distributions approximate DS_c distributions.

The KL divergence was defined as:

$$D_{KL}(P||Q) = \sum_i P(i)\log\left(\frac{P(i)}{Q(i)}\right)$$

Its value can be either positive or 0, a null value indicating that the two distributions are identical. KL divergence is not a metric, as it is neither symmetric nor does it satisfy the triangle inequality. It is nevertheless an accurate and most common way of comparing two probability distributions.

The random shuffling causes any codon specific features to be shared among synonyms, causing every codon distribution to reflect its amino acid distribution. If synonymous codons had equivalent distributions, the only observed variations would reflect some increase in the variance arising from splitting 20 amino acid distributions into 61 codon distributions. Therefore, values for $D_{KL}(D_c||P_c)$ would then be almost equal to values for $D_{KL}(DS_c||PS_c)$, and codons would cluster along the diagonal. However, the only codons on the diagonal were ATG(M), TGG(W) that have no synonyms, and TAT(Y), TAC(Y) (Fig. 2D) that have very similar distributions (Supplementary Fig. S1 and S2). This shows that codon distributions do

not mirror amino acid distributions. Moreover, of the 61 codons only 14 were below the diagonal while for 47 codons (77%), variations at the codon level were higher that variation at the amino acid level. Codons also did not cluster by amino acids along the $D_{KL}(D_c||P_c)$ diagonal showing that the level of divergence varied among synonymous codons. This shows that the breadth of synonymous codon variations cannot be explained by common amino acid features. Therefore, the variations observed when comparing positive and negative datasets at the codon level cannot be explained by variations at the amino acid level.

Notably, the codons showing the most divergent frequencies in the positive vs. negative datasets (i.e., located at the far-right in Fig. 2D) are coding for cysteine (C) and histidine (H). These two amino acids form the C2H2 motif found in the C2H2-ZF family of zinc finger proteins (36). This observation dovetails well with the significant enrichment in zinc finger proteins observed in the positive dataset (Fig. 2C). Taken together, these results show that local sequences adjacent to MIPs differ considerably from the rest of the transcriptome. On average, they tend to use less optimal codons and they bear distinct features to which zinc finger proteins are contributing to a significant extent. Our positional analyses of codon distributions indicate that, in most cases, differences between the positive and negative datasets reflect the specific impact of codons rather than of amino acids. Hence, our data unveil that codon bias in MIP flanking regions has a major role in MIP biogenesis.

4.7.4 Capturing codon bias using ANNs

To further assess the importance of codon usage in MIP biogenesis, we reasoned that if codons bear important information that is operative at the translational rather than the post-

translational level (e.g., during protein degradation), then: (i) ANNs trained to identify MCC-flanking regions should consistently perform better when trained on RNA sequences than on amino acid sequences, (ii) synonymous codons should have different effects on the prediction. To test these predictions, we designed a three-layer ANN depicted in Supplementary Fig. S3 using the machine learning framework Mariana (http://bioinfo.iric.ca/~daoudat/Mariana Fig./). The first (input) layer receives either MCC-flanking regions from the positive dataset or sequences of the same length contained in the negative dataset (Fig. 1, Supplementary Fig. S3). The second layer was a codon embedding layer similar to that introduced for neural language model (37). Embedding is a technique used in natural language processing to encode discrete words, and has been shown to greatly improve performances (26). In this technique, the user defines a fixed number of dimensions in which words should be encoded. When the training starts, each word receives a random vector-valued position (its embedding) in that space. The network then iteratively adjusts the words' embedding vectors during the training phase and arranges them in a way that optimizes the classification task. Notably, embeddings have been shown to represent semantic spaces in which words of similar meanings are arranged close to each other (26). In the present work, we treated codons as words: each codon received a set of random 2D coordinates that were subsequently optimized during training. The third (output) layer delivered the probability that the input sequence was an MCC-flanking region (rather than a sequence from the negative dataset).

To evaluate the consistency of our findings, we tested the performance of this architecture on several datasets corresponding to different lengths of flanking sequences (context sizes). The maximum context size that we used was 162 nucleotides (54 codons) on each side of the MCCs in the positive dataset and of central codons in the negative dataset.

Longer lengths would have excluded more than 25% of the transcripts from our datasets (Fig. 3A). For each context size, we randomly divided the positive and negative datasets into three subsets: the training subsets containing 60% of the positive and negative transcripts, as well as the test and validation subsets each containing 20% of the positive and negative transcripts. We used the transcripts of the training subsets to train our models and used the validation subsets to implement an early stopping strategy. The values for the area under the receiver operator characteristic curve (AUC) reported here were all obtained on the test subsets. These results show that increasing the context size has a positive effect on the performances, showing that codon usage regulates MIP presentation at different ranges (Fig. 3B). Performances on the training and validation subsets are presented in Supplementary Fig. S4.

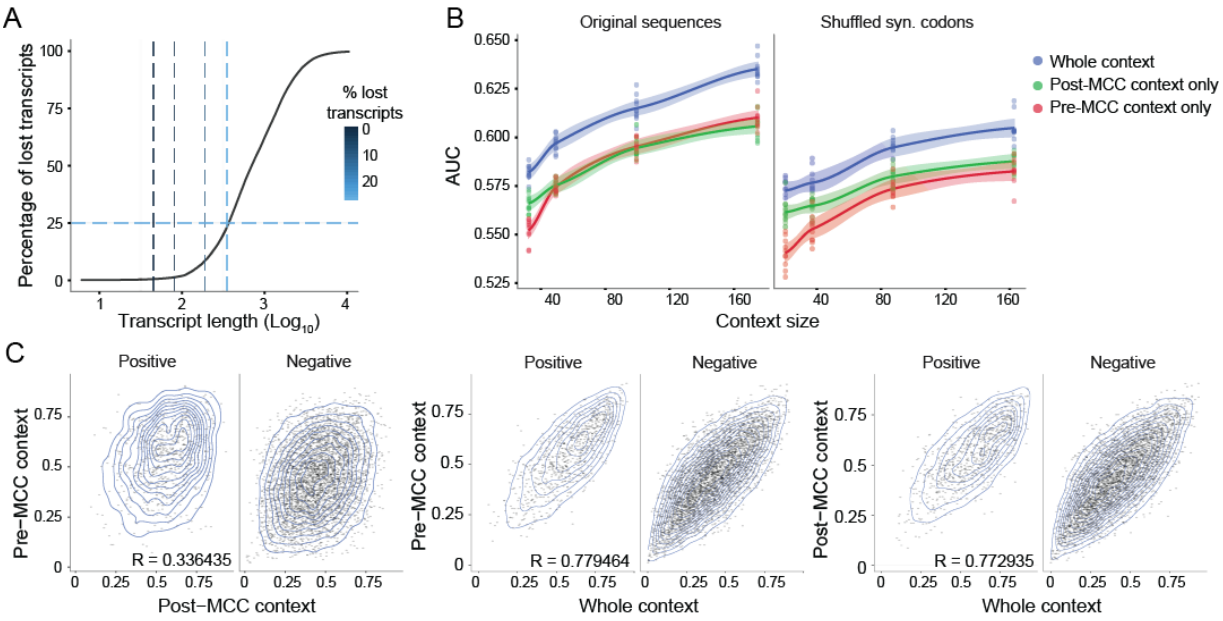


Figure 3. ANN predictions on MIP-flanking sequences. (**A**) Percentage of transcript ineligibility as a function of context size. Transcript length corresponds to *C x 2 + 27*, where *C* is the context size in nucleotides and 27 the length of the MCCs. (**B**) AUC for context sizes 9, 27, 81 and 162. Ten ANNs

where trained per condition, the areas around the curves represents 95% confidence intervals. Each point corresponds to a different ANN. (**C**) Correlation between ANN predictions, for a context size of 162 nucleotides: pre-MCC context vs the post-MCC context, pre-MCC context vs whole context, post-MCC context vs whole context. Blue lines represent 2D densities.

To remove codon specific information, we shuffled synonymous codons according to their frequencies in the human genome. We applied the transformation in the same way for the positive and negative datasets. We observed that predictions were consistently better when the ANNs received the original codons (Fig. 3B, left) than when they received shuffled synonymous codons (Fig. 3B, right). This further supports the concept that MIP biogenesis is regulated by the RNA sequences flanking MCCs. To evaluate whether any part of the context was particularly important to the prediction, we trained ANNs with either the mRNA sequence preceding or following the MCCs (red and green lines in Fig. 3B). In both cases performances suffered (Figure 3B). When comparing the predictions given by models trained with only the pre-MCC context to those trained with the post-MCC context, we noted that these predictions were weakly correlated ($r = 0.33$) (Fig. 3C). However, when we compared the predictions of either model to those obtained when training on full contextual sequences, the correlations were much higher ($r = 0.77$). Collectively, these data suggest that both contexts (pre- and post-MCCs) bear important but distinctive features relevant to MIP presentation.

4.7.5 ANNs unveil positional codon preferences

ANNs still carry the reputation of being undecipherable black boxes. It is true that the interpretation of the inner structures of deep ANNs is still in its infancy. On the other hand, simpler architectures, such as the one used herein, can be more easily probed to yield useful

information about the way predictions are being made. Indeed, a trained ANN remains a fixed set of mathematical transformations that can be studied, analyzed and, in theory, interpreted. In order to assess the effect of individual codons on the overall prediction, we therefore presented a single codon at a single position to the best model trained with a context size of 162 nucleotides. By running this setup for every codon at every position, while monitoring the prediction, we isolated the model preferences for individual codons (Fig. 4A,B). In other words, the probabilities retrieved when a specific codon is present at a specific position. A value of 0.5 denotes a neutral preference, while negative and positive preferences correspond to values below and above 0.5, respectively. Preferences for all codons are available in Supplementary Fig. S4.

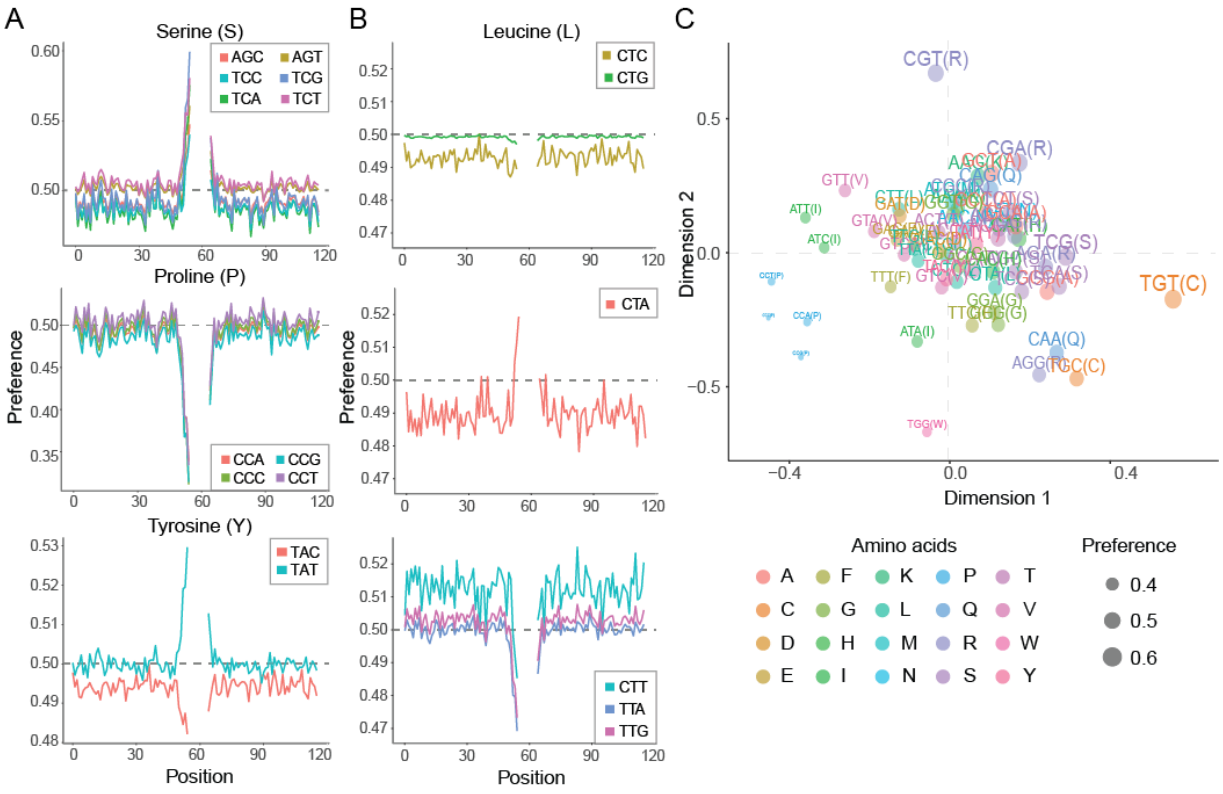


Figure 4. ANN interpretation of codon impact on MIP biogenesis. Preferences for a network trained on a context of 162 nucleotides for (**A**) serine, proline and tyrosine codons, (**B**) leucine codons.

(**C**) learned codon embeddings and preferences at the position directly preceding the MCCs. Proline codons are the only synonyms that form a conspicuous cluster. As indicated by the size of the dots, codons on the right-hand side increase the probability of the sequence being classified as source, whereas codons on the left-hand side of the graph have the opposite effect.

While codons at all positions contributed to the prediction, the most influential were those located around 4-5 positions before or 2-3 positions after the MCC. The presence of specific codons at those positions greatly increased (e.g. Serine codons) or decreased (e.g. Proline codons) the model's output probability (Fig. 4A). In this narrow region, preferences exhibited a strong symmetry centered around the MCCs, where an increase in preference before the MCCs was always matched with an increase after the MCCs and vice-versa. Interestingly, when located in the close vicinity of MCCs, prolines have been shown to decrease MIP biogenesis by preventing proteasomal cleavage (38), which is reflected by the lower preferences for all proline codons around the MCC. In other cases, we observed that synonymous codons had divergent impacts. Indeed, the ANN favored one tyrosine codon (TAT) but disfavored the other (TAC) (Fig. 4A, lower panel). The situation was even more complex for leucine, as two codons were considered neutral, whereas one was favored and three were disfavored by the ANN (Fig. 4B). While the ANN showed similar preference for several synonymous codons, the preference magnitude showed major discrepancies among synonymous codons. Examples of codons showing much higher variations than their synonyms are TGT for cysteine, GAT for aspartic acid, TTT for phenylalanine, CAT for histidine, AAG for lysine, AAT for asparagine, and ACG for threonine (Supplementary Fig. S4).

Using embeddings to encode codons has the advantage of arranging them into a semantic space, wherein codons with similar influences are positioned close to each other. Fig. 4C shows the resulting semantic space as well as the preferences for every codon for the position directly

preceding the MCCs. Conspicuously, synonymous codons do not form clusters, a notable exception being proline codons. This shows that the effect of a given codon can be closer to that of a non-synonymous codon than to that of a synonym. The evolution of preferences on the embedding space for every position in the sequence is depicted in Supplementary Movie S1. Altogether, these results further support our conclusion that codon choice plays a determining role in MIP biogenesis. Overall, the impact of codon usage on MIP biogenesis was observed when we analyzed the entire transcript (Fig. 2A), the 54 codons on either side of the MCCs (Fig. 3) and the 5 codons adjacent to the MCCs (Fig. 4). We therefore conclude that codon usage has short, medium and long-range effects on MIP processing.

4.8 Discussion

Each HLA allotype presents no more than 0.1% of the potential 9-mer peptides from human protein-coding genes (21). A recent report showed that the entire MIP repertoire presented by 27 HLA allotypes covered only 10% of the exomic sequences expressed in B lymphocytes (14). In line with this, less than 1% of expressed tumor mutations generate immunogenic MIPs (39). The need for peptides to be strong MHC binders in order to become MIPs severely constrains the diversity of the MIP repertoire. However, MHC binding is not the sole limiting factor. Indeed, while practically all proteins contain peptides that would be strong MHC binders (40), about 40% of proteins generate no MIPs while other proteins generate up to 64 MIPs/gene (14). Hence, some proteins are good sources of MIPs while others are not. Therefore, events that precede MHC binding must have a determinant influence on biogenesis of the immunopeptidome. Efforts to decipher the rules of MIP processing have heretofore focused on various post-translational events: cleavage by the proteasome and other proteases, binding to proteins such as TAP and ERAAP (15). However, seminal studies have demonstrated that MIP biogenesis is clearly regulated at the translation level, and that most MIPs originate from proteins that undergo proteasomal degradation co-translationally or in the minutes that follow translation (41). This pool of rapidly degraded proteins includes a large proportion of DRiPs that arise from errors in protein translation or folding.

Because codon usage regulates translation accuracy, efficiency and co-translational protein folding, we investigated whether codon choice might regulate MIP biogenesis. Our analyses of large datasets using diverse bioinformatics approaches provides compelling evidence that codon usage regulates MIP biogenesis via both short- and long-range effects. Over

their entire length, MIP-source transcripts use more low affinity codons than the rest of the transcriptome (Fig. 1A). More in-depth analyses of the 54 codons on each side of the MCCs revealed differential codon usage in the MCC flanking regions (Fig. 2 and 4). The salient finding was that in most cases, synonymous codons showed different usage distribution in MCC flanking regions compared to the rest of the transcriptome.

Our study also illustrates that ANNs can be used not only for prediction but also to extract relevant biological features from large datasets, and thereby provide mechanistic insights in such complex processes. Here we elected to use embeddings because their capacity to represent discrete inputs into an interpretable latent continuous space makes them especially well-suited for codon analysis. Three main points can be made from the performance of ANNs trained to discriminate between MCC flanking regions and regions randomly extracted from the transcriptome. First, the better prediction accuracy of ANNs trained with original codons rather than with shuffled synonyms supports the critical role of codon usage in MIP genesis (Fig. 3). Second, the interpretation of ANNs output and inner structure show that while positions distant from as much as 54 codons of the MCCs influence the prediction (Fig. 3B), positions directly adjacent to the MCCs disproportionately influence the output (Fig. 4). Third, synonymous codons have different effects on the prediction (Fig. 4). Thus, in codons adjacent to the MCCs, tyrosine codon TAT increased the probability of the sequence being classified as source, while TAC decreased it (Fig. 4A). Two inferences can be made from these results: synonymous codons have non-redundant effects on MIP biogenesis, and MIP biogenesis must be regulated, at least in part, during translation.

We have limited our studies to the most common type of MIPs: nonamers coded by the canonical reading frame of classic protein-coding genes (42). Further analyses of large datasets

will be needed to assess the full extent of codon usage on both classic MIPs, and MIPs derived from non-canonical reading frames (43). Likewise, further studies will be required to understand exactly how codon usage regulates MIP biogenesis. One attractive possibility would be that codon bias regulates the formation of DRiPs. A corollary of this work is that CD8 T cells are preferentially confronted to peptides encoded by transcripts with a codon bias. Whether this preference is biologically relevant to immunosurveillance against pathogens or transformed cells has yet to be explored.

A more direct implication of this work is that integrating both translational and post-translational events in predictive algorithms may greatly enhance the predictive modeling of the immunopeptidome. Particularly in the field of precision cancer immunotherapy, discovering suitable target antigens is a formidable challenge given the limitations inherent to all methods tested to date (10,22). Since RNA sequencing requires only a small number of primary tumor cells, predicting the immunopeptidome directly from transcriptomic data would have a transformative impact on the development of personalized cancer vaccines (7,8,10-13).

4.9 References

1. Granados DP, Laumont CM, Thibault P, Perreault C. The nature of self for T cells-a systems-level perspective. Curr Opin Immunol 2015;34:1-8

2. Caron E, Espona L, Kowalewski DJ, Schuster H, Ternette N, Alpizar A, et al. An open-source computational and data resource to analyze digital maps of immunopeptidomes. Elife 2015;4

3. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science 2015;348:69-74

4. Davis MM, Krogsgaard M, Huse M, Huppa J, Lillemeier BF, Li QJ. T cells as a self-referential, sensory organ. Annu Rev Immunol 2007;25:681-95

5. Caron E, Vincent K, Fortier MH, Laverdure JP, Bramoulle A, Hardy MP, et al. The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. Mol Syst Biol 2011;7:533

6. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. Mol Cell Proteomics 2015;14:3105-17

7. Coulie PG, Van den Eynde BJ, van der Bruggen P, Boon T. Tumour antigens recognized by T lymphocytes: at the core of cancer immunotherapy. Nat Rev Cancer 2014;14:135-46

8. Melief CJ, van Hall T, Arens R, Ossendorp F, van der Burg SH. Therapeutic cancer vaccines. J Clin Invest 2015;125:3401-12

9. Romero P, Banchereau J, Bhardwaj N, Cockett M, Disis ML, Dranoff G, et al. The Human Vaccines Project: A roadmap for cancer vaccine development. Sci Transl Med 2016;8:334ps9

10. Yarchoan M, Johnson BA, 3rd, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. Nat Rev Cancer 2017;17:209-22

11. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Lower M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature 2017;547:222-6

12. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 2017;547:217-21

13. Melief CJM. Cancer: Precision T-cell therapy targets tumours. Nature 2017;547:165-7

14. Pearson H, Daouda T, Granados DP, Durette C, Bonneil E, Courcelles M, et al. MHC class I-associated peptides derive from selective regions of the human genome. J Clin Invest 2016;126:4690-701

15. Vigneron N, Van den Eynde BJ. Insights into the processing of MHC class I ligands gained from the study of human tumor epitopes. Cell Mol Life Sci 2011;68:1503-20

16. Rammensee HG, Falk K, Rotzschke O. Peptides naturally presented by MHC class I molecules. Annu Rev Immunol 1993;11:213-44

17. Bassani-Sternberg M, Gfeller D. Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions. J Immunol 2016;197:2492-9

18. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med 2016;8:33

19. Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, et al. Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res 2008;36:W513-8

20. Nielsen M, Lundegaard C, Lund O, Kesmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics 2005;57:33-41

21. Abelin JG, Keskin DB, Sarkizova S, Hartigan CR, Zhang W, Sidney J, et al. Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity 2017;46:315-26

22. Capietto AH, Jhunjhunwala S, Delamarre L. Characterizing neoantigens for personalized cancer immunotherapy. Curr Opin Immunol 2017;46:58-65

23. Anton LC, Yewdell JW. Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. J Leukoc Biol 2014;95:551-62

24. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, Roth AC, et al. A role for codon order in translation dynamics. Cell 2010;141:355-67

25. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 2011;12:32-42

26. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44

27. Granados DP, Rodenbrock A, Laverdure JP, Cote C, Caron-Lizotte O, Carli C, et al. Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers. Leukemia 2016;30:1344-54

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

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12:2825-30

30. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 2008;36:W509-12

31. Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, Hou YM, et al. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Mol Syst Biol 2012;8:572

32. Yu CH, Dang Y, Zhou Z, Wu C, Zhao F, Sachs MS, et al. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol Cell 2015;59:744-54

33. Saunders R, Deane CM. Synonymous codon usage influences the local protein structure observed. Nucleic Acids Res 2010;38:6719-28

34. Zhao F, Yu CH, Liu Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res 2017

35. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res 2015;43:D1079-85

36. Pabo CO, Peisach E, Grant RA. Design and selection of novel Cys2His2 zinc finger proteins. Annu Rev Biochem 2001;70:313-40

37. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. J Mach Learn Res 2003;3:1137-55

38. Shimbara N, Ogawa K, Hidaka Y, Nakajima H, Yamasaki N, Niwa S, et al. Contribution of proline residue for efficient production of MHC class I ligands by proteasomes. J Biol Chem 1998;273:23062-71

39. Yadav M, Delamarre L. IMMUNOTHERAPY. Outsourcing the immune response to cancer. Science 2016;352:1275-6

40. Hoof I, van Baarle D, Hildebrand WH, Kesmir C. Proteome sampling by the HLA class I antigen processing pathway. PLoS Comput Biol 2012;8:e1002517

41. Dolan BP, Bennink JR, Yewdell JW. Translating DRiPs: progress in understanding viral and cellular sources of MHC class I peptide ligands. Cell Mol Life Sci 2011;68:1481-9

42. Trolle T, McMurtrey CP, Sidney J, Bardet W, Osborn SC, Kaever T, et al. The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. J Immunol 2016;196:1480-7

43. Laumont CM, Daouda T, Laverdure JP, Bonneil E, Caron-Lizotte O, Hardy MP, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat Commun 2016;7:10238

4.10 Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

4.11 Authors' Contributions

Conception and design: T. Daouda

Development of methodology: T. Daouda

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): T. Daouda, M. Dumont-Lagacé

Writing, review, and/or revision of the manuscript: T. Daouda, M. Dumont-Lagacé, P. Thibault, Y. Bengio, S. Lemieux, C. Perreault

Study supervision: S. Lemieux, C. Perreault

4.12 Grant support

This work was supported the Canadian Cancer Society (Grant number 701564 to C. Perreault) and the Quebec Breast Cancer Foundation (Major Research Grant to C. Perreault). M. Dumont-Lagacé. is supported by a studentship from the Canadian Institute of Health Research. Y. Bengio, P. Thibault and C. Perreault hold Canada Research Chairs.

4.13 Footnotes

Note: Supplementary data for this article include 4 supplementary figures and 1 movie.

4.14 Supplementary Informations

Les informations supplémentaires peuvent être retrouvées en annexe 3.

Conclusion

Dans ce travail, nous avons tous d'abord introduit de nouvelles méthodes d'identification des MIP par MS. Ces méthodes, spécifiques au génome du sujet, nous ont permis d'augmenter grandement la précision de l'identification, ce qui nous a permis de découvrir 34 nouveaux MiHAs. Ce travail fondateur en proteogénomique à introduit les bases de la collaboration entre immunologie, protéomique et génomique à l'origine de travaux ultérieures[5,12,20]. Nous avons ensuite introduit pyGeno, une base de données optimisée pour la création et l'exploration de génomes personnalisés.

Finalement, en utilisant un ensemble de données construit à l'aide des méthodes présentées dans les chapitres 2 et 3, nous démontrons que l'usage des codons influe sur la présentation de MIP. Plus spécifiquement, nous avons montré que différents codons ont une influence différente sur la présentation (même s'ils sont synonymes) et que cette influence dépend de la position du codon par rapport à l'ARN codant pour le MIP.

Les règles de présentation de MIP découlant de l'usage des codons que nous avons extrait à l'aide d'ANN sont indépendantes de l'allèle de MHC-I du sujet, comme le montre la Figure 1A. Ceci est confirmé par le fait que la corrélation entre l'affinité prédite de la séquence codant pour le MIP et la prédiction retournée par notre ANN est très basse (Figure 1A). Plus intéressant encore, un réseau entrainé sur des séquences d'ARN issues de séquences alignées de lymphoblastes B humains est capable de prédire des MIP murins à partir de lectures brutes (*reads* en anglais) du séquenceur. Comme le montre la Figure 1B, les distributions de prédictions sont très similaires entre l'humain et la souris. Les performances chez la souris sont néanmoins moins bonnes. Ceci pourrait être causé par des différences entre espèces et types cellulaires, ou bien du fait que les séquences de souris proviennent ici de lectures brutes du séquenceur, alors que notre réseau a été entrainé sur des séquences alignées, et donc filtrées. Néanmoins, ces résultats suggèrent fortement que le réseau a identifié des règles fortes de présentation des MIP conservées à travers les espèces.

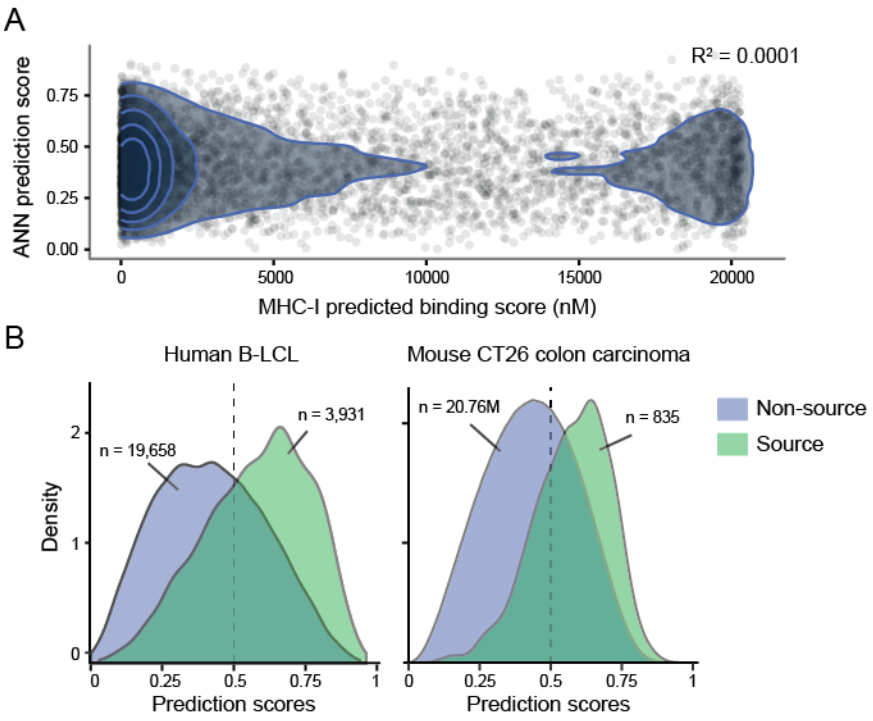


Figure 1. Les règles d'usage des codons sont indépendantes de l'affinité aux allèles de MHC-I. (A) Corrélations entre prédiction du réseau et affinité prédite de la séquence centrale pour les allèles MHC-I des sujets. La corrélation est très faible, indiquant que le règles d'usage de codons extraites par le réseau sont indépendantes de l'affinité au MHC-I de la séquence prédite comme codant pour un MIP. (B) Comparaison entre les sorties d'un réseau entrainé sur des cellules humaines B-LCL, sur : 1) l'ensemble de test, composé de séquences humaines de la même origine que les séquences de l'ensemble d'entrainement, mais non présentées durant l'apprentissage (à gauche) et 2) des lectures brutes non alignées d'un séquençage de cellule CT26 (cancer colorectal murin, à droite). La courbe verte représente la distribution des prédictions pour les séquences encadrant des MIP identifiés par MS. La courbe bleue représente la distribution des prédictions pour toutes les autres séquences. L'allure des courbes est similaire chez l'humain et la souris, suggérant que les règles identifiées par l'ANN sont conservées à travers les espèces et les types cellulaires.

Une implication majeure de notre découverte est qu'elle suggère fortement l'existence d'un ensemble de MIP candidats identifiables selon des règles conservées à travers les espèces (Figure 2). À partir de cet ensemble de candidats, les MIP présentés sont ensuite sélectionnés à partir de facteurs tels que leur affinité de liaison aux allèles MHC-I. Cet ensemble de MIP

candidats est particulièrement intéressant puisqu'il apparait avant le filtrage par les allèles du MHC-I, et constitue donc un ensemble de prédilection pour la recherche de MiHA humains. D'après les résultats de Pearson et al.[20], 10% des exons seraient à l'origine de MIP. Ceci suggére que l'ensemble des MIP candidats couvriraient au moins 10% des régions codantes.

Nos résultats suggèrent aussi fortement qu'incorporer de l'information sur les séquences d'ARN source améliorerait les capacités de prédictions de MIP, et par conséquent des MiHA et TSA. Cette capacité à prédire les MiHA et TSA directement à partir de séquences d'ARN accélérerait grandement la recherche en immunothérapie du cancer et rendrait possible le développement de thérapies cliniques hautement personnalisées. De plus, étant donné l'importance primordiale de l'immunopeptidome pour le système immunitaire adaptatif, l'existence de MIP candidats identifiables à travers les espèces grâce aux règles d'usage de codons pourrait avoir de profondes implications pour notre compréhension de phénomènes aussi importants que les maladies auto-immunes, l'évasion virale et la réponse immunitaire face au cancer.

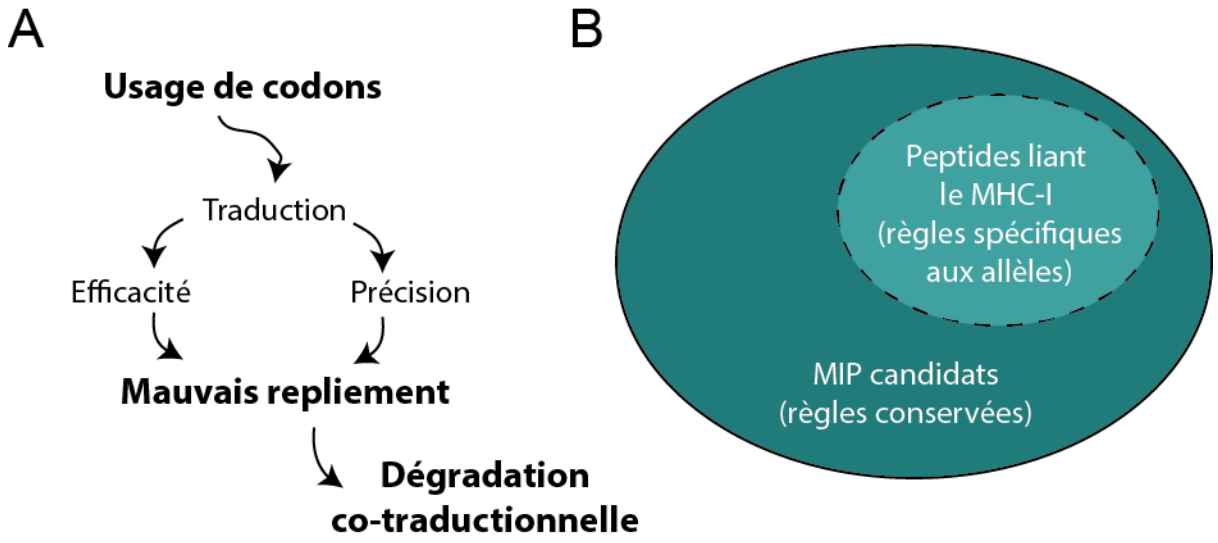


Figure 2. Théorie sur les mécanismes de l'influence de l'usage des codons sur la présentation des MIP. (A) L'usage de certains codons influence négativement la précision et l'efficacité de traduction, ce qui amène à des erreurs de repliement et à une dégradation rapide de la chaine naissante. (B) Nos résultats suggèrent fortement l'existence d'un ensemble de MIP candidats au sein desquels les MIP présentés sont sélectionnés en fonction de leur affinité aux molécules du MHC-I. Si cette théorie est confirmée, elle pourrait avoir un impact majeur sur notre compréhension de la genèse

des MIP, ainsi que sur leur prédictibilité par des moyens algorithmiques. Ces retombées influenceraient l'immunothérapie du cancer, la conception de vaccin antiviraux et la compréhension de maladies auto-immunes.

Ces règles de présentation sont-elles exactement les même pour tous les types cellulaires? L'expression des ARNt varie selon le type cellulaire[54,55], se pourrait-il que des dérèglements physiologiques localisés puissent modifier l'ensemble des candidats dans un type cellulaire seulement? Ces cellules présenteraient alors des MIP que les cellules de la médulla thymique seraient incapables de présenter. Ces MIP pourraient alors être reconnus par des lymphocytes T.

Les virus adaptent leur usage de codons à celui de leur hôte[56,57]. Serait-il possible que cette adaptation soit, au moins en partie, régie par ces règles de présentation de MIP? L'ensemble des MIP candidats est-il modifié par une infection virale? Étudier ces questions permettrai de mieux identifier les pressions sélectives à l'origine de l'évolution des virus et de d'avoir une meilleure compréhension de comment certains virus arrivent à déjouer le système immunitaire.

Il serait également intéressant de découvrir comment cet ensemble de candidats se retrouve modifié dans les cellules infectées et cancéreuses. Il a été montré que l'expression des ARNt est modifiée dans les cellules du cancer du sein[58,59]. Se pourrait-il que des mutations génétiques ou physiologiques apparaissant dans des cellules cancéreuses soient capables de modifier cet ensemble de candidats? A ce moment-là, ces cellules présenteraient un sous-ensemble de MIP non présents chez les cellules saines. Le plus intéressant est que les MIP en question pourraient provenir de régions non mutées dans le génome, mais qui ne sont pas source de MIP chez les cellules saines. Puisque ces MIP ne seraient alors présentés que par les cellules cancéreuses, il est raisonnable de suspecter que leur immunogénicité, et donc leur potentiel thérapeutique, pourrait être semblable à celui des TSA classiques.

Ici, nous ne nous sommes intéressés qu'aux peptides provenant de régions traduites dans le cadre de lecture canonique de transcrits exprimés. Les mêmes règles d'usage de codons s'appliquent-elles aussi à la présentation de MIP d'origine cryptique? Malgré le fait que le réseau montre un comportement similaire sur des séquences brutes et des séquences alignées (Figure 1B), nous ne pouvons trancher définitivement sur cette question. Étant donné le fait que

les MIP cryptiques peuvent être immunogéniques, et donc fournir des cibles d'immunothérapie[5], il serait intéressant d'investiguer si l'usage des codons influe similairement sur la présentation de MIP cryptiques.

En 2016, nous avons introduit une nouvelle méthode d'identification par MS permettant le séquençage de MIP cryptiques (article disponible en annexe 4)[5]. Les MIP cryptiques peuvent provenir de diverses régions non-codantes : d'introns, de régions inter-géniques, ou de jonctions entre exons et introns. L'alignement de lectures du séquenceur est une étape particulièrement complexe, durant laquelle les lectures issues du séquenceur peuvent être éliminées par l'algorithme d'alignement. Au lieu de travailler sur un génome aligné, nous avons élaboré une méthode qui utilise directement les lectures issues du séquenceur. Chaque lecture est traduite en plusieurs MIP potentiels, dans les 6 cadres de lecture possibles. Afin de diminuer le nombre faux positifs et d'éliminer les traductions résultant d'erreurs de séquençage, les séquences traduites sont ensuite filtrées pour ne garder que celles qui ont été vues un nombre minimal de fois. Ces séquences filtrées sont ensuite regroupées dans une base de données pour le spectromètre de masse. En utilisant cette méthode, nous avons été capables d'identifier 168 MIP d'origine non canonique chez un seul sujet. Les MIP cryptiques identifiés de cette façon présentent des enjeux particuliers puisqu'ils sont supportés par des lectures de séquenceurs (75-150 paires de bases) et non par des transcrits complets. La petite taille des séquences implique nécessairement que la taille du contexte disponible autour du MIP est plus restreinte. Néanmoins, il serait possible d'utiliser des techniques d'alignement *de novo* afin de rassembler les lectures codant pour les MIP en des séquences plus longues. Il serait alors possible d'utiliser les résultats issus de cette méthode d'identification pour déterminer si les règles qui régissent la présentation des MIP canoniques s'appliquent également aux MIP cryptiques.

Finalement, les réseaux de neurones artificiels utilisés dans notre dernière étude ont été conçus pour démontrer l'importance des codons dans le phénomène de présentation des MIP. Par conséquent, ces réseaux sont minimalistes afin d'en faciliter l'interprétation. Dans le but d'améliorer les performances de prédiction, des réseaux de neurones récurrents bidirectionnels, tels que ceux utilisés pour la transformation de séquences textuelles[48] seraient sans doute beaucoup plus adaptés à la réalité biologique. Ces réseaux ont non seulement plus de capacité

que ceux que nous avons utilisés, ils ont aussi l'avantage d'être capables de traiter des séquences de longueur variable.

Qu'est-ce qui expliquerait l'influence des codons sur la présentation des MIP? Notre hypothèse favorite est que ces MIP sont le résultat de la dégradation co-traductionnelle engendrée par les diverses influences des codons, discuté dans l'introduction.

Notre étude illustre également comment des réseaux de neurones artificiels peuvent, au-delà de la construction de prédicteurs, servir à extraire de l'information pertinente d'ensembles de données biologiques. Le fait qu'un réseau soit capable de prédire un phénomène à partir d'un autre, indique l'existence d'une relation de corrélation entre les données présentées en entrée, et celles prédites en sortie. L'analyse des structures internes permet ensuite de découvrir les règles qui permettent au réseau d'effectuer sa prédiction. Finalement, ces règles à leur tour permettent d'élaborer des théories sur les mécanismes biologiques qui expliquerai la relation observée. Ces dernières années ont vu à la fois une augmentation importante des données biologiques standardisées disponibles[60–62], ainsi que des avancées importantes en ce qui concerne l'interprétabilité des réseaux profonds, tel que des mécanismes d'attention qui permettent d'identifier les régions les plus importantes pour l'inférence [48]. La convergence de ces deux disciplines (bio-informatique et apprentissage machine), crée un terrain de plus en plus favorable à l'élaboration de méthodes d'analyses de données biologiques assistées par des techniques d'intelligence artificielle de pointe.

Références

1. Paham P. *The Immune System*. Vol 3rd ed. Garland Science; 2009.

2. Coley W. Immunotherapy What is immunotherapy ? *Am Cancer Soc*. 2011:1-31. www.cancer.org.

3. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science (80-)*. 2015;348(6230):69-74. doi:10.1126/science.aaa4971.

4. Caron E, Vincent K, Fortier MH, et al. The MHC i immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol Syst Biol*. 2011;7. doi:10.1038/msb.2011.68.

5. Laumont CM, Daouda T, Laverdure J, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun*. 2016;7:10238. doi:10.1038/ncomms10238.

6. Granados DP aola, Sriranganadane D, Daouda T, et al. Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat Commun*. 2014;5:3600. doi:10.1038/ncomms4600.

7. Mathis D, Benoist C. Aire. *Annu Rev Immunol*. 2009;27(1):287-312. doi:10.1146/annurev.immunol.25.022106.141532.

8. Eldershaw SA, Sansom DM, Narendran P. Expression and function of the autoimmune regulator (Aire) gene in non-thymic tissue. *Clin Exp Immunol*. 2011;163(3):296-308. doi:10.1111/j.1365-2249.2010.04316.x.

9. St-Pierre C, Trofimov A, Brochu S, Lemieux S, Perreault C. Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells. *J Immunol*. 2015;195(2):498-506. doi:10.4049/jimmunol.1500558.

10. Vincent K, Roy DC, Perreault C. Next-generation leukemia immunotherapy. *Blood*. 2011;118(11):2951-2959. doi:10.1182/blood-2011-04-350868.

11. Vincent K, Hardy MP, Trofimov A, et al. Rejection of Leukemic Cells Requires Antigen-Specific TCells with High Functional Avidity. *Biol Blood Marrow Transplant*. 2014;20(1):37-45. doi:10.1016/j.bbmt.2013.10.020.

12. Granados DP, Rodenbrock a, Laverdure J-P, et al. Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers. *Leukemia*. 2016;(October 2015):1344-1354. doi:10.1038/leu.2016.22.

13. Ren SC, Qu M, Sun YH. Investigating intratumour heterogeneity by single-cell sequencing. *Asian J Androl*. 2013;15(6):729-734. doi:10.1038/aja.2013.106.

14. Seoane J, De Mattos-Arruda L. The challenge of intratumour heterogeneity in precision medicine. *J Intern Med*. 2014;276(1):41-51. doi:10.1111/joim.12240.

15. Waclaw B, Bozic I, Pittman ME, et al. Intratumour Heterogeneity. *Nature*. 2016;525(7568):261-264. doi:10.1038/nature14971.A.

16. Mittal D, Gubin MM, Schreiber RD, Smyth MJ. New insights into cancer immunoediting and its three component phases—elimination, equilibrium and escape. *Curr Opin Immunol*. 2014;27:16-25. doi:10.1016/j.coi.2014.01.004.

17. Granados DP, Laumont CM, Thibault P, Perreault C. The nature of self for T cells-a systems-level perspective. *Curr Opin Immunol*. 2015;34:1-8. doi:10.1016/j.coi.2014.10.012.

18. Backert L, Kohlbacher O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*. 2015;7(1):119. doi:10.1186/s13073-015-0245-0.

19. Granados DP, Yahyaoui W, Laumont CM, et al. MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements. *Blood*. 2012;119(26):e181-91. doi:10.1182/blood-2012-02-412593.

20. Pearson H, Daouda T, Granados DP, et al. MHC class I-associated peptides derive from selective regions of the human genome. *J Clin Invest*. 2016;126(12):4690-4701. doi:10.1172/JCI88590.

21. Antón LC, Yewdell JW. Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. *J Leukoc Biol*. 2014;95(4):551-562. doi:10.1189/jlb.1113599.

22. Yewdell JW, Schubert U, Bennink JR. At the crossroads of cell biology and immunology: DRiPs and other sources of peptide ligands for MHC class I molecules. *J Cell Sci*. 2001;114(Pt 5):845-851. http://www.ncbi.nlm.nih.gov/pubmed/11181168.

23. Kim Y, Yewdell JW, Sette A, Peters B. Positional bias of MHC class I restricted T-cell epitopes in viral antigens is likely due to a bias in conservation. *PLoS Comput Biol*. 2013;9(1):e1002884. doi:10.1371/journal.pcbi.1002884.

24. Yewdell JW, Antbn LC, Bennink JR. Defective Ribosomal Products. 1996;7:1823-1826.

25. Schubert U, Antón LC, Gibbs J, Norbury CC, Yewdell JW, Bennink JR. Rapid

degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*. 2000;404(6779):770-774. doi:10.1038/35008096.

26. Rock KL, Farfán-Arribas DJ, Colbert JD, Goldberg AL. Re-examining class-I presentation and the DRiP hypothesis. *Trends Immunol*. 2014;35(4):144-152. doi:10.1016/j.it.2014.01.002.

27. Eisenlohr LC, Huang L, Golovina TN. Rethinking peptide supply to MHC class I molecules. *Nat Rev Immunol*. 2007;7(5):403-410. doi:10.1038/nri2077.

28. Kevei É, Pokrzywa W, Hoppe T. Repair or destruction-an intimate liaison between ubiquitin ligases and molecular chaperones in proteostasis. *FEBS Lett*. 2017;591(17):2616-2635. doi:10.1002/1873-3468.12750.

29. Duttler S, Pechmann S, Frydman J. Principles of cotranslational ubiquitination and quality control at the ribosome. *Mol Cell*. 2013;50(3):379-393. doi:10.1016/j.molcel.2013.03.010.

30. Rock KL, Gramm C, Rothstein L, et al. Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell*. 1994;78(5):761-771. doi:10.1016/S0092-8674(94)90462-6.

31. Perreault C. The origin and role of MHC class I-associated self-peptides. *Prog Mol Biol Transl Sci*. 2010;92:41-60. doi:10.1016/S1877-1173(10)92003-6.

32. Kraft C, Peter M, Hofmann K. Selective autophagy: ubiquitin-mediated recognition and beyond. *Nat Cell Biol*. 2010;12(9):836-841. doi:10.1038/ncb0910-836.

33. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of

codon bias. *Nat Rev Genet*. 2011;12(1):32-42. doi:10.1038/nrg2899.

34. Cannarozzi G, Cannarrozzi G, Schraudolph NN, et al. A role for codon order in translation dynamics. *Cell*. 2010;141(2):355-367. doi:10.1016/j.cell.2010.02.036.

35. Presnyak V, Alhusaini N, Chen YH, et al. Codon optimality is a major determinant of mRNA stability. *Cell*. 2015;160(6):1111-1124. doi:10.1016/j.cell.2015.02.029.

36. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell*. 2015;59(2):149-161. doi:10.1016/j.molcel.2015.05.035.

37. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol*. 2013;20(2):237-243. doi:10.1038/nsmb.2466.

38. Berg OG, Kurland C. Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol*. 1997;270(4):544-550. doi:10.1006/jmbi.1997.1142.

39. Pechmann S, Willmund F, Frydman J. Review The Ribosome as a Hub for Protein Quality Control. *Mol Cell*. 2013;49(3):411-421. doi:10.1016/j.molcel.2013.01.020.

40. Boël G, Letso R, Neely H, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*. 2016;529(7586):358-363. doi:10.1038/nature16509.

41. Bengio Y, Courville A, Vincent P. Representation Learning : A Review and New Perspectives. 2012;(1993):1-34.

42. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539.

43. Schmidhuber J. Deep Learning in neural networks: An overview. *Neural Networks*. 2015;61:85-117. doi:10.1016/j.neunet.2014.09.003.

44. Park Y, Kellis M. Deep learning for regulatory genomics. *Nat Biotechnol*. 2015;33(8):825-826. doi:10.1038/nbt.3313.

45. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform*. 2016:bbw068. doi:10.1093/bib/bbw068.

46. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*. 2015;33(8):831-838. doi:10.1038/nbt.3300.

47. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res*. 2008;36(Web Server issue):W509-12. doi:10.1093/nar/gkn202.

48. Goodfellow I, Bengio Y, Courville A. Deep Learning. *MIT Press*. 2016;521(7553):800. doi:10.1038/nmeth.3707.

49. Bishop CM. *Pattern Recognition and Machine Learning*. Vol 4. 2006. doi:10.1117/1.2819119.

50. Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Trans Neural Networks*. 1994;5(2):157-166. doi:10.1109/72.279181.

51. Laumont CM, Perreault C. Exploiting non-canonical translation to identify new targets

for T cell-based cancer immunotherapy. *Cellular and Molecular Life Sciences*. 2017:1-15.

52. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42. doi:10.1093/nar/gkt1196.

53. Granados DP, Rodenbrock A, Laverdure JP, et al. Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers. *Leukemia*. 2016;30(6):1344-1354. doi:10.1038/leu.2016.22.

54. Dittmar KA, Goodenbour JM, Pan T. Tissue-specific differences in human transfer RNA expression. *PLoS Genet*. 2006;2(12):2107-2115. doi:10.1371/journal.pgen.0020221.

55. Sagi D, Rak R, Gingold H, et al. Tissue- and Time-Specific Expression of Otherwise Identical tRNA Genes. *PLoS Genet*. 2016;12(8). doi:10.1371/journal.pgen.1006264.

56. Butt AM, Nasrullah I, Qamar R, Tong Y. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg Microbes Infect*. 2016;5(10). doi:10.1038/emi.2016.106.

57. Wong EHM, Smith DK, Rabadan R, Peiris M, Poon LLM. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol Biol*. 2010;10(1). doi:10.1186/1471-2148-10-253.

58. Pavon-Eternod M, Gomes S, Geslain R, Dai Q, Rosner MR, Pan T. tRNA over-expression in breast cancer and functional consequences. *Nucleic Acids Res*. 2009;37(21):7268-7280. doi:10.1093/nar/gkp787.

59. Goodarzi H, Nguyen HCB, Zhang S, Dill BD, Molina H, Tavazoie SF. Modulated

expression of specific tRNAs drives gene expression and cancer progression. *Cell*. 2016;165(6):1416-1427. doi:10.1016/j.cell.2016.05.046.

60. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653.

61. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol*. 2015;1A:A68-A77. doi:10.5114/wo.2014.47136.

62. Shao W, Pedrioli PGA, Wolski W, et al. The SysteMHC Atlas project. *Nucleic Acids Res*. 2017. doi:10.1093/nar/gkx664.

Annexes

Annexe 1	Pearson, H., Daouda, T., Granados, D. P., Durette, C., Bonneil, E., Courcelles, M., ... & Perreault, C. (2016). MHC class I–associated peptides derive from selective regions of the human genome. The Journal of clinical investigation, 126(12), 4690. (Article)
Annexe 2	Supplementary Informations from Granados, D. P., Sriranganadane, D., Daouda, T., *et al*. (2014). Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nature communications*, 5.
Annexe 3	Supplementary Informations from Daouda, T., Dumont-Lagacé, M., Thibault, P., Bengio, Y., Lemieux, S., Perreault, C. Codon usage regulates biogenesis of human MHC class I-associated peptides. Submitted to *Cancer Immunology Research*, 2017.
Annexe 4	Laumont, C. M., Daouda, T., Laverdure, J. P., Bonneil, É., Caron-Lizotte, O., Hardy, M. P., ... & Perreault, C. (2016). Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nature communications, 7, 10238. (Article)

Annexe 1 - MHC class I–associated peptides derive from selective regions of the human genome

Pearson, H., Daouda, T., Granados, D. P., Durette, C., Bonneil, E., Courcelles, M., ... & Lemieux, S. (2016). *The Journal of clinical investigation*, 126(12), 4690.

MHC class I–associated peptides derive from selective regions of the human genome

Hillary Pearson,[1,2] Tariq Daouda,[1,3] Diana Paola Granados,[1,2] Chantal Durette,[1] Eric Bonneil,[1] Mathieu Courcelles,[1] Anja Rodenbrock,[1] Jean-Philippe Laverdure,[1] Caroline Côté,[1] Sylvie Mader,[1,3] Sébastien Lemieux,[1,4,5] Pierre Thibault,[1,5,6] and Claude Perreault[1,2,5,7]

[1]Institute for Research in Immunology and Cancer, [2]Department of Medicine, [3]Department of Biochemistry, [4]Department of Informatics and Operational Research, [5]Canadian National Transplant Research Program, and [6]Department of Chemistry, Université de Montréal, Montreal, Quebec, Canada. [7]Division of Hematology-Oncology, Hôpital Maisonneuve-Rosemont, Montreal, Quebec, Canada.

MHC class I–associated peptides (MAPs) define the immune self for CD8⁺ T lymphocytes and are key targets of cancer immunosurveillance. Here, the goals of our work were to determine whether the entire set of protein-coding genes could generate MAPs and whether specific features influence the ability of discrete genes to generate MAPs. Using proteogenomics, we have identified 25,270 MAPs isolated from the B lymphocytes of 18 individuals who collectively expressed 27 high-frequency HLA-A,B allotypes. The entire MAP repertoire presented by these 27 allotypes covered only 10% of the exomic sequences expressed in B lymphocytes. Indeed, 41% of expressed protein-coding genes generated no MAPs, while 59% of genes generated up to 64 MAPs, often derived from adjacent regions and presented by different allotypes. We next identified several features of transcripts and proteins associated with efficient MAP production. From these data, we built a logistic regression model that predicts with good accuracy whether a gene generates MAPs. Our results show preferential selection of MAPs from a limited repertoire of proteins with distinctive features. The notion that the MHC class I immunopeptidome presents only a small fraction of the protein-coding genome for monitoring by the immune system has profound implications in autoimmunity and cancer immunology.

Introduction

MHC class I (MHCI) molecules present thousands of peptides at the surface of nucleated somatic cells (1). These MHCI-associated peptides (MAPs), collectively referred to as the immunopeptidome, regulate each step in the development and function of CD8⁺ T cells (2, 3). Indeed, real-time monitoring of the immunopeptidome is a vital process that allows CD8⁺ T cells to discriminate between self and nonself and to swiftly reject infected or transformed cells (4–6). Genesis of the immunopeptidome can be broadly divided into 2 events: (a) the processing of MAPs and (b) their binding to MHCI molecules (7, 8). The rules that regulate the second event, binding of MAPs to MHCI, are well defined: MHCI alleles are highly polymorphic, and each allotype has a specific peptide-binding motif that can be accurately predicted by several algorithms (9, 10). However, the first event, processing of MAPs, is a complex multistep process whose overall outcome cannot be predicted (1). Some proteins appear to generate more MAPs than others, but the mechanistic underpinning for these discrepancies remains elusive (11).

Classic biochemical studies have shown that MAP processing is initiated in the cytoplasm by proteasomal protein degradation followed by further trimming by cytosolic peptidases, transport in the ER, and final trimming by ER peptidases (8, 12–15). According to the dominant paradigm, MAPs preferentially originate from defective ribosomal products (DRiPs) which can be created by several mechanisms such as nonsense-mediated decay (NMD), mRNA destabilization, or noncanonical translation in the cytosol or the nucleus (16–20). Large-scale mass spectrometry (MS) offers the sole direct approach to analyzing the global molecular composition of the immunopeptidome. Previous large-scale MS studies of MAPs presented by one or a few MHCI allotypes have shown that thousands of proteins located in all cell compartments can be the source of MAPs (21–24). However, the rules of MAP processing cannot be figured out by studying the immunopeptidome presented by individual HLA allotypes because each allotype can only bind peptides containing a specific motif (25, 26).

The goals of our study were to assess the extent of MAP generation from the entire set of protein-coding genes and to determine whether specific features influence the ability of discrete genes to generate MAPs. We used a well-validated high-throughput proteogenomic approach to identify MAPs presented by 27 HLA-A and HLA-B allotypes on B lymphoblastoid cell lines (B-LCLs) derived from 18 subjects. Overall, we identified 25,270 nonredundant MAPs, which derived from 6,195 out of the 10,575 genes expressed in B-LCLs. Hence, while 59% of genes were the source of 1–64 MAPs per gene, 41% of expressed genes were not represented in the immunopeptidome. Overall, we estimate that the immunopeptidome presented by 27 alleles covered only 10% of exomic sequences expressed in B-LCLs. We then used a series of bioinformatic tools to understand how identifiable features of genes, transcripts, and proteins could influence MAP generation. With these data we built a logistic regression model that was able to predict whether or not a given gene will produce MAPs with a receiver operating characteristic

▶ Related Commentary: p. 4399

Conflict of interest: The authors have declared that no conflict of interest exists.
Submitted: May 13, 2016; **Accepted:** September 30, 2016.
Reference information: *J Clin Invest.* 2016;126(12):4690–4701. doi:10.1172/JCI88590.

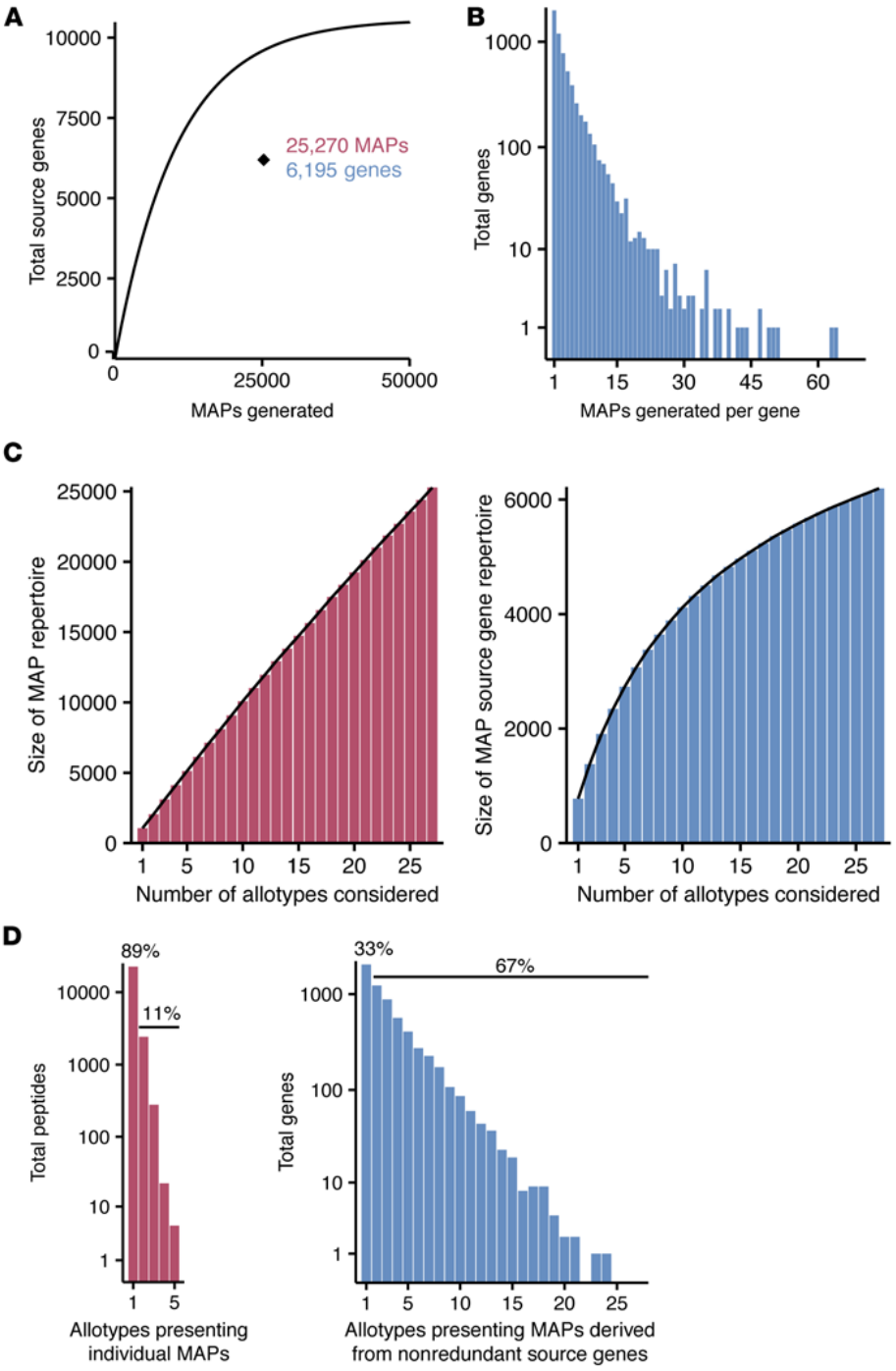


Figure 1. The immunopeptidome presented by 27 HLA allotypes. (**A**) Total number of nonredundant MAPs and their source genes in the immunopeptidome of 18 B-LCLs compared with an expected binomial distribution. The curve depicts the expected number of source genes if all genes had a similar ability to generate MAPs. The black diamond shows the actual number of source genes (n = 6,195) observed for 25,270 MAPs ($P < 1 \times 10^{-250}$, binomial test). (**B**) Histogram showing the number of MAPs generated per MAP source gene (range = 1–64). (**C**) The number of unique identifications of MAPs (left panel) and MAP source genes (right panel) was counted for various numbers of randomly selected HLA allotypes. Results show the average of 1,000 simulations. (**D**) The promiscuity of antigen presentation for MAPs (left panel) and their source genes (right panel). Histograms show the number of allotypes associated with each peptide or gene.

(ROC) AUC of 0.81 ± 0.02 (95% CI). Our results show that the immunopeptidome is forged from a limited repertoire of gene products with distinct features influencing transcription, translation, and proteasomal degradation.

Results

Proteogenomic-based definition of the MAP repertoire presented by 27 HLA allotypes. To obtain a comprehensive representation of the immunopeptidome presented by HLA-A and HLA-B molecules, we applied a well-validated high-throughput proteogenomic approach that hinges on a combination of next-generation sequencing and high-throughput MS (20, 27, 28). Transcriptome and exome-sequencing data were used to build personalized protein databases for B-LCLs of 18 subjects using the Python package pyGeno (29). These personalized databases were used for peptide identification by MS. MAPs were eluted from the cell surface by mild acid elution, and stringent quality filters were applied to the list of MAPs assigned by MS: (a) a peptide length of 8–14 amino acids, (b) a 1% false discovery rate based on searches against concatenated target/decoy databases (30), (c) assignment to single genetic origin among the 10,575 protein-coding genes expressed and annotated in B-LCLs, and (d) a predicted MHCI IC_{50} of less than or equal to 1,250 nM according to the NetMHC or NetMHCcons algorithms (31, 32) (Supplemental Figure 1C, see details in Methods; supplemental material available online with this article; doi:10.1172/JCI88590DS1). About 99.8% of individuals of European descent bear at least one of the 27 HLA-A,B allotypes studied (33).

We identified 25,270 nonredundant MAPs derived from 6,195 source genes in, to the best of our knowledge, the largest set of MHCI-associated peptides reported to date (Figure 1A and Supplemental Tables 1 and 2). Strikingly, only 59% expressed and annotated genes in B-LCLs were capable of generating detectable MAPs. MAP source genes produced up to 64 individual MAPs, and 68% of these genes produced more than 1 MAP (Figure 1B). To estimate the diversity of a multiallelic immunopeptidome, we computed the size of the MAP repertoire and MAP source gene repertoire as a function of the number of HLA allotypes considered (Figure 1C). We counted the number of unique identifications when a given number of randomly selected allotypes was consid-

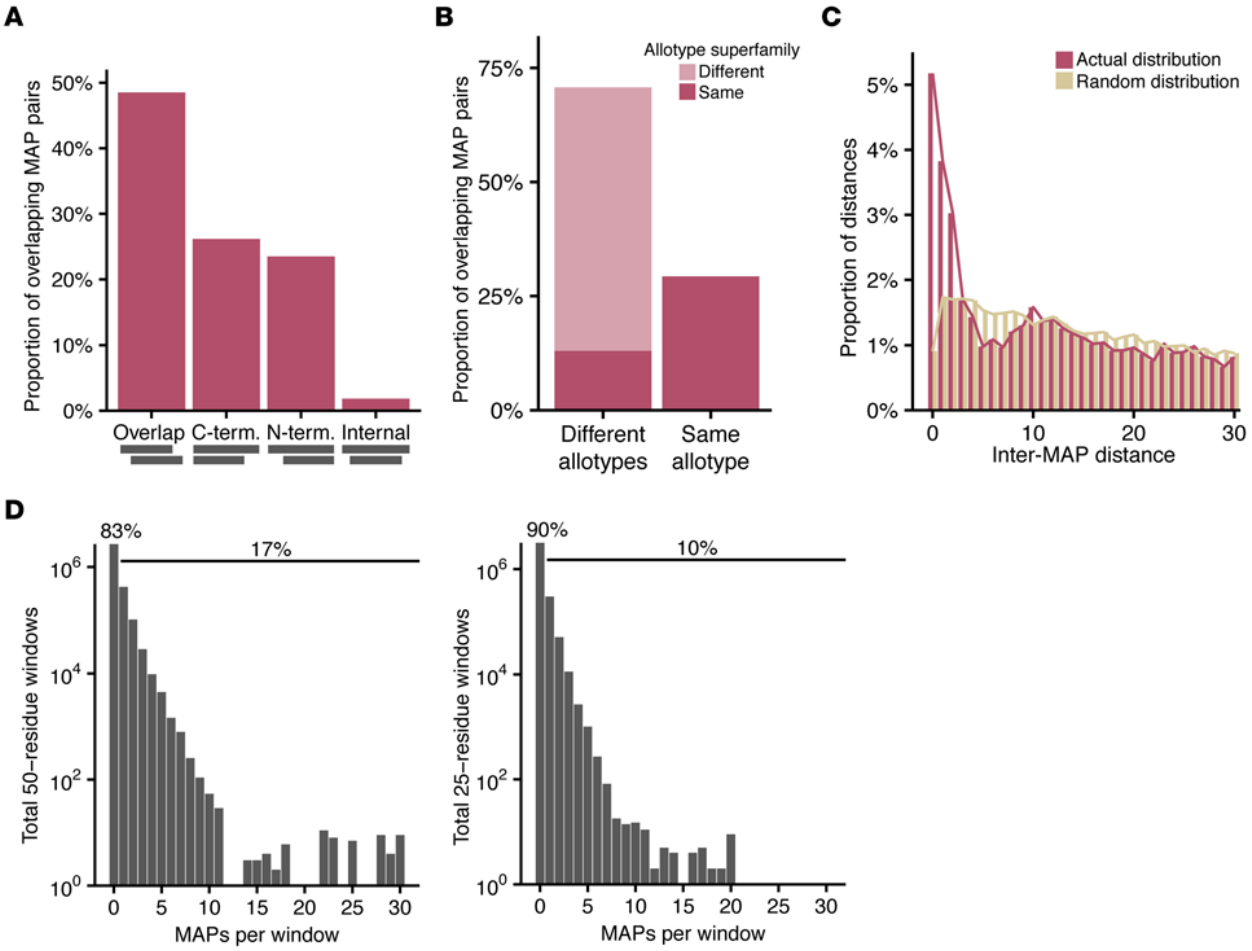


Figure 2. Spatial distribution of MAPs along source proteins. (**A**) Distribution of overlap types for 3,682 pairs of overlapping MAPs formed by 5,046 individual peptides: pairs with any overlapping residues and no common ends; pairs with a common C terminus (C term); pairs with a common N terminus; and pairs with 1 peptide contained within the other. (**B**) Proportion of overlapping MAP pairs presented by the same allotype or different allotypes. For MAP pairs presented by different allotypes, whether the 2 allotypes belong to the same superfamily is indicated (34). (**C**) Distances between MAP start sites along proteins generating more than 1 MAP compared with a matched, random distribution. Distances are shown up to 30 residues. Distances are significantly shorter in the actual distribution (Wilcoxon rank sum test, $P = 7 \times 10^{-52}$). (**D**) Exome coverage by the immunopeptidome. A window of 50 or 25 amino acids (left and right panel, respectively) was moved residue by residue along proteins of the transcribed exome of B-LCLs. Histograms show the number of MAPs found in each window; the proportion of windows containing 0 versus at least 1 MAP is indicated.

ered. For MAPs, the nearly linear nature of this relationship demonstrated little redundancy in the peptides presented by different allotypes (Figure 1C). Conversely, the redundancy of the genes generating MAPs across all 27 HLA allotypes was much greater (Figure 1C). As more allotypes were considered, a diminishing number of additional genes were represented in the immunopeptidome. A simulation examining the size of the peptide and gene repertoires as various numbers of subjects were considered showed similar results (Supplemental Figure 1, A and B). Most MAPs (89%) were presented by a single HLA allotype (Figure 1D). The few promiscuous binders were presented by HLA allotypes with similar peptide-binding motifs (i.e., same "superfamily"), such as A*03:01 and A*11:01 (34). However, the majority of MAP source genes (67%) produced MAPs for multiple allotypes, some for up to 24 of the 27 allotypes studied (Figure 1D).

Since MS analyses can be subject to some stochastic variations, would it be possible that no MAPs were assigned to 41% of expressed genes because these MAPs were missed by MS? We reasoned that if our MS analyses randomly missed some MAPs, the proportion of MAP source versus nonsource genes should nevertheless follow a binomial distribution where the number of source genes increases as a function of the number of detected MAPs (Figure 1A). Notably, we found that the 25,270 MAPs that we identified by MS derived from significantly fewer genes ($n = 6{,}195$) than predicted by a binomial distribution (Figure 1A, exact binomial test $P < 10^{-250}$). Hence, random failure to detect some MAPs cannot explain that only 59% of genes were found to generate MAPs. In addition, we used internal standard triggered–parallel reaction monitoring (IS-PRM) in order to compare the detection threshold for 2 sets of stable isotopically labeled synthetic peptides (35). Peptides AEIEQKIKEY, EEINLQRNI, EEIPVSSHY and EESAVPERSW (underlined residues indicate ^{13}C, ^{15}N-labeled amino acids) had the amino acid sequence of MAPs presented by B*44:03. The other synthetic peptides AESQELLTF, EESHLNRHF, HESAEGKEY, and TESSDITEY corresponded to amino acid sequences from nonsource genes (i.e., not detected in our initial shotgun MS analyses)

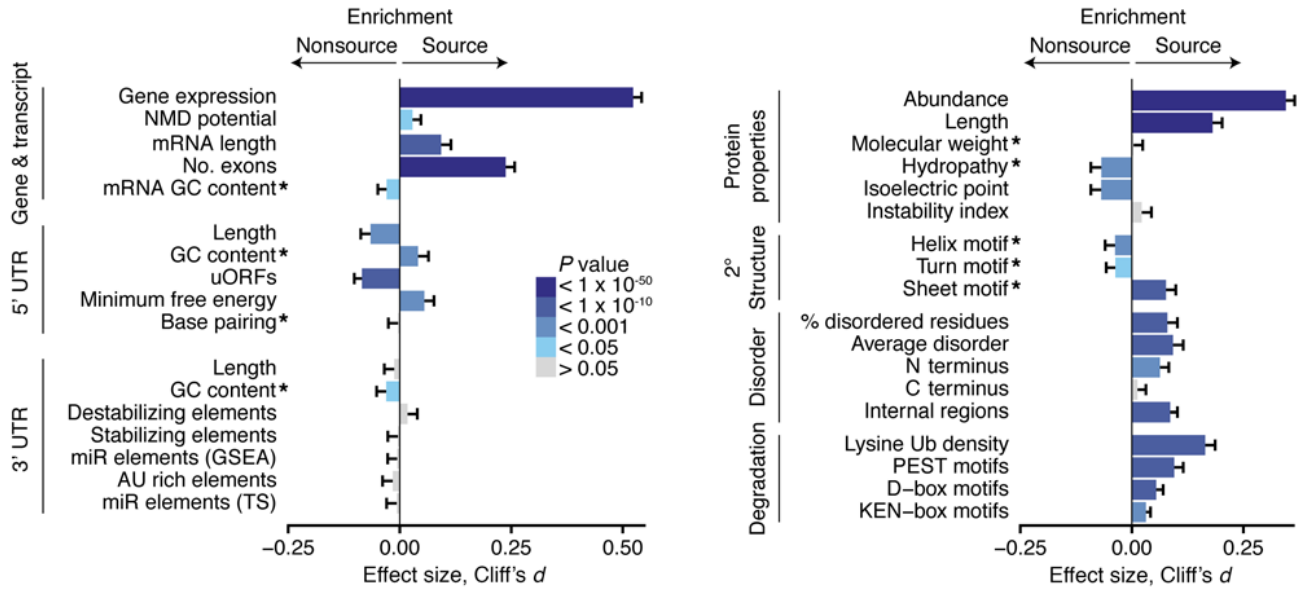


Figure 3. Features of MAP source and nonsource genes, transcripts, and proteins. Error bars represent a 95% CI based on bootstrapping for Cliff's *d* value, a nonparametric measurement of effect size. *P* values derived from 2-sided Wilcoxon tests; 6,195 source and 4,380 nonsource genes and gene products were studied for each comparison. * indicate features that were normalized for the respective transcript, UTR, or protein lengths. See Methods for details of how each feature was calculated. miR, microRNA; TS, TargetScan software; Ub, ubiquitination site.

that were randomly chosen among peptides predicted to be strong binders for B*44:03 (IC_{50} < 50 nM). These synthetic peptides were spiked (100 fmoles each) in mild acid elution extracts from 3 different B-LCLs to correlate the identification of the corresponding endogenous MAPs. IS-PRM analyses showed that the detection threshold was similar for the 2 groups of synthetic peptides (Supplemental Figure 2). Furthermore, none of the selected B*44:03 strong binders coded by nonsource genes were detected in the 3 different B-LCLs using IS-PRM. In contrast, endogenous peptides from source genes presented by B*44:03 were all correlated with their corresponding synthetic peptides (Supplemental Figure 2). These results provide compelling evidence that failure to detect MAPs from nonsource genes cannot be ascribed to MS bias against the product of nonsource genes.

Two major points can be made from these data: (a) a distinct subset of genes produced most MAPs, and (b) our method captured the majority of MAP source genes (Figure 1C). As a corollary, these results suggest a model whereby a common pool of source proteins selectively enters the antigen-processing pathway and can generate MAPs with suitable motifs for most MHCI allotypes.

Discrete protein regions are preferential sources of MAPs. We next asked whether there might be "hot spots" in MAP source genes, i.e., regions or domains that provide disproportionately high amounts of MAPs. To this end, we analyzed the spatial distribution of MAPs along proteins that generated more than one MAP. We first identified 3,682 pairs of overlapping MAPs formed by 5,046 individual peptides (20% of the entire data set). In a given pair, MAPs differed from each other at their N and/or C terminus (Figure 2A). These pairs may result from differential trimming of a common precursor by various peptidases in the cytosol and ER. Notably, 71% of MAP pairs bound different allotypes; of these, 82% bound allotypes from different superfamilies (Figure 2B). Hence, from the perspective of an MHCI allotype, generation of overlapping MAP pairs is generally not redundant: members of a pair are seldom presented by the same MHCI allotype. At the population level, the net result is that some protein regions are included in the immunopeptidome of many people who do not share the same HLA alleles.

To further evaluate whether selected protein regions were preferential sources of MAPs, we analyzed the spatial distribution of MAPs along proteins. For each protein, distances between adjacent MAP start sites were calculated. A control distribution was generated by randomly placing the same number of MAPs along the same protein length. We found that MAPs colocalized along proteins more than expected ($P = 7 \times 10^{-52}$, Figure 2C). The fact that no MAPs were assigned to 41% of expressed protein-coding genes, together with the clustering of MAP-coding sequences in source genes, suggests that the immunopeptidome covers a limited portion of the whole exome. To estimate global exome coverage, (a) we moved a walking window of 150 base pairs (50 amino acids) along the exome coding for the 10,575 genes expressed in B-LCLs, and (b) we calculated the number of MAPs seen in each window. We found that 83% of windows generated no MAPs, whereas 17% of windows covered 1–30 MAPs per window (Figure 2D). When we reduced the window size to 75 base pairs, only 10% of windows were a source of MAPs (Figure 2D). From this, we conclude that the immunopeptidome presented by 27 HLA-A,B allotypes covers an unexpectedly small portion of the whole transcribed exome.

Gene expression cannot solely account for differential ability of genes to generate MAPs. Understanding the genetic origins of the immunopeptidome is of paramount importance fundamentally and in the search for MAPs that could be used as therapeutic targets. What distinguishes the 6,195 genes that were capable of generating MAPs compared with the other 41% of genes from

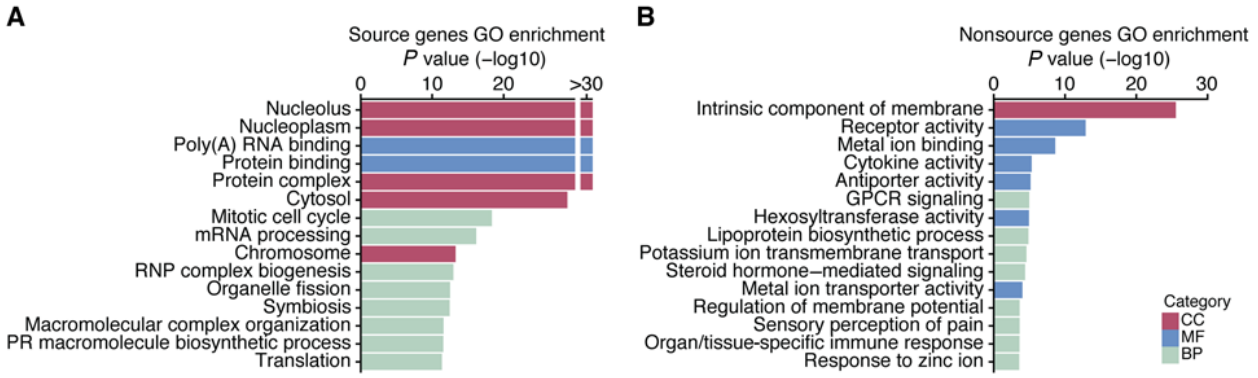


Figure 4. GO analysis of source and nonsource genes. Enrichment in source (**A**) and nonsource (**B**) groups was calculated on a background of both groups using the topGO algorithm to eliminate redundancies (60). The top 15 most enriched functions are shown for each group including all 3 ontology categories. For all GO terms significantly enriched in source and nonsource gene categories, see Supplemental Tables 3 and 4. PR, positive regulation; RNP, ribonucleoprotein; CC, cellular component; MF, molecular function; BP, biological process.

which no MAPs were detected? To answer this question, we applied a variety of analyses and prediction algorithms to study the features of MAP source and nonsource genes, transcripts, and proteins. We first asked whether MAP source proteins simply contained more potential HLA-binding peptides, i.e., peptides with the right binding motif for the 27 HLA allotypes considered here. This was not the case: the density of predicted nonamer MHCI binders was no greater in source genes than nonsource genes (Supplemental Figure 1D). Since the difference between MAP source and nonsource genes is unrelated to the number of potential MHC binders, it must therefore involve discrepancies in the processing of MAP source proteins.

Whether gene expression influences MAP generation is a controversial issue, as shown by previous studies based on smaller data sets. According to some reports, MAPs derive preferentially from highly abundant mRNAs or proteins (19, 25, 36), but other reports cast some doubts on this contention (23, 37). By analyzing RNA-sequencing data of the 18 B-LCLs studied herein, we found that the average gene expression was significantly higher for MAP source genes (Figure 3). However, expression alone provided an incomplete portrait of antigen presentation: some highly expressed genes generated no MAPs, and some lowly expressed genes were capable of generating MAPs. Since the proteome is an imperfect mirror of the transcriptome (38, 39), we also analyzed the relationship between protein abundance in human B cells (40) and MAP generation. MAP source proteins are more abundant than nonsource proteins (Figure 3), yet the fact that some proteins with similar expression belonged to source or nonsource groups suggested that other factors were at play.

MAP source transcripts are enriched in features conferring greater translation efficiency. Ultimately, MAP generation must be regulated at the level of translation and protein degradation (41). To gain further insights into the mechanisms regulating MAP generation, we analyzed the potential role of factors regulating protein metabolism. We first asked whether features enhancing translation efficiency and transcript stability may distinguish source from nonsource transcripts. Coherent with the concept that NMD is a source of MAPs (18), we observed that the proportion of genes with at least one transcript with an NMD biotype (determined with the ENSEMBL regulatory build) was higher in source relative to nonsource genes (Figure 3). Also, consistent with the positive correlation between the number of exons and translation efficiency (42), we found that MAPs derived from transcripts composed of more exons than nonsource transcripts (Figure 3), even when normalized for transcript length ($P = 5 \times 10^{-49}$).

We next examined features of the 5′ UTR for evidence of translational regulation of antigen processing. Upstream open reading frames (uORFs) tend to negatively influence translation by destabilizing transcripts and by acting as physical obstacles that slow ribosomal scanning (43). The 5′ UTRs of MAP source transcripts were significantly shorter and contained fewer predicted uORFs. Similarly, the secondary structure predicted with Vienna RNAfold (44) revealed greater free energy scores in spite of enriched GC content for MAP source 5′ UTRs. No definitive differences between the amount of base pairing in 5′ UTR structures were found (Figure 3). These findings suggest that MAP source 5′ UTRs are structurally fluid and contain fewer obstacles to translation.

The 3′ UTR is a critical site of translational control containing regulatory elements such as adenylate-uridylate–rich (AU-rich) elements and binding sites for microRNAs and RNA-binding proteins (45). Despite this regulatory potential, we initially remarked no difference in the lengths of 3′ UTRs (Figure 3). The density of AU-rich elements was similar; however, our analyses could not take into account the distinction between AU elements involved in rapid decay and finer stability regulation (46). Greater GC content was found in nonsource 3′ UTRs and along the entire mRNA transcript in general; this may reflect a positive association with mRNA levels in a degradation-independent manner (47). Stabilizing and destabilizing regulatory elements were queried in the 3′ UTRs of all transcripts (48) and revealed similar prevalence in source and nonsource transcripts (Figure 3). Moreover, we were unable to confirm previous results that MAPs derive preferentially from transcripts with microRNA-binding sites using 2 different tools: Gene Set Enrichment Analysis (GSEA) and TargetScan (19, 49, 50) (Figure 3). However, our negative findings regarding binding sites for microRNAs and RNA-binding proteins must be considered with some reservations. First, microRNA regulation is highly cell-type specific, while the methods used to predict microRNA involvement operate at an organism-wide level (50). Second, since the effects of

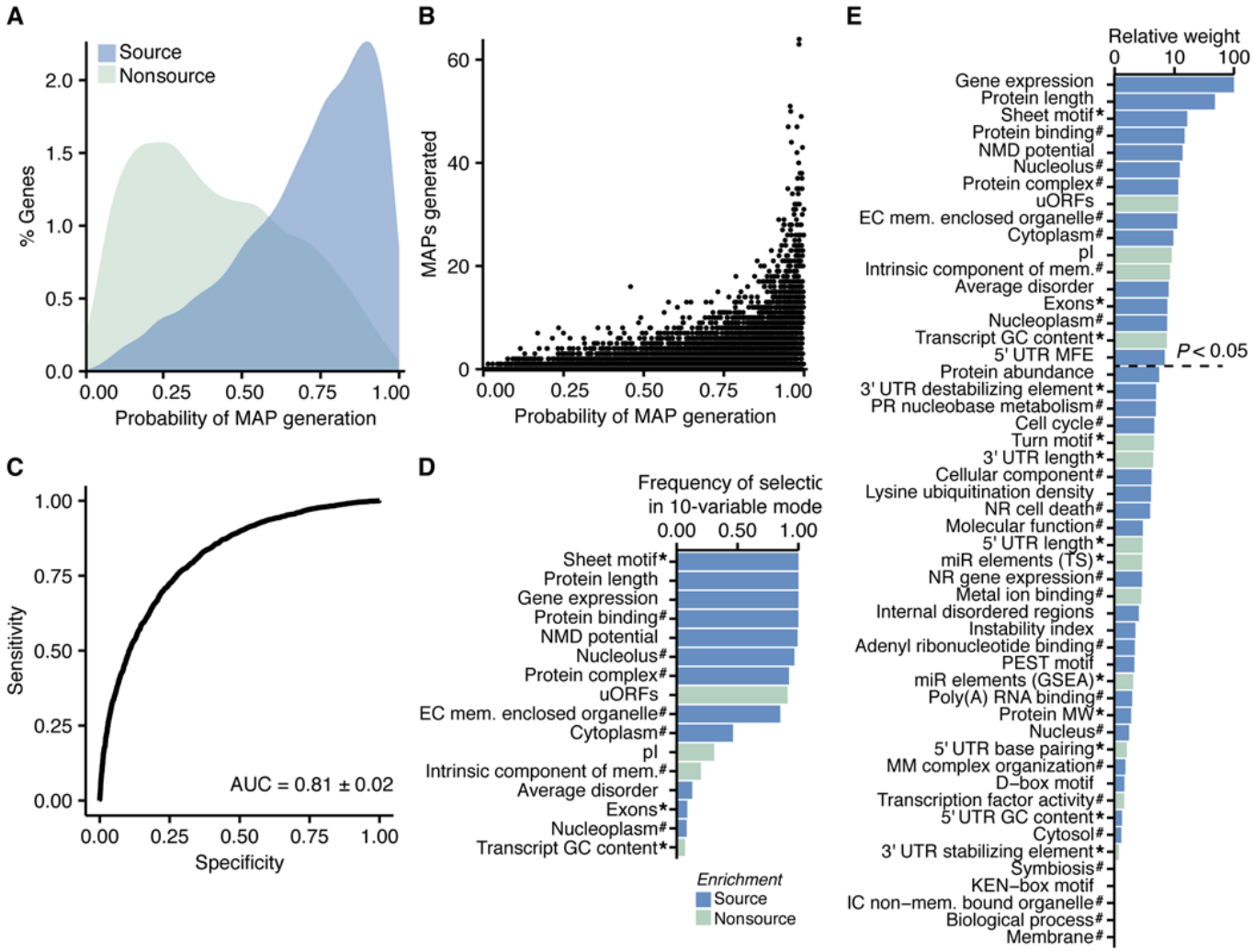


Figure 5. A logistic regression model to predict whether or not a gene will generate MAPs. (**A**) Prediction scores for each gene grouped by experimentally defined source classification. (**B**) Prediction scores for each gene and the number of MAPs generated. (**C**) Model performance measured by a ROC plot of sensitivity (the rate of true positives) as a function of specificity (the rate of true negatives); the AUC is 0.81 ± 0.02 (95% CI). (**D**) Frequency of input variable selection in a logistic regression model using recursive feature elimination; frequencies above 0.05 are shown. (**E**) The relative weight of all input variables in the 2-class logistic regression model. Variables normalized by the length of the corresponding UTR, transcript, or protein are denoted by * and GO terms denoted by #. EC, extracellular; IC, intracellular; Mem., membrane; MFE, minimum free energy; MM, macromolecular; NR, negative regulation of; PR, positive regulation of; TS, TargetScan software. All metrics are averaged over 1,000 models (see Methods for details).

3′ UTR regulatory elements are heavily context dependent (45), the role of 3′ UTR regulation in MAP generation may be obscured by the specific activity of predicted elements in B-LCLs. Nonetheless, in contrast with the 5′ UTR, these findings indicate at least limited regulation of MAP generation by 3′ UTR elements.

Notably, features enriched in MAP source transcripts and UTRs had minimal correlations with protein abundance (absolute Spearman's rank correlation coefficient r of 0.22 for number of exons and $r < 0.14$ for others, Supplemental Figure 3). This led us to postulate that gene expression and transcript features may provide nonredundant information for the modeling of MAP generation.

The primary and secondary structure of proteins regulates MAP generation. Next, we assessed the electrochemical and structural features of MAP-generating proteins. We confirmed previous reports that longer proteins generate more MAPs (25, 36) (Figure 3). This may reflect that, relative to shorter proteins, longer proteins (a) contain more MHCI-binding sequences, (b) have a greater chance of forming DRiPs, and (c) bind more ribosomes (25,42). MAP source proteins had lower hydropathy scores, indicating more polar amino acid composition. Furthermore, the predicted isoelectric point revealed greater acidic composition of source proteins (Figure 3). At the next level of complexity, the predicted secondary structure of MAP source proteins showed distinct contributions of helix, turn, and sheet motifs. In particular, MAP source proteins showed a conspicuous enrichment in sheet motifs (Figure 3).

The ubiquitin proteasome system is a key entry point for proteins into the MHCI-processing pathway (7, 51). We first examined MAP proteins for proteasomal degradation motifs. We found that, compared with nonsource proteins, MAP source proteins contained higher frequencies of (a) KEN-box and D-box motifs targeted by the anaphase-promoting complex ubiquitin ligase (52), (b) PEST motifs, which serve as proteolytic signals for the proteasome and other proteases (53), and (c) canonical lysine ubiquitination sites (54) (Figure 3).

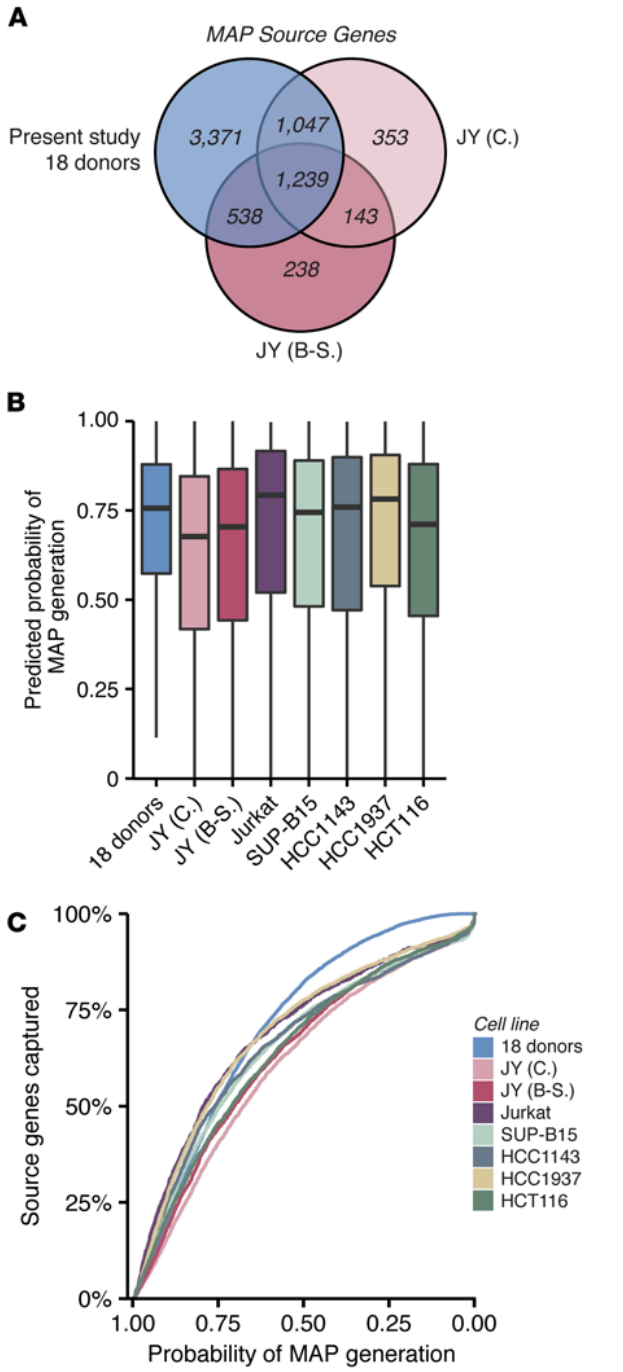


Figure 6. Evaluation of model performance with independent data sets on human cancer cell lines. (**A**) Overlap in source gene identifications between the present study and 2 independent studies of JY B-LCLs using different MS techniques: JY (C.) and JY (B-S.). (**B**) Distribution of prediction scores for MAP source genes in B-LCLs and cancer cell lines (details in Table 1); median value is shown with whiskers extending to the extremes of the interquartile range x 1.5; outliers are hidden. (**C**) Proportion of MAP source genes captured as a function of prediction score threshold.

Unstructured protein regions serve as initiation sites for proteasomal degradation (55), and intrinsically disordered segments favor proteasome degradation (56). Therefore, to analyze the potential influence of protein disorder on MAP generation, we computed the disorder status of proteins in our data set with the neural network predictor PONDR VLXT (57). Whether the proportion of disordered residues, the average disorder of all residues, the length of N-terminal disorder, or the presence of internally disordered regions longer than 30 residues was considered, MAP source proteins consistently contained greater disorder compared with nonsource proteins (Figure 3). Similar results were obtained using 2 other disorder predictors: DISOPRED (58) and IUPRED (59) (Supplemental Figure 4B). We conclude that primary and secondary structure of proteins and particularly features linked to proteasomal degradation have a strong influence on MAP generation.

GO terms analysis. We next compared the enrichment of gene ontology (GO) terms in MAP source and nonsource genes using the topGO algorithm (60) to eliminate redundancies (Figure 4). Our findings here confirm and extend reports based on smaller data sets (19, 22, 25). The source gene population was highly enriched in genes coding for intracellular proteins interacting with DNA, RNA, and other proteins (Figure 4A and Supplemental Table 3). This may have resulted from a relatively greater expression of genes implicated in housekeeping functions, such as poly(A) RNA binding, mitotic cell cycle, and mRNA processing. Non–mutually exclusive hypotheses are that these genes have a preferential access to the MHC-processing machinery, for example, via "immunoribosomes," or that components of macromolecular complexes have a greater propensity to form DRiPs (17). Nonsource proteins were enriched in membrane components and related signaling processes, demonstrating that proteins traversing the secretory pathway are poorly represented in the MHCI immunopeptidome (Figure 4B and Supplemental Table 4).

Modeling MAP generation. Having identified features that differentiate MAP source versus nonsource genes, we asked whether it might be possible to build a model for predicting whether a given gene generates MAPs. Taking into account features listed in Supplemental Table 5 (see also Supplemental Figure 3), we trained a logistic regression model on 80% of our data set using 10-fold cross-validation and tested its ability to discriminate MAP source versus nonsource genes on the remaining 20% of our data set. The process was repeated 1,000 times with randomly divided training and test data sets (see Methods for details). Prediction scores, falling between 0 and 1, demonstrated a considerable ability to correctly discriminate between MAP source and nonsource genes (Figure 5A). Although the model was blind to the number of MAPs produced by source genes, we found that the predictions corresponded to the rate of MAP production (Figure 5B).

To assess the overall predictive power of the model, we constructed ROC plots with averaged prediction scores and found an AUC of 0.81 ± 0.02 (95% CI) (Figure 5C). By examining the parameters of the model, we assessed the relative contribution of each feature to learning (Figure 5E). We found that gene expression was the most informative variable, followed by protein length, the presence of sheet motifs, and various GO terms. Features of genes, transcripts, and proteins were included in the group of relatively less important but significant variables, suggesting that a wide range of fine-tuning processes contribute to MAP generation. Since estimates of relative importance can be influenced by related variables, we used a second method to assess feature importance. We assessed the predictive capacity of a logistic regression model, selecting only the top 10 most infor-

Table 1. Features of human B-LCLs and cancer cell lines used to evaluate model performance

Cell line	Type	Method	AUC	*n*
18 Donors	B lymphoblast cell line	MAE/DDA	0.81	6,195
JY (C.)	B lymphoblast cell line	IP/DIA	0.83	2,782
JY (B-S.)	B lymphoblast cell line	IP/DDA	0.83	2,185
Jurkat	T cell lymphoblast leukemia	IP/DIA	0.82	959
SUP-B15	Acute lymphoblastic leukemia	IP/DDA	0.85	2,997
HCC1143	Breast carcinoma	IP/DDA	0.79	3,136
HCC1937	Breast carcinoma	IP/DDA	0.83	4,546
HCT116	Colorectal carcinoma	IP/DDA	0.83	2,900

The MS method used to detect MAPs and the number of MAP source genes identified in each sample (n) are indicated. For each cell line, an AUC derived from predictions by the 2-class logistic regression model is reported. DDA, data-dependent acquisition; DIA, data-independent acquisition; MAE, mild acid elution.

mative features. Despite this constraint, the model demonstrated comparable predictive power. The frequency with which features were selected in this model (Figure 5D) coincided with the relative weight when all input variables were considered (Figure 5E).

A 2-class distinction of MAP source and nonsource genes does not take into consideration that some source genes generate up to 64 nonredundant MAPs, while other genes produce only one (Figure 1B). To integrate these findings, we produced a nuanced version of the classification model that made predictions for 3 ordered groups: none (no MAPs), low (1–2 MAPs), and high (≥ 3 MAPs). Predictions were most accurate for the high category, which obtained an AUC of 0.86 ± 0.02, while the low and none groups had AUCs of 0.64 ± 0.01 and 0.81 ± 0.02, respectively (Supplemental Figure 5A). Clearly, the model had difficulty distinguishing the low group, for which its predictions reached a maximum probability of 0.43 compared with 0.99 for the high and none categories (Supplemental Figure 5C). Interestingly, when we compared the relative contribution of different input parameters between the 2-class and 3-class models, we found a very similar hierarchy (Figure 5E and Supplemental Figure 5B). We conclude that no particular feature within the model distinguishes genes that generate few versus numerous MAPs.

Model validation with independent data sets and human cancer cell lines. The various strategies used for high-throughput MS analyses of the immunopeptidome present strengths and limitations (61). In the present study, MAPs were isolated from 18 B-LCLs by mild acid elution and analyzed by data-dependent MS. To gauge the robustness of the model, we tested it on MAPs identified by 2 other groups in JY B-LCLs. MAPs in these 2 data sets were isolated by MHCI immunoprecipitation; one study used data-dependent MS (36), and the other used data-independent MS (21). While our data set contained MAPs presented by 27 HLA-A,B allotypes, JY B-LCLs express just 2 of these: HLA-A*02:01 and HLA-B*07:02. Transcriptomic data from JY B-LCLs (62) defined a set of candidate genes on which we performed predictions with the 2-class logistic regression model. Notably, 82% of source genes for the 2 other data sets were included in our own set of source genes (Figure 6A). Moreover, our model effectively predicted MAP origin in these 2 independent data sets (Table 1 and Figure 6, B and C) despite differences in methods of MAP isolation and MS analyses.

We further challenged the model trained on 18 donor B-LCLs to predict MAP generation in 5 human cancer cell lines: 2 leukemias, 2 breast carcinomas, and 1 colorectal carcinoma (Table 1). To evaluate the performance of our model, we used previous analyses of the transcriptome (63) and the immunopeptidome (21, 36) of these cell lines. The models' ability to predict MAP source genes was very good for the 5 cancer cell lines, with ROC AUC ranging from 0.79 to 0.85, similar to the accuracy observed with B-LCLs (Table 1). The distribution of the prediction scores for MAP source genes was similar in the various cell lines (Figure 6B), though the rate at which source genes were captured at different probabilities of MAP generation revealed slight divergence at the lower prediction scores (Figure 6C). These data suggest that MAP processing follows consistent rules with limited variations between cell types. Overall, we conclude that our prediction model is robust for cells of various lineages and that its accuracy is not biased by the methods used for MAP isolation or identification.

Discussion

To the best of our knowledge, this study reports the largest data set of MAPs to date. Several points can be made from our comprehensive analyses of 25,270 MAPs presented by 27 HLA-A,B allotypes, which illustrate how there can be "strength in numbers" (64). Indeed, while analyses of smaller data sets suggested that individual genes were represented in the immunopeptidome by only a single MAP (25), we found that MAP source genes generated up to 64 nonredundant MAPs. Importantly, we found that MAPs presented by 27 MHCI allotypes together cover an unexpectedly small fraction of the protein-coding exome (10%) because (a) 41% of genes did not generate detectable MAPs, and (b) MAPs derived from the same gene tend to originate from adjacent sequences. At the population level, one implication is that even though HLA allotypes have different peptide-binding motifs, a large fraction of MAPs presented by different subjects (2 to 4 HLA-A,B allotypes/subject) will originate from common genomic regions. Further studies are certainly warranted in order to explore whether, relative to the whole exome, MAP "hot spots" have distinctive features that would make their monitoring by T cells of special importance or whether these regions are simply opportunistically captured. For instance, are these hot spots preferential sites of somatic mutations in cancer cells or do they resemble viral genes? Notably, we observed that some features enriched in MAP source genes are also common in viral genes (e.g., internal disorder and 5′ UTR secondary structures). Indeed, disorder is prevalent in viral genomes (65), and several viral transcripts contain complex 5′ UTR secondary structures that stall ribosomal translation (66).

Our results suggest that at the systems level, MAP generation is regulated by specific features of transcripts and proteins that often affect translation and proteasomal degradation. For example, features of the 5′ UTR, such as shorter length, looser secondary structure, and fewer uORFs, which are easier for ribosomes to navigate, may confer efficient translation and consequently greater MAP generation. The importance of proteasomal process-

ing is underscored by the prevalence of disorder and degradation motifs in MAP source proteins. Additionally, that MAPs originate preferentially from abundant transcripts is consistent with the fact that the immunopeptidome is different from one cell lineage to another and is affected by the metabolic status of cells (5, 51). The relation between transcript abundance and MAP presentation may also be relevant to the establishment of self-tolerance in the thymic medulla. Indeed, central self-tolerance depends on promiscuous gene expression by medullary thymic epithelial cells that collectively express almost all protein-coding genes (67, 68). Remarkably, this promiscuous gene expression follows a mosaic pattern: individual medullary thymic epithelial cells promiscuously express a limited number of genes, but at a high level (67, 69). A mosaic pattern of highly expressed genes may be instrumental in increasing the breadth of the MAP repertoire that can thereby induce central self-tolerance.

By taking into account the various features enriched in MAP source genes, we were able to build a logistic regression model that predicts whether or not a given gene will produce MAPs with a ROC AUC of 0.81 ± 0.02. The robustness of this model was validated by predicting MAP generation in 7 independent data sets. Notably, when the model was applied to predict MAP generation in 5 human cancer cell lines, it performed comparably well, suggesting strong potential for predicting MAP generation in a clinical context. Would it be possible to build an in silico antigen-processing machine that would predict with even greater accuracy sources and sites of MAP generation? We speculate that this may be possible if we trained the model with more quantitative data and more accurate assessment of features. Indeed, there are certain limitations to a rather coarse 2-class output, not the least of which is a lack of precision for the number of MAPs produced and their location along a protein. Recent developments in MS now enable quantification of MAPs in terms of number of copies per cell (61). High-throughput quantitative analyses of immunopeptidomes could thereby pave the way to the development of improved predictive models, and community-based efforts to achieve this goal should be encouraged (21).

Our demonstration that the immunopeptidome covers only a small fraction of the protein-coding exome has special relevance to cancer immunology. There is a general consensus that cancer-specific neo-MAPs derived from somatic mutations represent ideal targets for cancer immunotherapy (70). However, discovery of cancer-specific MAPs is currently fraught with major difficulties. Typically, neo-MAP discovery strategies adopt the following path: exome sequencing, identification of mutations, and selection of mutations located in peptide regions predicted to have a good MHC-binding affinity. However, when putative neo-MAPs are tested experimentally, by MS or immune assays, the hit rate is below 10% (71–73). Our contention is that this low success rate is simply due to the fact that few mutations are strategically located in MAP-generating regions and that most mutations are in exomic sequences that are not covered by the immunopeptidome. We believe that progress in the field of neo-MAP discovery will be greatly facilitated by large-scale analyses of cancer cell immunopeptidomes.

Methods

Cell lines. Peripheral blood mononuclear cells (PBMCs) were isolated from blood samples of 18 volunteers. High-resolution HLA genotyping was performed at the Hôpital Maisonneuve-Rosemont using 500 ng of genomic DNA. B-LCLs were derived from PBMCs as described (27).

Proteogenomic identification of MAPs derived from B-LCLs. We applied our previously described proteogenomic approach to isolate and sequence MAPs. The methods of cell culture, transcriptome sequencing, mild acid elution, and MS have been described previously (27, 28). RNA-sequencing data were mapped using kallisto version 0.42.5 to ENSEMBL assembly 37.75 (NCBI Bioproject database http://www.ncbi.nlm.nih.gov/bioproject/; accession PRJNA286122) (74, 75). Transcriptome sequencing revealed no genetic polymorphisms in the regions coding for the mature (active) form of PSMB5 and PSMB8, the proteasome subunits that are mainly responsible for MAP processing (data not shown). We defined the B-LCL transcriptome as 10,575 expressed (averaged transcripts per million > 2) and annotated protein-coding genes. To mitigate the risk of false positives, stringent quality filters were applied to the list of identified MAPs: a peptide length of 8–14 amino acids; a 1% false discovery rate; and a predicted IC_{50} of 1250 nM or less. The binding affinity threshold was chosen to optimize inclusivity and stringency; a less stringent threshold of 5,000 nM included 8.6% more MAPs and 2.3% more source genes (Supplemental Figure 1C). When possible, binding affinities were predicted with NetMHC 3.4 (21 allotypes); otherwise, NetMHCcons 1.1 was applied (6 allotypes). For each individual, peptides were assigned to the allele with the strongest binding affinity. Peptides were mapped to proteins using pyGeno (29, 74). We applied further filtering steps to facilitate bioinformatic analysis; peptides assigned to more than one gene origin, transcripts with incomplete 5′ and 3′ annotation, and proteins with internal stop codons were all excluded. Where multiple isoforms were identified for a gene, MAPs were assigned to the most abundant transcript. Estimates of HLA allele frequency were derived from the European Caucasian population registered in the National Marrow Donor Program (33). The MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (PXD004023). In addition, the list of MAP sequences was deposited in the Immune Epitope Database (http://www.iedb.org/; 1000704). For IS-PRM analyses, nonsource peptides were chosen randomly from the lowest predicted quintile of nonsource genes generating a peptide that bound HLA-B*44:03 with an affinity IC_{50} of less than 50 nM. IS-PRM analyses for 2 sets of stable isotopically labeled synthetic peptides were performed as described (35).

Simulations of the redundancy in MAP and MAP source gene repertoires. HLA allotypes were randomly ordered, and either peptides or genes were considered. The number of nonredundant identifications was counted, considering the repertoires of each subsequent allotype. The simulation was repeated 1,000 times; average repertoire sizes are shown. The same simulation considering subjects instead of allotypes was also performed (Supplemental Figure 1, A and B). We noted greater redundancy in this simulation due to some subjects sharing the same allotypes.

Spatial localization of MAPs along source proteins. Every pair of overlapping MAPs was extracted for each protein generating more than 1 MAP. Overlapping MAP pairs were classified as sharing the same beginning "C-terminal extensions," sharing the same end "N-terminal extensions," being contained within another peptide, "Internal," or sharing at least 1 amino acid, "Overlap." Alleles presenting each peptide pair and their superfamilies were compared (34). Distances between adjacent MAP start sites on the same protein were computed for the actual distribution. For the random distribution, an equivalent number of MAPs was

randomly placed within the same protein length and adjacent distances between start sites computed. To estimate exome coverage, a window of 50 amino acids or 25 amino acids was moved residue by residue along each of the 10,575 proteins expressed in our B-LCLs; the number of MAPs seen in each window was counted.

Evaluating features of transcripts and proteins. To ensure the quality and relevance of our source and nonsource gene sets, we considered all genes expressed on average more than 2 transcripts per million. For each gene, the most expressed protein-generating transcript with complete HAVANA annotation and the corresponding protein were selected. Feature assembly was executed in Python version 2.7.10; pyGeno was used to extract transcript and protein sequences (29). Annotation translation was determined with the ENSEMBL BioMart extension (74). To calculate MAP density, NetMHC was used to predict the binding affinity of overlapping nonamers from each protein for all 27 allotypes expressed by the B-LCLs. NetMHC 3.4 was applied preferentially to predict binding affinities for 21 allotypes; NetMHCcons 1.1 was applied for the remaining 6 allotypes. The fraction of 9mers binding any of the 27 allotypes with an affinity of 1,250 nM or less was calculated for each protein.

B cell protein abundance in average spectral counts per gene evaluated by MS analysis of whole cell extracts was extracted from the Human Proteome Map (40). Genes with at least 1 transcript with an NMD biotype based on Vega annotation in ENSEMBL were considered to have NMD potential, that is, if a coding sequence finished more than 50 bp from a downstream splice site using any exon structure (http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html). uORFs were defined as nonoverlapping sequences within the 5′ UTR beginning with the cognate start codon AUG and ending with an in-frame stop codon. 5′ UTR secondary structure was predicted using RNAfold within the ViennaRNA Package version 2.1.7 (44). The percentage of AU-rich elements was defined as the number of A and/or U sequence of at least 5 nucleotides in length within the 3′ UTR. Stabilizing and destabilizing elements identified by Zhao et al. were queried in the 3′ UTR (48). TargetScan 7.0 was employed to predict microRNA-binding sites within the 3′ UTR (50). 3′ UTRs were prepared by removing ORFs; the number of nonoverlapping microRNA-binding sites was computed for all families of microRNAs. MicroRNA-binding sites were retrieved from the Molecular Signatures Database of GSEA (http://software.broadinstitute.org/gsea/msigdb/) and queried in all 3′ UTRs. To analyze the structural features of proteins, we used BioPython's package SeqUtils (specifically the ProtParam tool) to predict the proportion of residues conforming to a helix, turn, or sheet motif as well as the isoelectric point, instability index, and hydropathy for each protein sequence (76).

Protein degradation prediction softwares. Anaphase-promoting complex target sequences were predicted using GPS-ARM version 1.0 using default thresholds for D-box and KEN-box motif (52). PEST motifs were predicted using the function *epestfind* within EMBOSS version 6.5.7 (77). Ubiquitination sites were predicted with UbiProber (54) with a stringency of 70%. Three disorder prediction software programs were selected for the complementarity of their approaches: PONDR VLXT is a neural network predictor trained on missing residues in x-ray structures as well as known terminal and long disordered segments; DISOPRED, version 3.16, is a support vector machine and neural network predictor also trained on missing residues in x-ray structures; and IUPRED, version 1.0, is a biophysical model based on local interaction energies (78). Where residues were assigned to be disordered or not, disorder cutoff values were determined to equate the total disorder of the B-LCL proteome for PONDR-VLXT, DISOPRED, and IUPRED at 0.7, 0.3, and 0.5 respectively (Supplemental Figure 4A). Where lengths of N or C terminus disorder or internally disordered regions were computed, stretches up to 3 aa of ordered residues were allowed.

Data visualization. Graphics were made in R, version 3.2.2, using ggplot2, version 1.0.0 (https://cran.r-project.org/web/packages/ggplot2/index.html).

GO analysis. We compared either source or nonsource genes on a background of both groups using the R package topGO (60). The Fisher weight algorithm was used to reduce redundancies and compute *P* values.

Statistics. To generate Figure 1A, a binomial distribution of the probability of detecting between 0 and 50,000 peptides in a repertoire of 10,575 genes was computed. An exact binomial test was used to compare the expected distribution with the experimental values. A nonparametric effect size measure, Cliff's d, was used to compare enrichment of features in source and nonsource groups. The 95% CI was calculated based on 100 bootstraps using the orddom package, version 3.1, in R (https://cran.r-project.org/web/packages/orddom/orddom.pdf). Unless otherwise noted, we employed 2-sample Wilcoxon rank sum tests to compare continuous variables for robustness. *P* values of less than 0.05 were considered significant. All statistical analyses were performed in R, version 3.2.2.

Modeling. The variables listed in Supplemental Table 5 were used as input variables for logistic regression models run with the R packages caret and MASS (79, 80). The top 50 most enriched GO terms from the source and nonsource groups were included (Supplemental Tables 3 and 4). Near-zero variance parameters were excluded; this excluded the majority of GO terms. To limit the extent of correlation in input variables that can obscure their relative weight, further variables were excluded. Spearman's rank correlation coefficient *r* was calculated for each pair of input parameters, and a maximum absolute *r* of 0.6 was permitted (Supplemental Figure 3). As noted, input variables were also normalized by length of the appropriate UTR, transcript, or protein. The data were divided into training and testing sets containing 80% and 20% of genes, respectively. A logistic regression model with or without recursive feature elimination was built with centered and scaled training data using 10-fold cross-validation. The model then predicted the probability of generating MAPs for each gene in the testing set. Relative variable weight was computed based on the *t* statistic for all model parameters. An ordered logistic regression model with 3-class outcomes was built using the same protocol; categories were selected to optimize class balance (number of genes: 4,380, none; 3,164, low; 3,031, high). All metrics reported are averages of 1,000 iterations of data division and model building. An R script is included in Supplemental Methods (source code) that trains and applies the 2-class logistic regression model using the data frame in Supplemental Table 6 and that reproduces the panels of Figure 5.

Validation in independent data sets and human cancer cell lines. Transcriptomic data for JY (62) and 5 other human cancer cell lines (63) were combined with the respective immunopeptidomes described by other groups (21,36). Transcriptomic mapping was performed with kallisto, version 0.42.5, and the most expressed transcript for each gene was selected for analysis (75). Features of each gene and its gene products were annotated. Protein abundance was extracted from the

Human Proteome Map (40) defined by the closest matching tissue (B cells for the JY and SUP-B15 cell lines, CD4 cells for Jurkat cells, and adult colon for HCT116) or using cell line–specific data for HCC1143 and HCC1937 (81). The 2-class logistic regression model using all features trained using 10-fold cross-validation on all genes from 18 B-LCL samples was used to predict MAP generation in each cell line.

Study approval. This study was approved by the Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont (permit number CÉR 14095). Volunteers provided written, informed consent.

Author contributions

HP was responsible for conception and design, acquisition of data, analysis and interpretation of data, and drafting and revising the article. DPG, TD, CD, EB, MC, AR, JPL, CC were responsible for acquisition of data, analysis and interpretation of data, and revising the article. SL, PT, CP were responsible for conception and design, analysis and interpretation of data, and revising the article. SM was responsible for analysis and interpretation of data and revising the article.

Acknowledgments

This work was supported by grants from the Quebec Breast Cancer Foundation (to CP and SM) and from the Genome Canada Innovation Network (to PT). We are most grateful to our blood donors. CP and PT hold Canada Research Chairs in Immunobiology, and Proteomics and Bioanalytical Spectrometry, respectively. SM holds the CIBC Breast Cancer Research Chair at Université de Montréal. The CP lab is supported in part by the Katelyn Bedard Bone Marrow Association.

Address correspondence to: Pierre Thibault or Claude Perreault, Institute for Research in Immunology and Cancer, Université de Montréal, P.O. Box 6128, Station Centre-ville, Montréal, Quebec, Canada H3C 3J7. Phone: 514.343.6126; Email: pierre.thibault@umontreal.ca (P. Thibault); claude.perreault@umontreal.ca (C. Perreault).

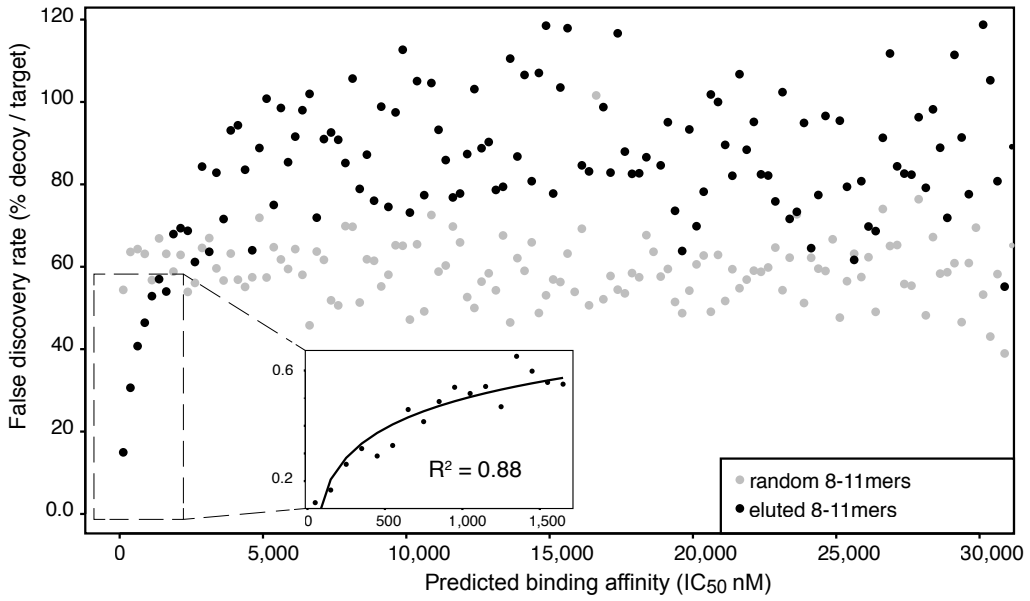
1. Granados DP, Laumont CM, Thibault P, Perreault C. The nature of self for T cells-a systems-level perspective. *Curr Opin Immunol*. 2015;34:1–8.
2. Govern CC, Paczosa MK, Chakraborty AK, Huseby ES. Fast on-rates allow short dwell time ligands to activate T cells. *Proc Natl Acad Sci U S A*. 2010;107(19):8724–8729.
3. Chakraborty AK, Weiss A. Insights into the initiation of TCR signaling. *Nat Immunol*. 2014;15(9):798–807.
4. Butler TC, Kardar M, Chakraborty AK. Quorum sensing allows T cells to discriminate between self and nonself. *Proc Natl Acad Sci U S A*. 2013;110(29):11833–11838.
5. Caron E, et al. The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol Syst Biol*. 2011;7:533.
6. Vrisekoop N, Monteiro JP, Mandl JN, Germain RN. Revisiting thymic positive selection and the mature T cell repertoire for antigen. *Immunity*. 2014;41(2):181–190.
7. Yewdell JW, Reits E, Neefjes J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat Rev Immunol*. 2003;3(12):952–961.
8. Hammer GE, Kanaseki T, Shastri N. The final touches make perfect the peptide-MHC class I repertoire. *Immunity*. 2007;26(4):397–406.
9. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999;50(3-4):213–219.
10. Kim Y, Sidney J, Buus S, Sette A, Nielsen M, Peters B. Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinformatics*. 2014;15:241.
11. de Verteuil D, Granados DP, Thibault P, Perreault C. Origin and plasticity of MHC I-associated self peptides. *Autoimmun Rev*. 2012;11(9):627–635.
12. Eisenlohr LC, Huang L, Golovina TN. Rethinking peptide supply to MHC class I molecules. *Nat Rev Immunol*. 2007;7(5):403–410.
13. Vigneron N, Van den Eynde BJ. Proteasome subtypes and the processing of tumor antigens: increasing antigenic diversity. *Curr Opin Immunol*. 2012;24(1):84–91.
14. Rock KL, Farfán-Arribas DJ, Colbert JD, Goldberg AL. Re-examining class-I presentation and the DRiP hypothesis. *Trends Immunol*. 2014;35(4):144–152.
15. Blum JS, Wearsch PA, Cresswell P. Pathways of antigen processing. *Annu Rev Immunol*. 2013;31:443–473.
16. Goodenough E, et al. Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3' UTR. *Proc Natl Acad Sci U S A*. 2014;111(15):5670–5675.
17. Antón LC, Yewdell JW. Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. *J Leukoc Biol*. 2014;95(4):551–562.
18. Apcher S, Millot G, Daskalogianni C, Scherl A, Manoury B, Fåhraeus R. Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. *Proc Natl Acad Sci U S A*. 2013;110(44):17951–17956.
19. Granados DP, et al. MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements. *Blood*. 2012;119(26):e181–e191.
20. Laumont CM, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun*. 2016;7:10238.
21. Caron E, et al. An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife*. 2015;4:e07661.
22. Hickman HD, et al. Toward a definition of self: proteomic evaluation of the class I peptide repertoire. *J Immunol*. 2004;172(5):2944–2952.
23. Mester G, Hoffmann V, Stevanović S. Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands. *Cell Mol Life Sci*. 2011;68(9):1521–1532.
24. Hassan C, et al. The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol Cell Proteomics*. 2013;12(7):1829–1843.
25. Hoof I, van Baarle D, Hildebrand WH, Keşmir C. Proteome sampling by the HLA class I antigen processing pathway. *PLoS Comput Biol*. 2012;8(5):e1002517.
26. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA class I alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity. *J Immunol*. 2013;191(12):5831–5839.
27. Granados DP, et al. Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat Commun*. 2014;5:3600.
28. Granados DP, et al. Proteogenomic-based discovery of minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers. *Leukemia*. 2016;30(6):1344–1354.
29. Daouda T, Perreault C, Lemieux S. pyGeno: A Python package for precision medicine and proteogenomics. F1000Research 2016;5(381). https://f1000research.com/articles/5-381/v2. Accessed October 31, 2016.
30. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007;4(3):207–214.
31. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res*. 2008;36(Web Server issue):W509–W512.
32. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*. 2012;64(3):177–186.
33. González-Galarza FF, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res*. 2015;43(Database issue):D784–D788.
34. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. *BMC Immunol*. 2008;9:1.
35. Gallien S, Kim SY, Domon B. Large-Scale Targeted Proteomics Using Internal Standard Triggered-Parallel Reaction Monitoring (IS-PRM). *Mol Cell Proteomics*. 2015;14(6):1630–1644.
36. Bassani-Sternberg M, Pletscher-Frankild S, Jensen LJ, 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*. 2015;14(3):658–673.
37. Weinzierl AO, et al. Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface. *Mol Cell Proteomics*. 2007;6(1):102–113.
38. Jovanovic M, et al. Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science*. 2015;347(6226):1259038.
39. Liu Y, Aebersold R. The interdependence of transcript and protein abundance: new data--new complexities. *Mol Syst Biol*. 2016;12(1):856.
40. Kim MS, et al. A draft map of the human proteome. *Nature*. 2014;509(7502):575–581.
41. Princiotta MF, et al. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity*. 2003;18(3):343–354.
42. Floor SN, Doudna JA. Tunable protein synthesis by transcript isoforms in human cells. *Elife*. 2016;5:e10921.
43. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A*. 2009;106(18):7507–7512.
44. Lorenz R, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*. 2011;6:26.
45. Szostak E, Gebauer F. Translational control by 3'-UTR-binding proteins. *Brief Funct Genomics*. 2013;12(1):58–65.
46. Schott J, Stoecklin G. Networks controlling mRNA decay in the immune system. *Wiley Interdiscip Rev RNA*. 2010;1(3):432–456.
47. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*. 2006;4(6):e180.
48. Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ. Massively parallel functional annotation of 3' untranslated regions. *Nat Biotechnol*. 2014;32(4):387–391.
49. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
50. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. *Elife*. 2015;4:e05005.
51. de Verteuil D, et al. Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules. *Mol Cell Proteomics*. 2010;9(9):2034–2047.
52. Liu Z, et al. GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. *PLoS One*. 2012;7(3):e34370.
53. Rechsteiner M, Rogers SW. PEST sequences and regulation by proteolysis. *Trends Biochem Sci*. 1996;21(7):267–271.
54. Chen X, Qiu JD, Shi SP, Suo SB, Huang SY, Liang RP. Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics*. 2013;29(13):1614–1622.
55. Prakash S, Tian L, Ratliff KS, Lehotzky RE, Matouschek A. An unstructured initiation site is required for efficient proteasome-mediated degradation. *Nat Struct Mol Biol*. 2004;11(9):830–837.
56. van der Lee R, et al. Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep*. 2014;8(6):1832–1844.
57. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins*. 2001;42(1):38–48.
58. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 2015;31(6):857–863.
59. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 2005;21(16):3433–3434.
60. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0. https://bioconductor.org/packages/release/bioc/html/topGO.html. Accessed October 5, 2016.
61. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol Cell Proteomics*. 2015;14(12):3105–3117.
62. Concha M, et al. Identification of new viral genes and transcript isoforms during Epstein-Barr virus reactivation using RNA-Seq. *J Virol*. 2012;86(3):1458–1467.
63. Klijn C, et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol*. 2015;33(3):306–312.
64. Benoist C, Germain RN, Mathis D. A plaidoyer for 'systems immunology'. *Immunol Rev*. 2006;210:229–234.
65. Longhi S. Structural disorder in viral proteins. *Protein Pept Lett*. 2010;17(8):930–931.
66. Murat P, Tellam J. Effects of messenger RNA structure and other translational control mechanisms on major histocompatibility complex-I mediated antigen presentation. *Wiley Interdiscip Rev RNA*. 2015;6(2):157–171.
67. Sansom SN, et al. Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Res*. 2014;24(12):1918–1931.
68. St-Pierre C, Trofimov A, Brochu S, Lemieux S, Perreault C. Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells. *J Immunol*. 2015;195(2):498–506.
69. Brennecke P, et al. Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat Immunol*. 2015;16(9):933–941.
70. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science*. 2015;348(6230):69–74.
71. Robbins PF, et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med*. 2013;19(6):747–752.
72. Yadav M, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*. 2014;515(7528):572–576.
73. Blankenstein T, Leisegang M, Uckert W, Schreiber H. Targeting cancer-specific mutations by T cell receptor gene therapy. *Curr Opin Immunol*. 2015;33:112–119.
74. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. *Genome Biol*. 2015;16:56.
75. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525–527.
76. Wilkins MR, et al. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol*. 1999;112:531–552.
77. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16(6):276–277.
78. Dosztányi Z, Mészáros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinformatics*. 2010;11(2):225–243.
79. Kuhn M, et al. caret: Classification and Regression Training. https://github.com/topepo/caret/. Published August 5, 2016. Accessed October 5, 2016.
80. Venables WN, Ripley BD. *Modern applied statistics with S*. Fourth edition. New York: Springer-Verlag; 2002.
81. Lawrence RT, et al. The proteomic landscape of triple-negative breast cancer. *Cell Rep*. 2015;11(4):630–644.

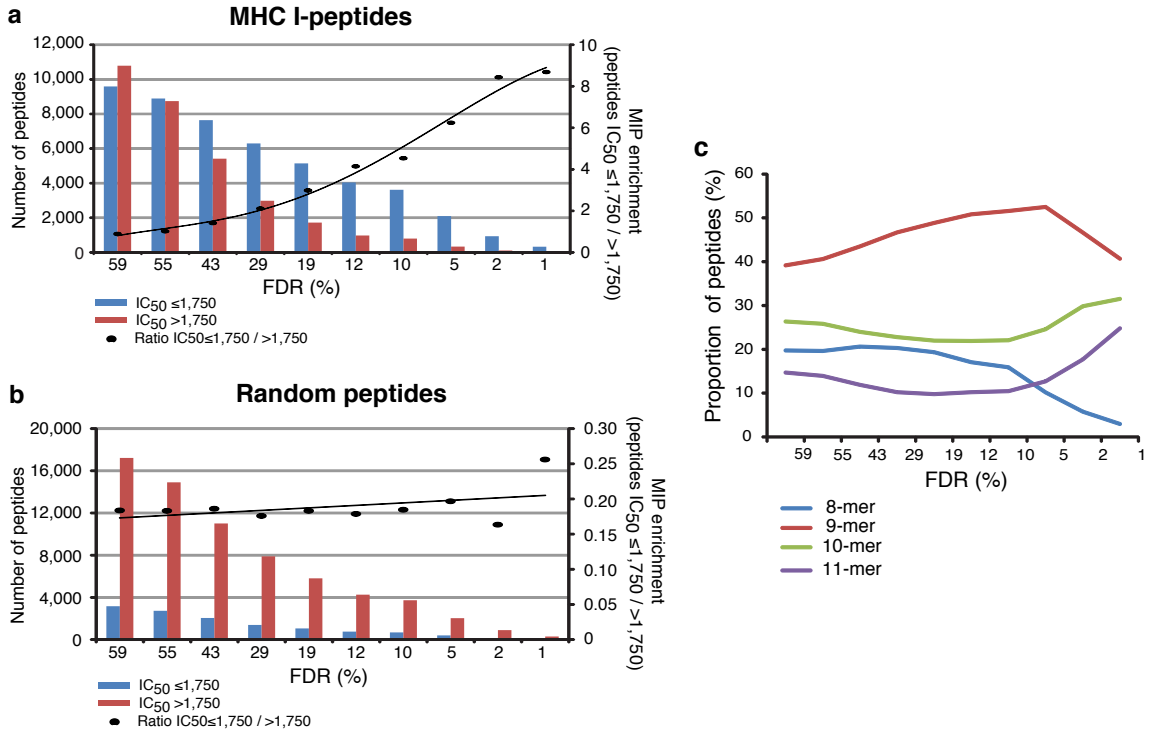
Annexe 2 – Supplementary Informations

Granados, D. P., Sriranganadane, D., Daouda, T., et al. (2014). Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nature communications*, 5.

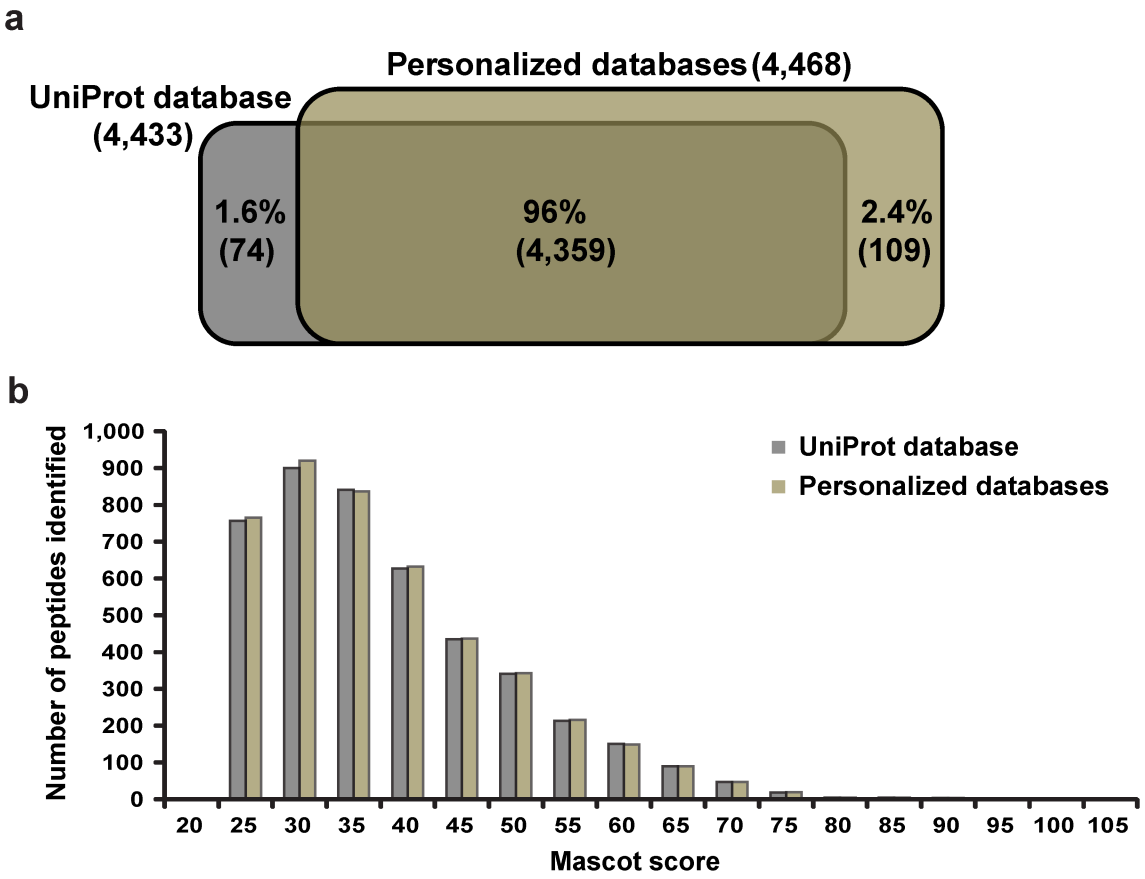
Supplementary Data 3, 4 and 4 can be found online at https://www.nature.com/articles/ncomms4600#supplementary-information



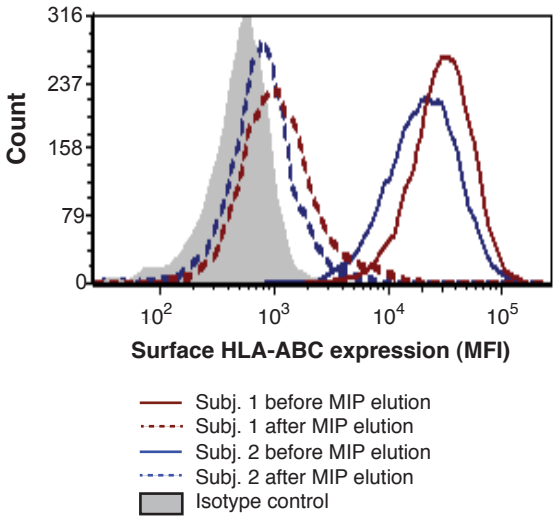
Supplementary Figure 1. Global false discovery rate (FDR) and predicted binding affinity allow discrimination between MIPs and contaminant (non-MIP) peptides. Predicted binding affinity to the relevant HLA molecules of 20,380 eluted peptides (black) and 20,380 random peptides generated from the personalized protein databases of subjects 1 and 2 (grey) was calculated using NetMHCcons. Binding score categories were generated by intervals of 500 nM (IC_{50}). Each dot represents the mean predicted binding affinity for peptides in a given bin. For each category, the level of accuracy in the peptide identification is shown as the FDR. The FDR (database of 20,319 target/12,011 decoy peptides identified by Mascot and defined by the clustering) was calculated for the eluted and the random peptides. The inset shows a high correlation between FDR values < 60% and PBA values < 1,750 nM. The distribution of random peptides shows no correlation between the predicted binding affinity and the FDR.



Supplementary Figure 2. The global false discovery rate (FDR) allows enrichment of MIPs and affects the proportion of small (8-9mers) and long peptides (10-11mers) identified. (**a**) We calculated the predicted binding affinity (IC50) for the 8-11mer peptides obtained after applying different FDR thresholds. FDR values were calculated from a dataset of 20,319 eluted peptides (target) and 12,011 reverse peptide versions (decoy). Without filtering any 8-11mer peptide identified by Mascot, the FDR value corresponds to 59% (maximal FDR). Bars show the number of peptides with a predicted binding affinity ≤ 1,750 nM (blue) or > 1,750 nM (red). The second y-axis shows the enrichment in MHC I-peptides calculated as the ratio of peptides with a predicted binding affinity ≤1,750 / >1,750 nM. (**b**) We performed the same analysis by randomly generating the same amount of peptides for each FDR threshold. The figure shows that lower FDR thresholds increase the proportion of eluted peptides with high predicted binding affinity (a) but have no impact on random peptides (b). (**c**) Proportion of 8 – 11mers identified by applying different FDR thresholds. Low FDR values favor the identification of long peptides and disfavor the identification of short peptides.

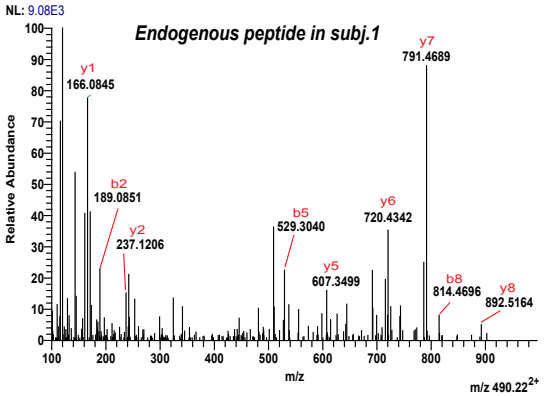


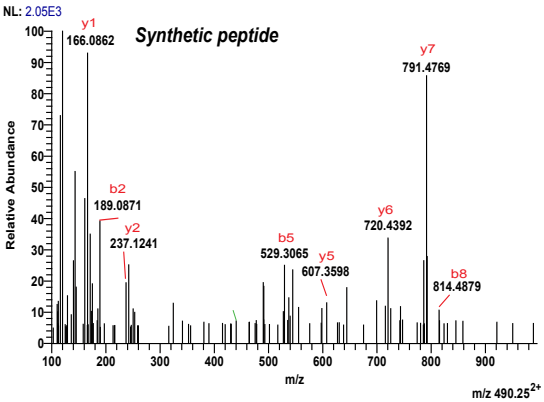
Supplementary Figure 3. Comparison of MIPs identified using UniProt vs. personalized databases built with next generation sequencing data. (**a**) Venn diagram comparing the number and percentage of unique and common MIP sequences found using UniProt vs. personalized databases. (**b**) Mascot score of MIPs identified with each database.



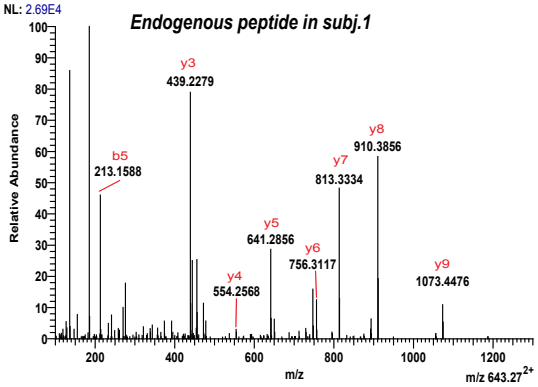
Supplementary Figure 4. Quantification of surface HLA-ABC before and after peptide elution. B-LCLs from subjects 1 and 2 were stained with PE anti human HLA-ABC monoclonal antibody (clone W6/32, Cedarlane) or the corresponding isotypic control and the mean fluorescence intensity (MFI) was analyzed by flow cytometry before or after mild acid elution of peptides. The histogram shows similar levels of HLA-ABC surface expression before and after peptide elution in representative samples of subjects 1 and 2.

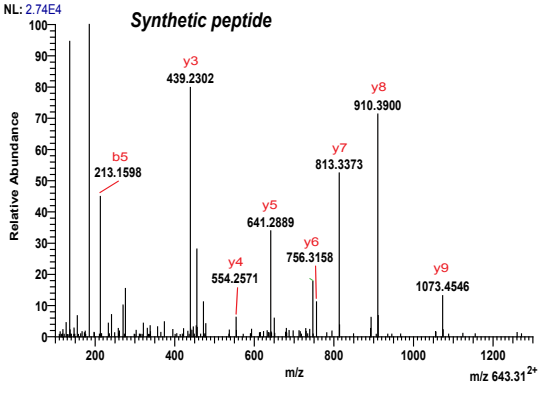
STALRLTAF_*ITGAL*



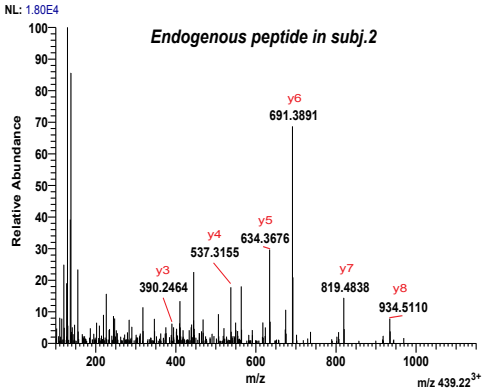


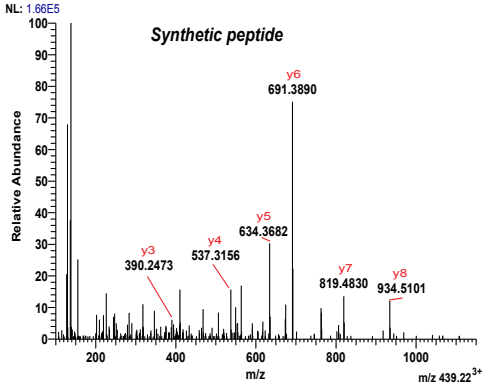
VIYPGDSDTRY_*IGHV5*



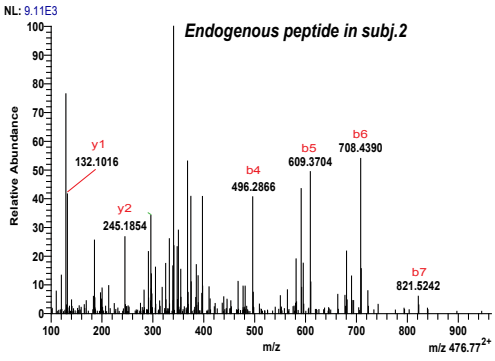


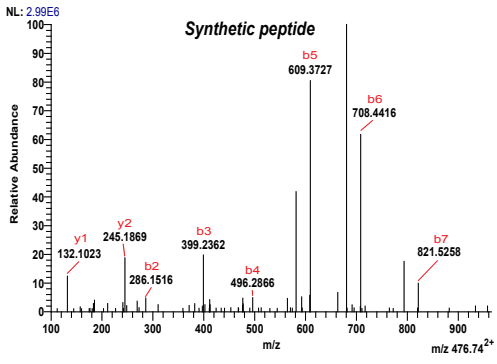
AMYDKGPFRSK_*NQ01*



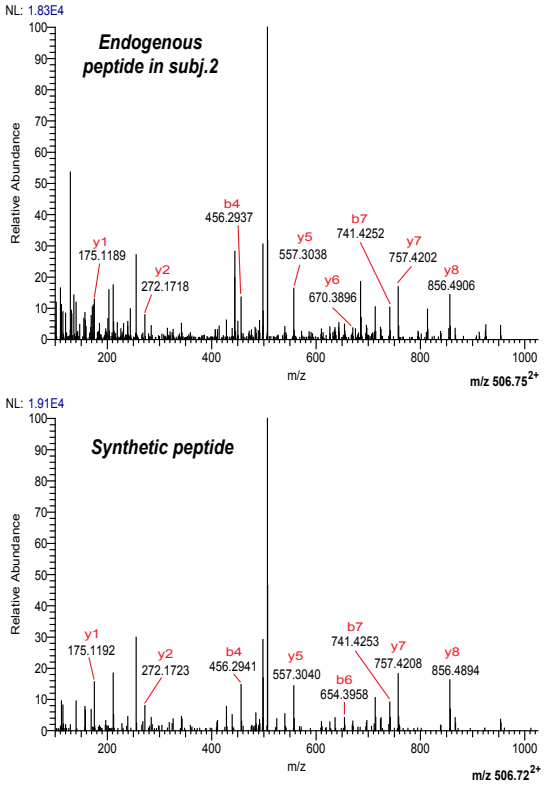


RELPLVLL_*GRP*

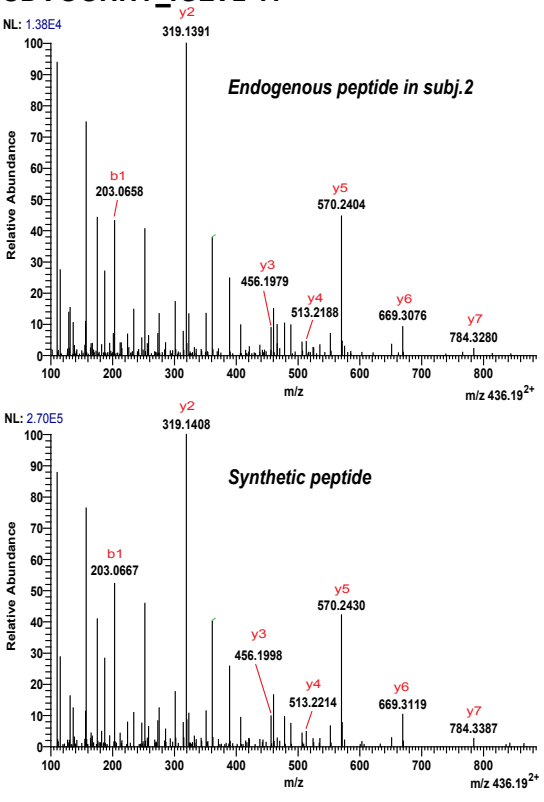




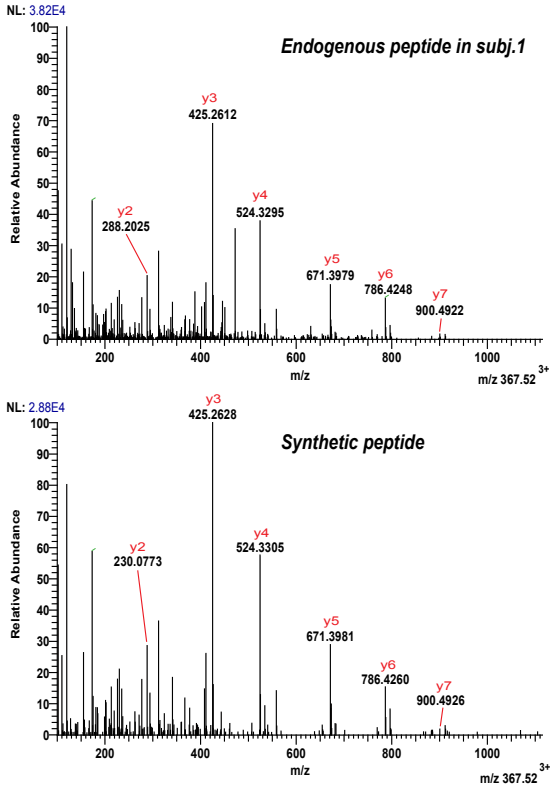
RVSLPTSPR_*C13Orf18*



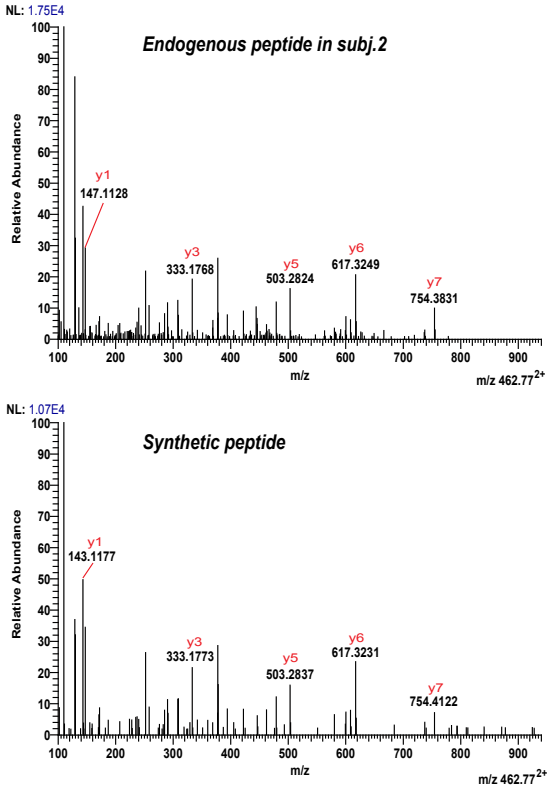
SDVGGHHY_*IGLV2-11*



AENDFVHRI_*R3HCC1*

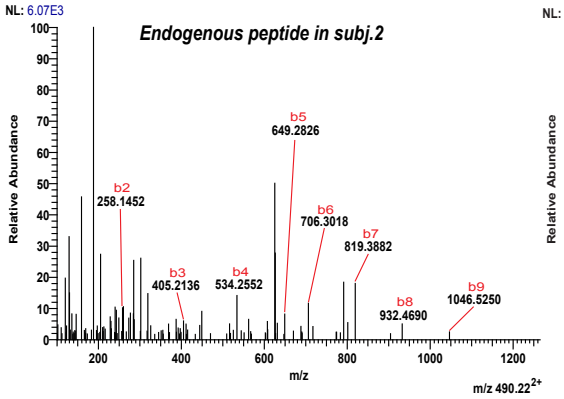


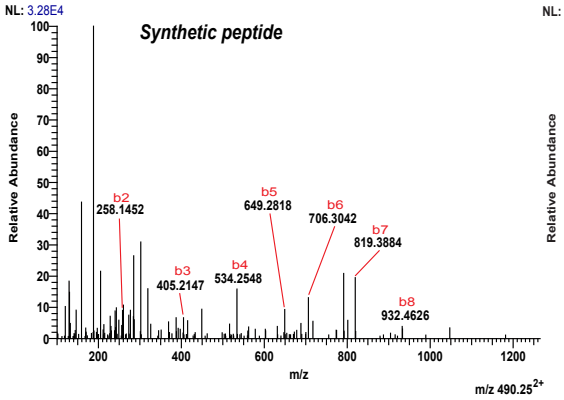
AVHNGLGEK_*NADK*



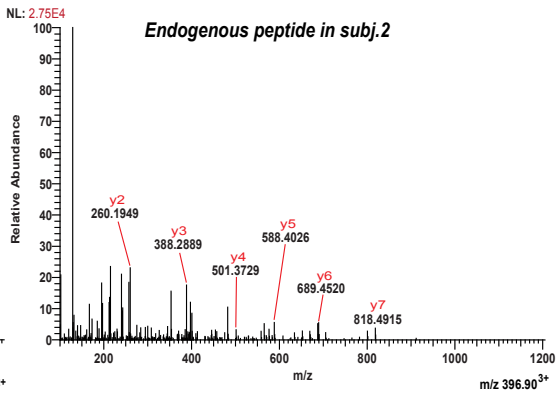
5a

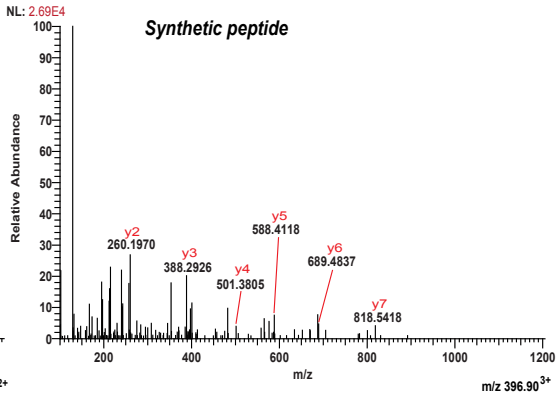
KEFEDGIINW_*BCL2A1*





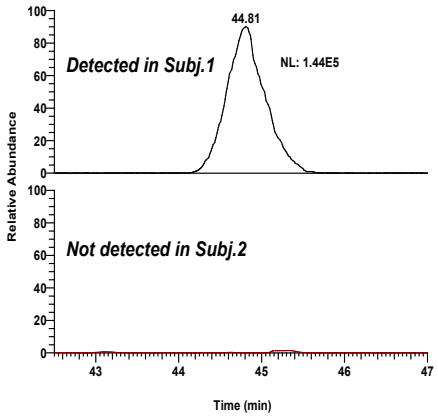
QELETSIKKI_*KIF20B*

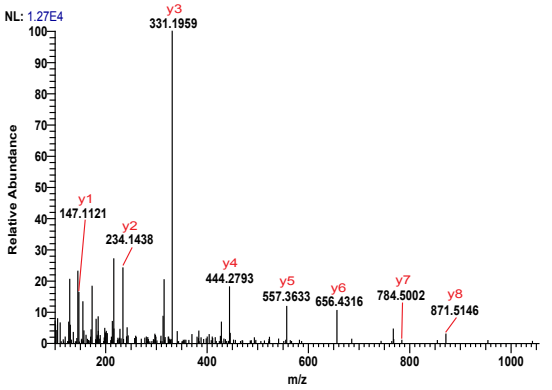




5a

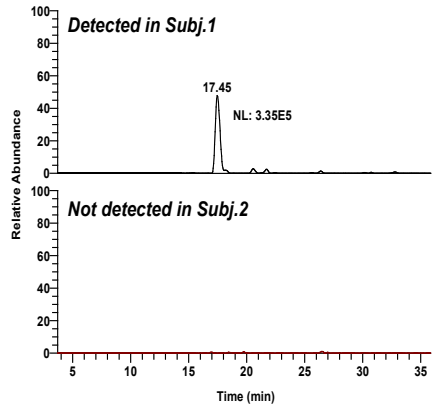
ATSQVLLPSK_*IGHM*

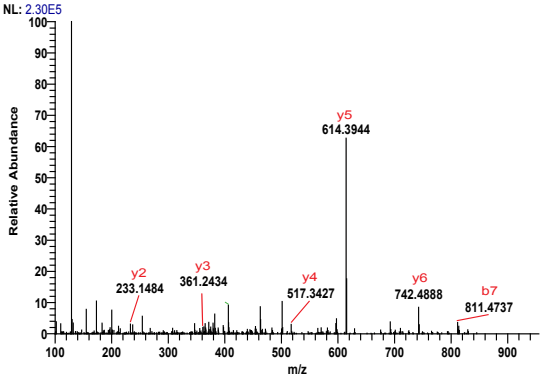




Fragment mass accurracy < 0.02Da

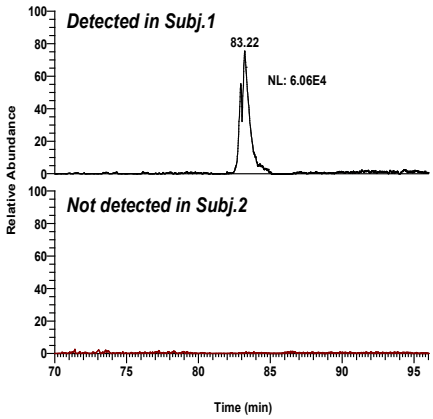
EAKPRKTL_*FAM21C*

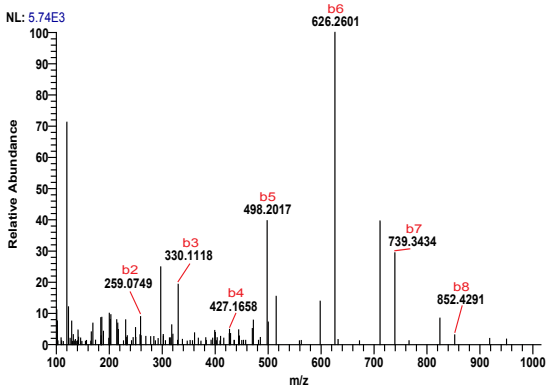




Fragment mass accurracy < 0.02Da

EEAPAQLLQ_*LRRC37B*

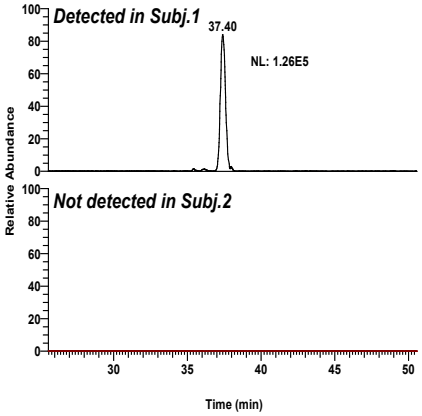


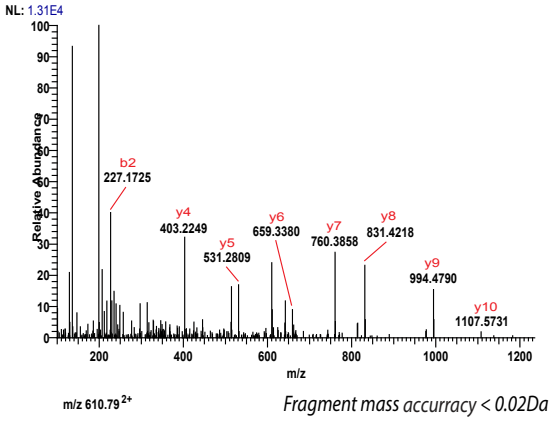


Fragment mass accurracy < 0.02Da

5b

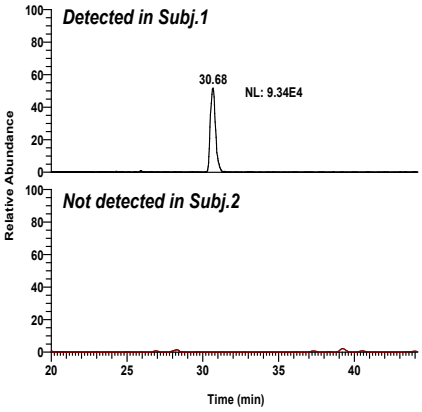
LLYATQQGQAK_*MTRR*

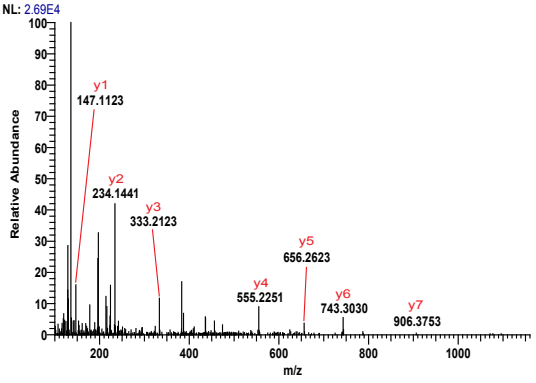




Fragment mass accurracy < 0.02Da

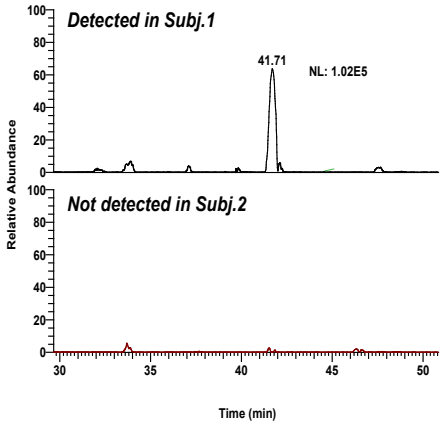
QIYSTC(cyst)VSK_*KIF21B*

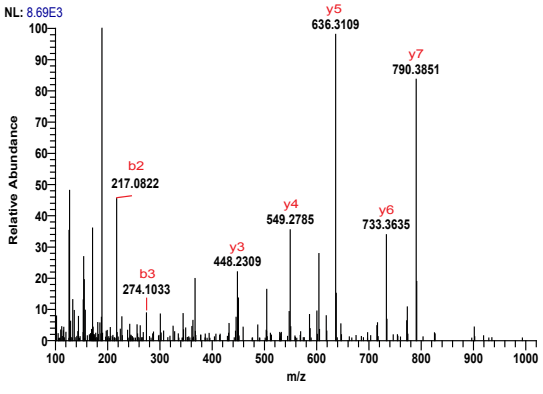




Fragment mass accurracy < 0.02Da

SEGPSTSRW_*FOXM1*

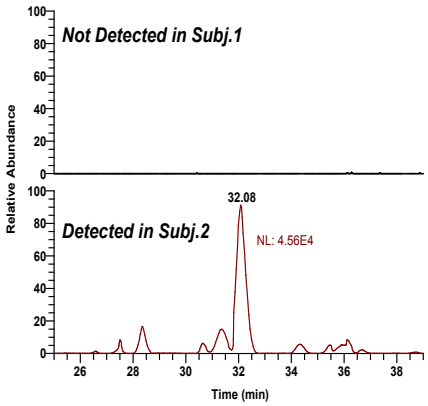


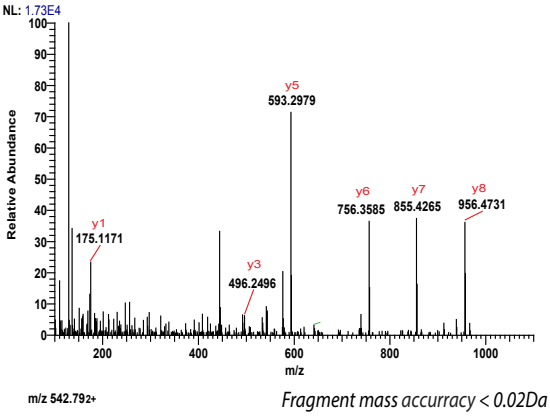


Fragment mass accurracy < 0.02Da

5b

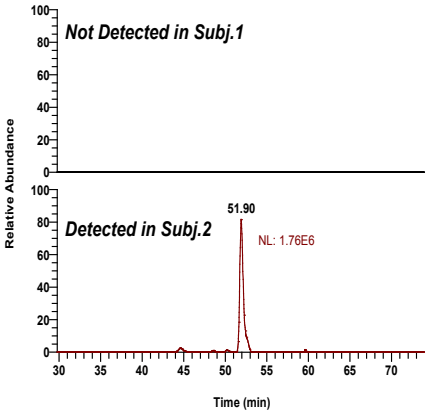
KTVYPASYR_*CDT1*

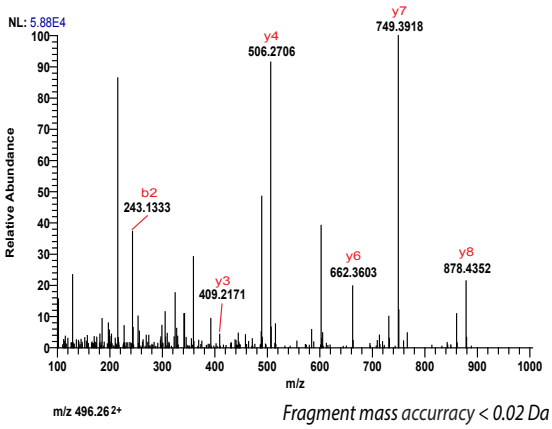




Fragment mass accurracy < 0.02Da

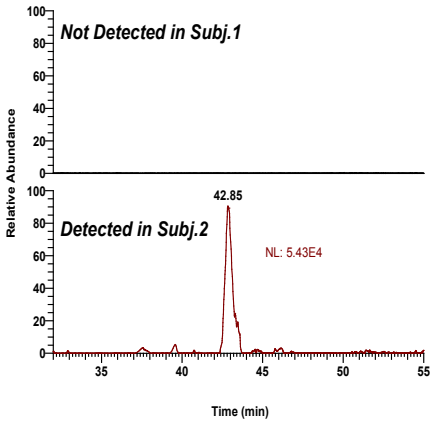
LESGVPSRF_*IGKV1D*

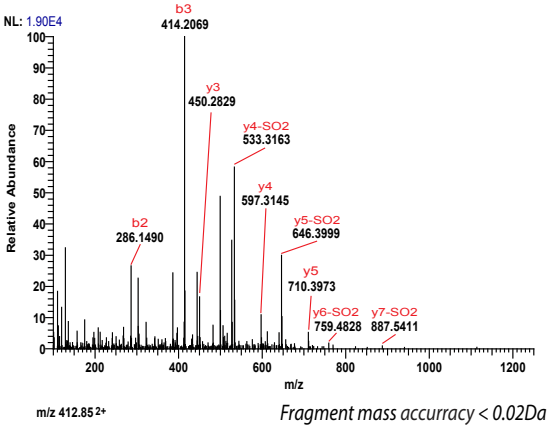




Fragment mass accurracy < 0.02 Da

REQILM(ox)KRF_*PWP2*

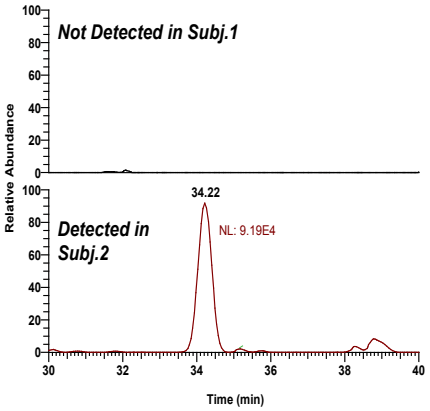


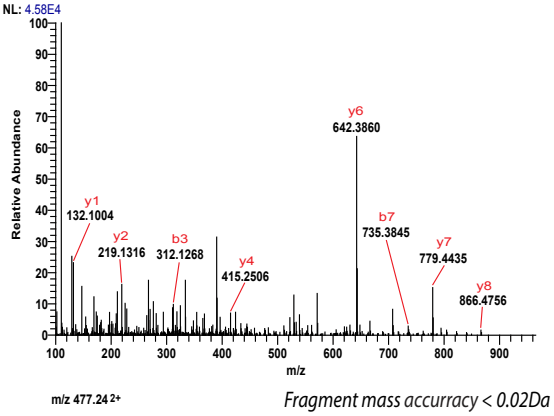


Fragment mass accurracy < 0.02Da

5b

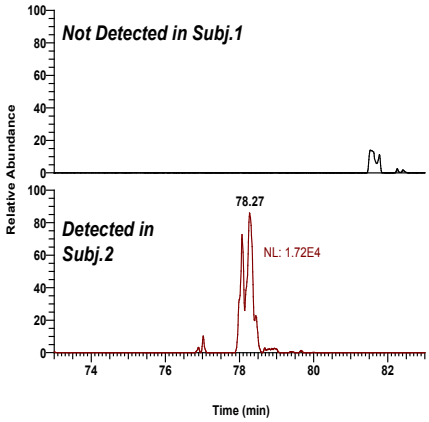
SSHQVVPSL_*PTPRZ1*

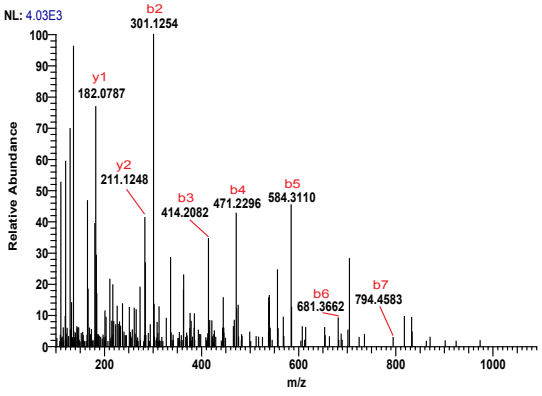




Fragment mass accurracy < 0.02Da

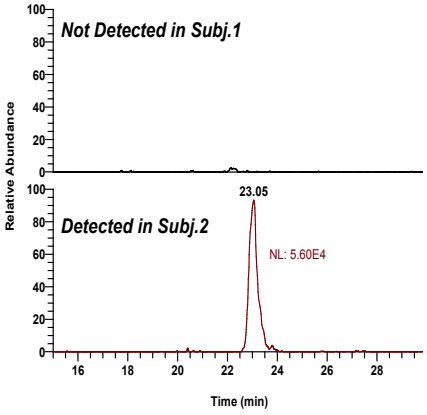
YHIGIPLTY_*VPS16*

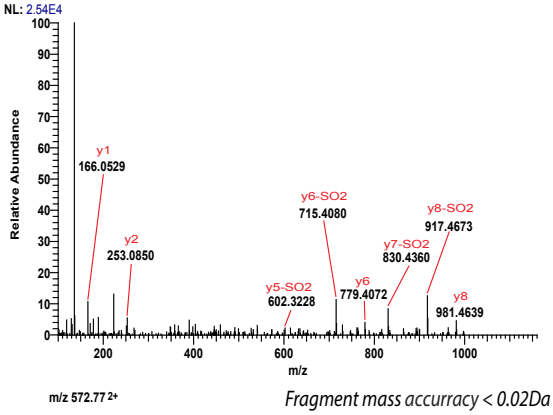




Fragment mass accurracy < 0.02Da

YSDLRKESM(ox)_*MCM2*

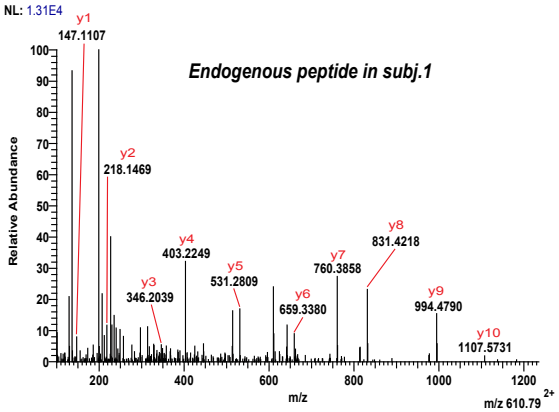


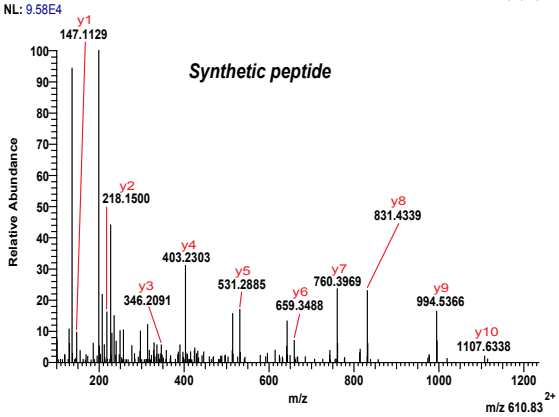


Fragment mass accurracy < 0.02Da

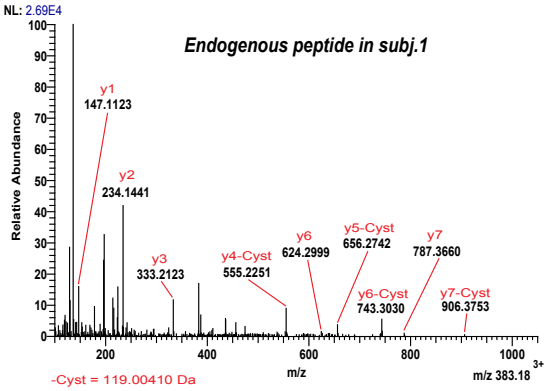
5b

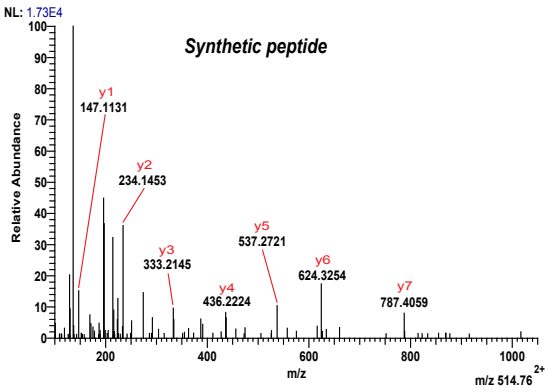
LLYATQQGQAK_MTRR



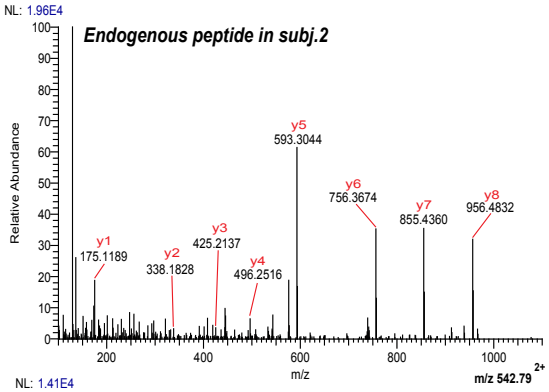


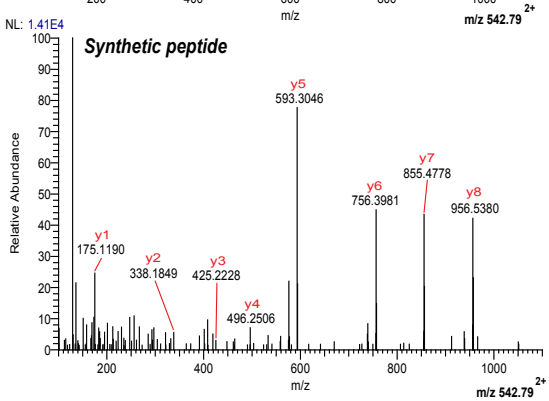
QIYSTC(cyst)VSK_KIF21B





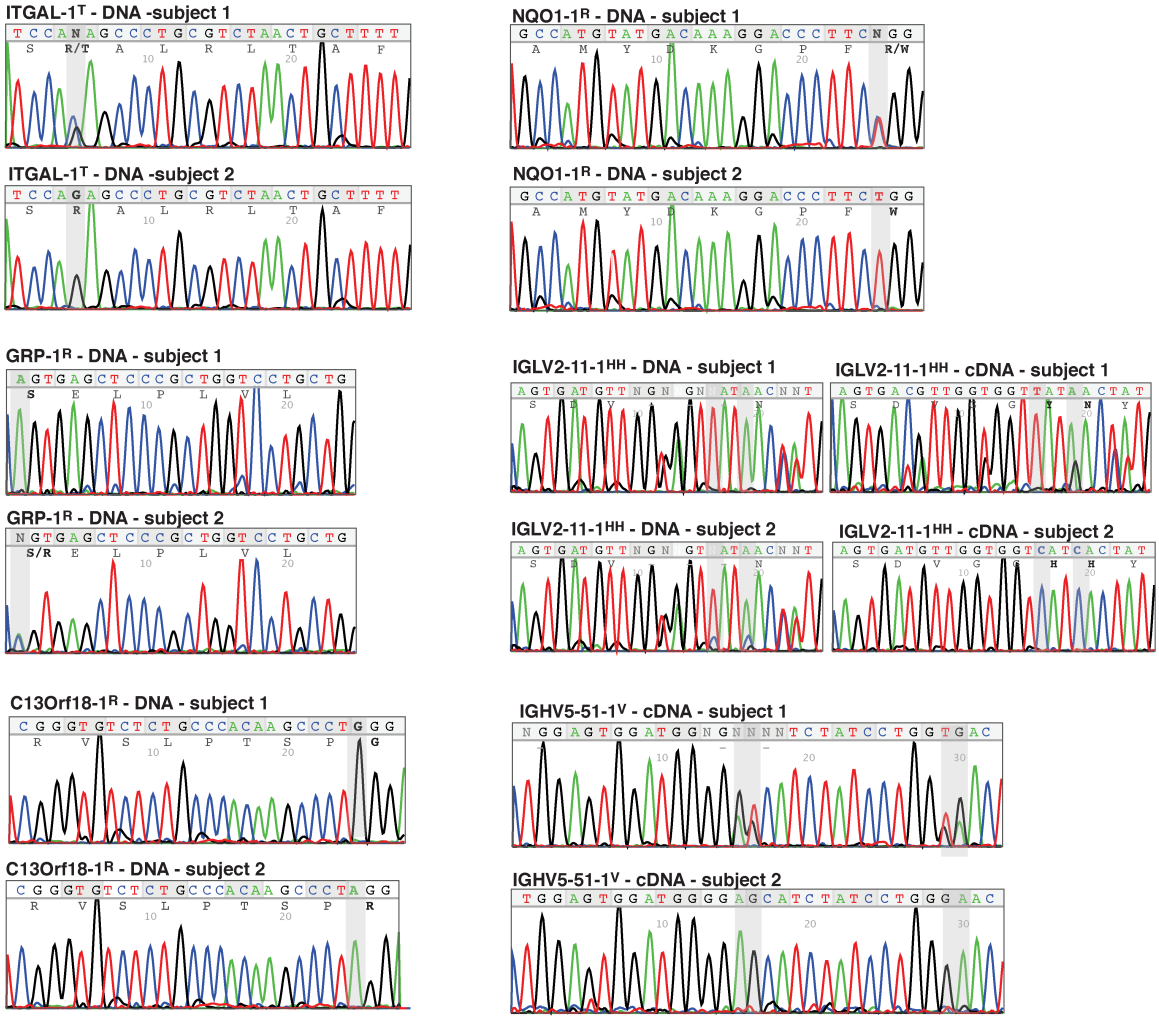
KTVYPASYR_*CDT1*



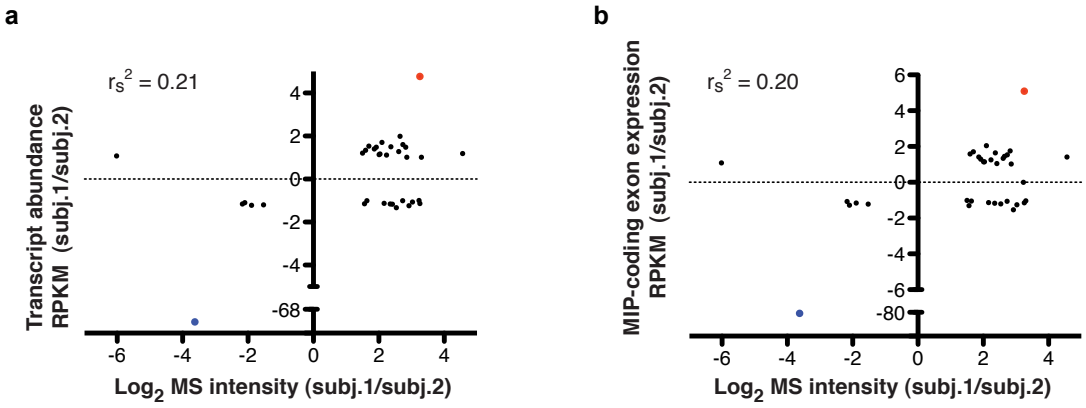


5c

Supplementary Figure 5. MS validation of polymorphic and non polymorphic peptides. (**a**) Polymorphic MiHAs resulting from ns-SNPs in the MIP-coding region, encoded by one single non-HLA gene and exclusively detected in one of the two subjects (corresponding to Table 1) were validated using the MS/MS spectra of the corresponding synthetic peptides. (**b**) Twelve peptides exclusively detected in subject 1 or 2 (6 in subject 1 and 6 in subject 2) were randomly selected among those that were present in 3 or 4 replicates. The extracted ion chromatograms of MiHA peptides for both subjects are presented along with the corresponding MS/MS spectra. Comparison of extracted ion chromatograms conrmed the selective detection of MiHA peptides in one of the two subjects. (**c**) Three non polymorphic peptides (Supplementary Data 3) used in cytotoxicity assays were also validated using MS/MS of their respective synthetic peptides.



Supplementary Figure 6. Validation of 6 MiHA-coding sequences by Sanger sequencing. Chromatograms obtained after Sanger sequencing of PCR-amplified DNA or cDNA encoding 6 MiHAs. The primers used are shown in Supplementary Table 1. Polymorphic loci are highlighted in grey. The IGLV2-11-1HH MiHA results from 2 nucleotide changes at the transcript level in subject 2.



Supplementary Figure 7. Differential expression of non-polymorphic MIPs does not correlate with differences in MIP-coding genes or exons between subjects. For 41 non-polymorphic MIPs that were exclusively detected in one sibling, we calculated the fold difference in intensity of the MIP and the fold difference in expression of the underlying MIP-coding gene (**a**) or exon (**b**) measured in Reads Per Kilobase per Million mapped Reads (RPKM). In only 2 cases, MIP abundance differences reflect MIP-coding transcript differences (dots in red and blue). The calculated Spearman r value (rs^2) shows no correlation.

Supplementary Table 1. List of PCR primers used for PCR amplification of DNA and cDNA segments containing the region coding for the MiHAs between subjects 1 and 2.

Peptide sequence (amino acids)	Gene symbol	Target nucleic acid	Forward (5' to 3')	Reverse (5' to 3')
STALRLTAF	ITGAL	DNA	TTATTTCTTTCTGGCCCACCA	AGCATCTTCTTCCAAGTTACTCAG
STALRLTAF	ITGAL	cDNA	GAAACCTGGGAGATCCCTTT	TCACATTGGCGTGCAATTC
VIYPGDSDTRY	IGHV5-51	DNA and cDNA	ACAGTAATACATGGCGGTGTC	GCTGTTCTCCAAGGTCAGTC
AMYDKGPFRSK	NQO1	DNA	AGGAATGGGAAAGGTGTGAAG	GGGAAGCTCCATCTCAAACAA
AMYDKGPFRSK	NQO1	cDNA	GCCCAGATATTGTGGCTGAA	GAAGCCACAGAAATGCAGAATG
RELPLVLL	GRP	DNA	TCTGCTCTTCCCAGCCTCTC	GCCAGGGAAACGCAAAGAAATG
RVSLPTSPR	C13orf18	DNA	TCTGAGGATACCACAGACTCC	GGTGTGAACAGAGAGGAATGAG
RVSLPTSPR	C13orf18	cDNA	CAACGTTGTCTGAGGATACCAC	GCACAAATACCTCTGGTGAGAA
SDVGGHHY	IGLV2-11	DNA and cDNA	GAGTGTGTTTCTCCCTCTTTCC	CCTGCATATGAGCAGCAGTAA
SDVGGHHY	IGLV2-11	DNA and cDNA	TGATCCTTGGTCTCCTGCT	GAAAGTGTAGCTGCCTGCATA

Supplementary Data 1. Exome and transcriptome sequencing and mapping statistics

Parameter	RNA-seq		Exome capture	
	subject 1	subject 2	subject 1	subject 2
Total number of reads	70877244	71142828	62794394	25556222
Number of reads passing Ilumina's quality filter	68933712	68672060	61549102	25080214
% reads passing filter	97.3	96.5	98.0	98.1
Number of mapped reads	54509785	54323949	50500441	20789030
% mapped reads	79.1	79.1	82.1	82.9
Number of reads in genes	50798312	51384167	41526778	17144575
% reads in genes	93.2	94.6	82.2	82.5
Mean RPKM in genes	16.85	17.4		
Number of reads in introns	4390772	4120751		
Number of reads in exons	46407540	47263416	31833586	13225156
% reads in exons	85.1	87.0	63.0	63.6
Mean annotated (UCSC, Hg19) exome coverage (X) (>= 5 reads)	131	130		
Number of reads in TruSeq targets			31757236	13196324
% of TruSeq targets covered			98.8	98.1
Mean Truseq targets exome coverage (X) (>= 5 reads)			158	66
Mean RPKM in exons	18.53	19.27		

Number and % of bases that were covered by at least 5 reads. Reads that mapped to intergenic and intronic regions are not excluded in the statistics, with the exception of "annotated exome statistics".

Parameter	subject 1	subject 2
Total number of bases covered	147473522	118075139
Number of bases covered only by exome-seq	85326528	58635152
% of bases covered only by exome-seq	57.9	49.7
Number of bases covered only by RNA-seq	24080010	24856658
% of bases covered only by RNA-seq	16.3	21.1
Number of bases covered by both exome-seq and RNA-seq	38066984	34583329
% of bases covered by exome-seq and RNA-seq	25.8	29.3
Total number of bases of the annotated exome (UCSC, Hg19) covered	53522898	50010572
% of bases of the annotated exome (UCSC, Hg19) covered	81.13	75.8

Supplementary Data 2. Comparison of synonymous and non-synonymous SNPs identified in the genome of these siblings (all regions included: introns, exons, UTRs, etc.) by exome capture and/or RNA-seq vs. SNPs reported in

Parameter	Subject 1		Subject 2	
	Number	%	Number	%
Total number of SNPs **(synonymous and non-synonymous)**	171298	100.0	140772	100.0
Detected by exome capture	128752	75.2	98417	69.9
Detected by RNA-seq	81183	47.4	80053	56.9
Detected by both exome capture and RNA-seq	38637	22.6	37698	26.8
Detected only by exome capture	90114	52.6	60719	43.1
Detected only by RNA-seq	42546	24.8	42355	30.1
Detected by exome capture and in dbSNP	126104	97.9	96105	97.7
Detected by RNA-seq and in dbSNP	65354	80.5	64242	80.2
Detected by exome capture and/or RNA-Seq and in dbSNP	153192	89.4	123159	87.5
Detected by both exome and RNA-seq and in dbSNP	38166	98.8	37188	98.6

Comparison of ns-SNPs that were included in the personalized protein data after quality filtering and combining variant calling by exome capture and known SNPs (dbSNP, buid 137)

Comparison	Total number non-synonymous SNPs	Non-synonymo
		Number
Subject 1 vs. Reference	8648	8452
Subject 2 vs. Reference	11644	10848
Subject 1 vs. Subject 2	4833	4203
Subjects 1 and/or 2 vs. Reference	11660	10861

abases (identified
by RNA-Seq) with

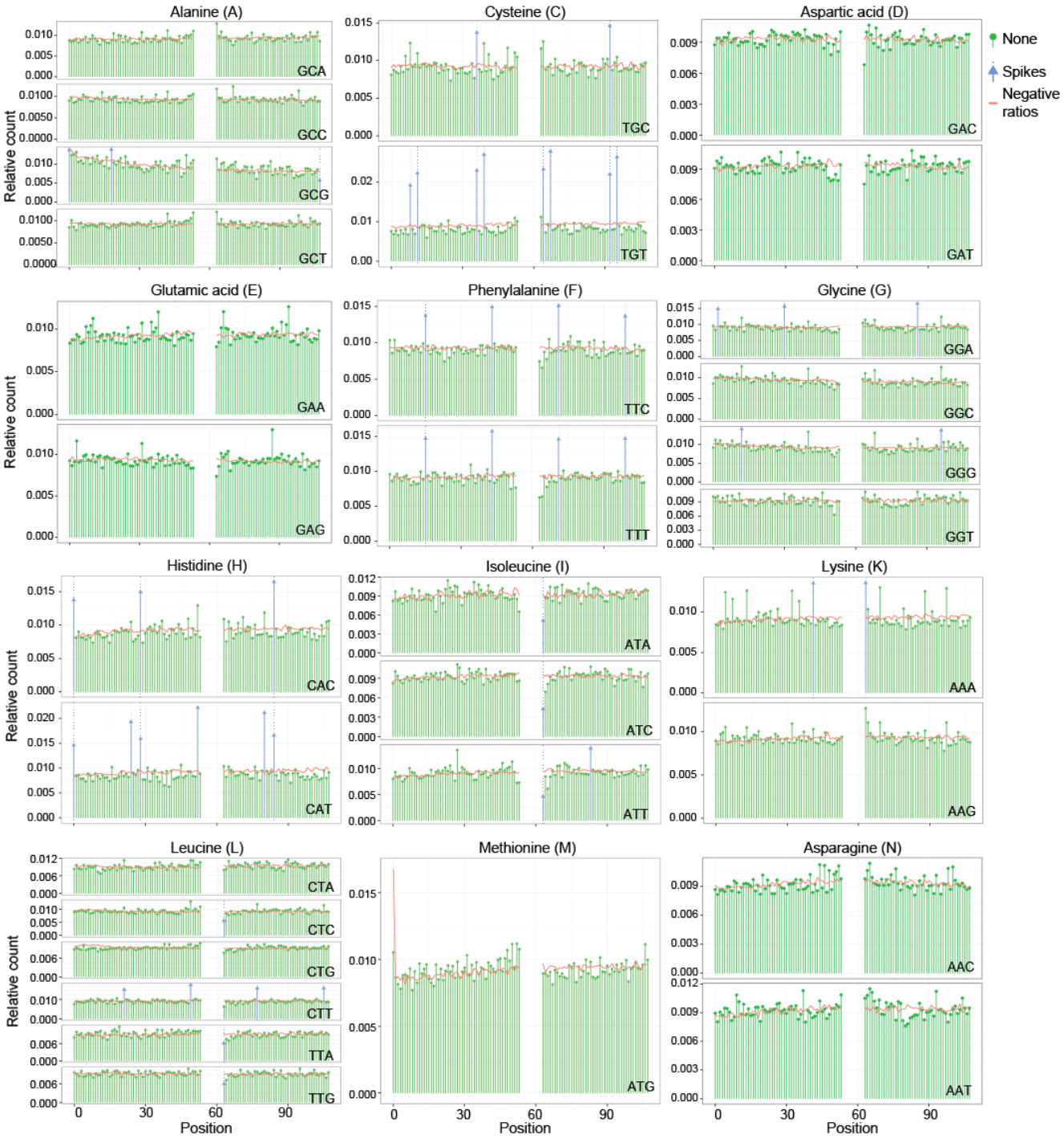
ous SNPs in dbSNP
%
97.7
93.2
87.0
93.1

Annexe 3 – Supplementary Informations

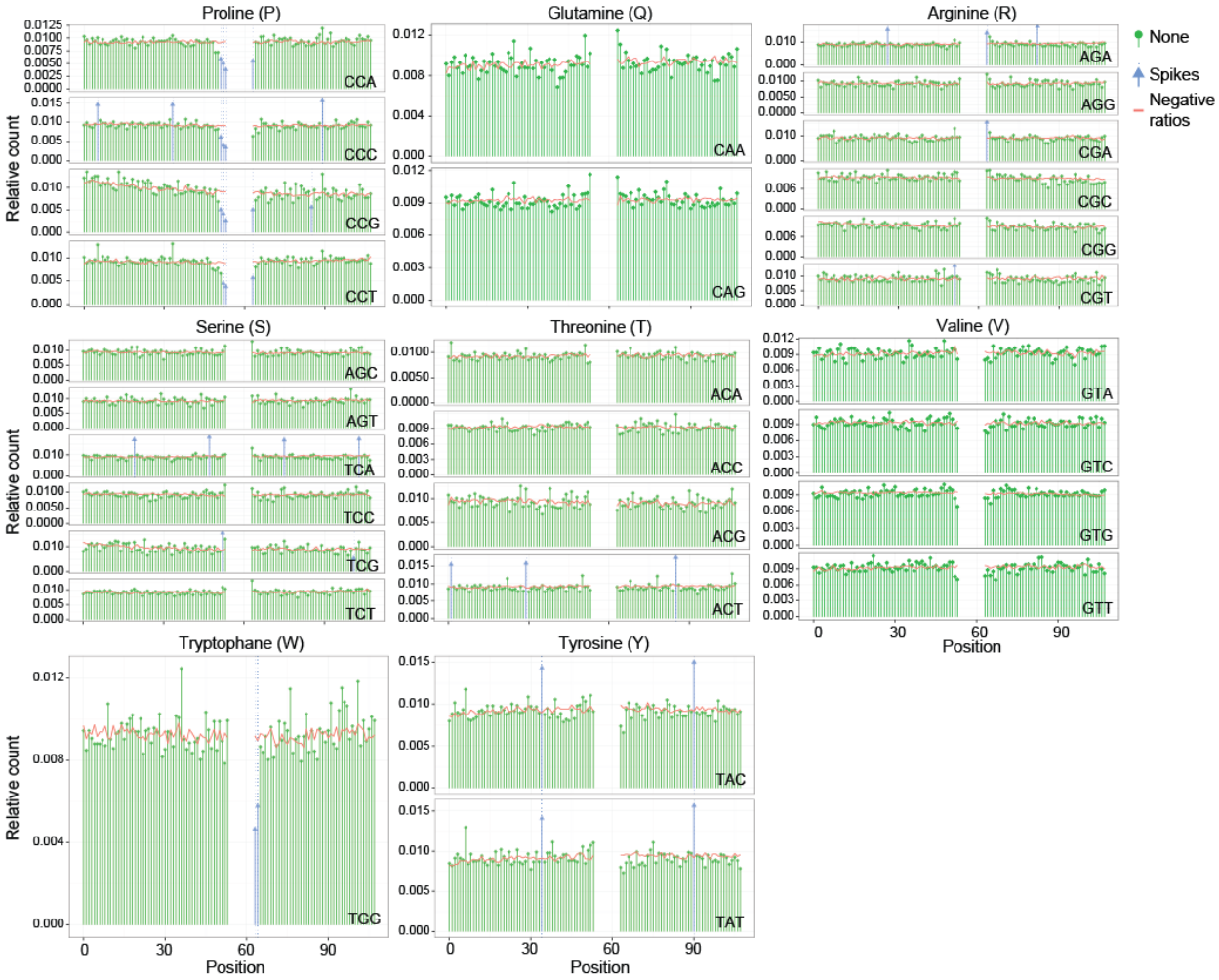
Daouda, T., Dumont-Lagacé, M., Thibault, P., Bengio, Y., Lemieux, S., Perreault, C. Codon usage regulates biogenesis of human MHC class I-associated peptides. Submitted to *Cancer Immunology Research*, 2017.

Supplementary Data

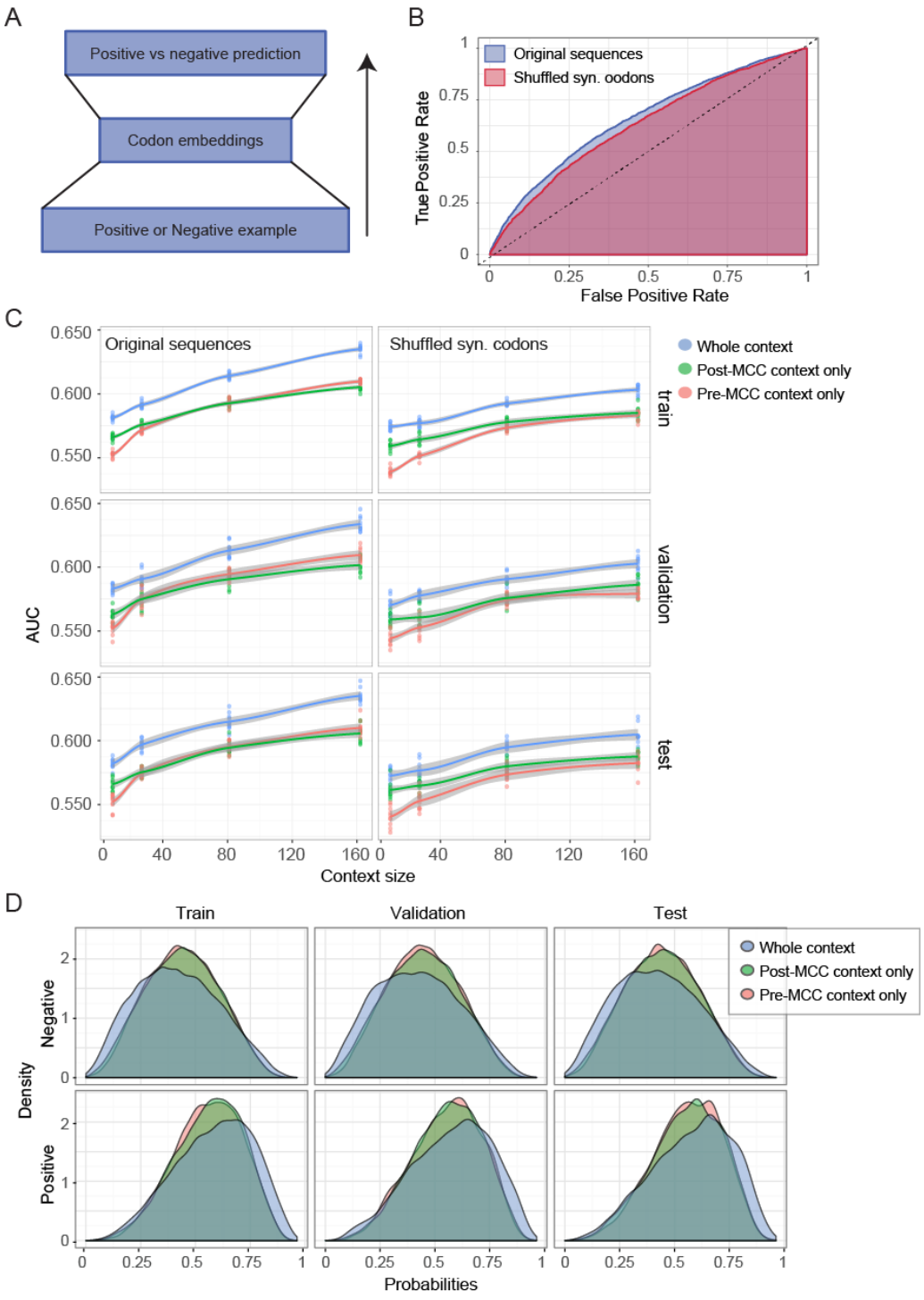
Supplementary Movie 1. Time-lapse of ANN preferences presented on the embedding space. Each frame in the video corresponds to an increase in the position. ANN preference is represented by the dot sizes.



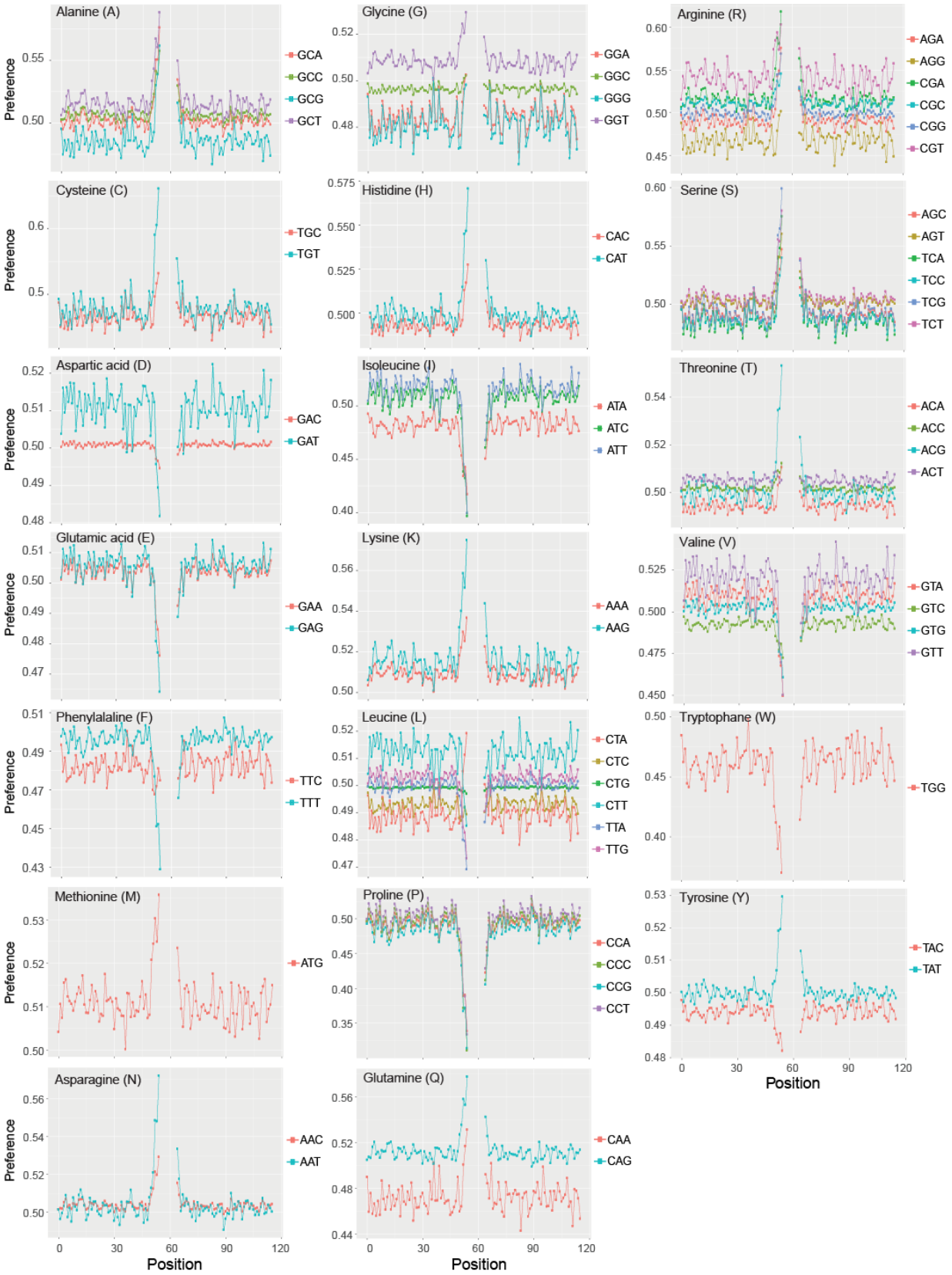
Supplementary Figure 1. Distribution of alanine, cysteine, aspartic acid, glutamic acid, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine and asparagine codons in positive and negative datasets. Spikes and drops are represented in blue.



Supplementary Figure 2. Distribution of proline, glutamine, arginine, serine, threonine, valine, tryptophan and tyrosine codons in positive and negative datasets. Spikes and drops are represented in blue.



Supplementary Figure 3. ANN architecture and detailed predictions. (A) Architecture of the ANNs used in this work. (B) ROC curves for an ANN trained on a context size of 162 nucleotides on original sequences or sequences with shuffled synonyms. (C) Results for the AUC on all train, validation and test subsets. Grey areas represent the 95% confidence intervals. (D) Distributions of output probabilities of the ANNs used to calculate correlations in Figure 3D.



Supplementary Figure 4. ANN preferences for all codons.

Annexe 4 - Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames.

Laumont, C. M., Daouda, T., Laverdure, J. P., Bonneil, É., Caron-Lizotte, O., Hardy, M. P., ... & Perreault, C. (2016). *Nature communications*, 7, 10238.

nature COMMUNICATIONS

ARTICLE

Received 13 Apr 2015 | Accepted 16 Nov 2015 | Published 5 Jan 2016

DOI: 10.1038/ncomms10238 OPEN

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

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

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

[1] Institute for Research in Immunology and Cancer, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7. [2] Department of Medicine, Faculty of Medicine, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7. [3] Department of Computer Science and Operations Research, Faculty of Arts and Sciences, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7. [4] Department of Chemistry, Université de Montréal, PO Box 6128 Station Centre-Ville, Montreal, Quebec, Canada H3C 3J7. [5] Division of Hematology, Hôpital Maisonneuve-Rosemont, 5415 de l'Assomption Boulevard, Montreal, Quebec, Canada H1T 2M4. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.P. (email: claude.perreault@umontreal.ca).

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

Proteogenomics has rapidly revolutionized our vision of the proteome of cells from numerous living organisms, including normal and neoplastic human cells[4,5]. A fundamental issue tackled by proteogenomics is the landscape of genomic regions that are expressed at the protein level. Ribosome-profiling experiments have provided strong evidence for pervasive translation outside of annotated protein-coding genes[6]. However, the definite proof of a genomic locus being protein-coding is the detection of its corresponding protein[7]. Accordingly, one salient concept emerging from proteogenomic analyses is that the proteome is more complex than previously thought. The proteome contains peptides arising from a variety of RNAs that were not supposed to encode proteins (noncoding RNAs) and are therefore not included in annotated protein databases. Many long noncoding RNAs, short open reading frames (ORFs) and pseudogenes, mislabelled as 'noncoding', were ultimately found to code for peptides[2–9]. Moreover, numerous peptides originate from non-canonical reading frames with non-AUG start codons[10].

We therefore hypothesised that proteogenomics might allow us to elucidate a fundamental question: the contribution of proteins derived from non-canonical transcripts to the repertoire of major histocompatibility complex (MHC) class I-associated peptides (MAPs). Endogenous MAPs are collectively referred to as the immunopeptidome and represent the essence of self for CD8 T lymphocytes[11,12]. Despite the fundamental importance of the immunopeptidome, its genesis remains ill-defined[13,14]. MAPs derive from proteolytic degradation of proteins found in all cell compartments; however, the immunopeptidome is not a random sample of the proteome: many abundant proteins do not generate MAPs, while some low-abundance proteins generate large amounts of MAPs[13–16]. In a series of seminal studies, Shastri and colleagues made startling observations showing that, similarly to the proteome, the immunopeptidome might be more complex than anticipated. Using an alloreactive T-cell clone as a probe, they screened a splenic cDNA library in transfected antigen-presenting cells (APCs) and isolated a cDNA clone that encoded the MAP recognized by the T-cell clone. The salient finding was that this MAP derived from a non-canonical reading frame initiated with a non-AUG start codon[17]. They discovered that synthesis of this peptide was initiated with a CUG codon decoded as a leucine rather than a methionine[18]. Studies by other groups provided evidence that MAPs could arise not only from alternate translational reading frames but also from untranslated regions (UTRs or introns)[19,20]. However, the structure of only a handful of these 'cryptic MAPs' has been confirmed with MS[20,21]. Therefore, in the absence of proteomic evidence, the existence of most reported cryptic MAPs must be considered with some scepticism because their identification relied on indirect methods fraught with high false discovery rates (FDRs). We therefore developed a novel proteogenomic approach to define the landscape of the cryptic immunopeptidome and answer the following questions: what proportion of MAPs derives from non-canonical reading frames and how are they generated? To this end, we performed an all-frames translation of the transcriptome of human B lymphoblastoid cell lines to generate databases of predicted peptides/proteins. These databases were used to identify MAPs using high-throughput MS sequencing. Integration of transcriptomic and proteomic data revealed that cryptic MAPs constitute ~10% of the immunopeptidome and that their biogenesis and properties differ in many ways from those of conventional MAPs.

Results

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

The control database corresponds to the canonical proteome of the B-LCL and was generated as follows (Fig. 1a, left): RNA-seq reads were mapped on the reference genome (version GRCh37.75) to identify subject 1-specific high-quality non-synonymous single-nucleotide polymorphisms (ns-SNPs), which were then integrated in the reference genome to build the personalized genome of subject 1. All putative protein-coding genes were then *in silico* translated in their conventional reading frame to obtain the canonical proteome of the B-LCL. The all-frames database was built using the six-frame translation of RNA-seq data from the B-LCL (Fig. 1a, right): reads passing the Illumina quality filters were *in silico* translated into six possible reading frames using a sliding window of 33 base pairs (bp), since the vast majority of MAPs are known to be 8–11 amino acids long and only rare MAPs contain more than 11 residues. Translation products having a length inferior to eight amino acids, due to the presence of a stop codon within the sliding window, were excluded. By not aligning the reads before translating them, we are able to leverage the whole output of the sequencer, including reads resulting from rare elongation events that might otherwise be discarded. However, this approach also prevents us from using established filtering approaches such as coverage measures or base quality filters. To address the necessity of sequence filtering, we computed for each translated peptide an *S*-value (or Seen-value) that represents the number of

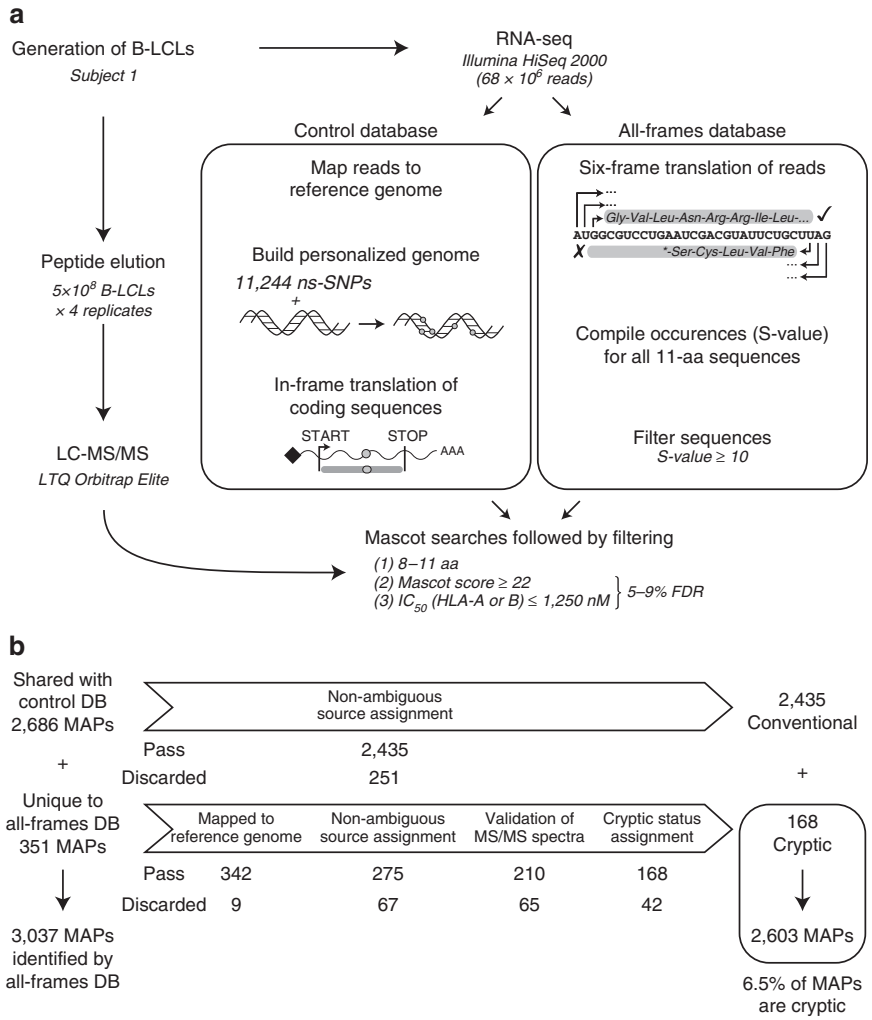


Figure 1 | Proteogenomic workflow used for high-throughput identification of cryptic MAPs. (**a**) General overview of the proteogenomic workflow used to identify conventional (Conv.) and cryptic (Crypt.) MAPs. Peptides were eluted from the cell surface of subject 1's B-LCL and were sequenced with liquid chromatography-MS/MS (LC-MS/MS). To determine the amino-acid (aa) sequence of those peptides, we built two databases (DBs), both derived from the analysis of RNA-seq data obtained from subject 1's B-LCL: the control DB and the all-frames DB (see Methods and Supplementary Fig. 1). (**b**) Peptides solely identified by the all-frames DB were considered as Crypt. MAP candidates and further filtered to remove ambiguous and false-positive identifications. See also Supplementary Figs 2 and 3.

times a peptide was seen following the *in silico* translation (Supplementary Fig. 1a). The higher the *S*-value, the more confidence we have that the peptide sequence is indeed not due to a sequencing error. We therefore elected to use a stringent approach and kept only peptides having an *S*-value ≥ 10 to (i) obtain a database whose size was manageable using the Mascot search engine (Supplementary Fig. 1b) and (ii) to minimize the risk of false discovery[4,5,24].

In our search for cryptic MAPs, the key question was whether the all-frames database would lead to the identification of MAPs missed with the control database, which only contains the *in silico* translation of sequences assumed to be translated (for example, protein-coding transcripts). Out of 3,037 MAPs identified by the all-frames database, 2,686 MAPs were also identified by the control database among which 2,435 were unambiguously assigned to a single gene (Fig. 1b). However, the salient finding is that 351 MAPs were solely identified by the all-frames database. After these 351 putative cryptic MAPs were subjected to four stringent filtering and validation steps (see Methods), we found that 168 of them were unambiguously assigned to a single genomic region (Fig. 1b). We validated 18 cryptic MAPs using the synthetic version of them (Supplementary Fig. 2). Furthermore, we found that the Mascot score distribution (the confidence level of a peptide assignation using MS) and the transcriptomic coverage of the peptide-coding regions (PCRs) were similar for these 168 cryptic and the 2,435 conventional MAPs (Supplementary Fig. 3). It should be noted that the multiple filtering steps were designed to be particularly stringent. We therefore expect that some of the 183 discarded peptides may, nevertheless, be genuine cryptic MAPs (Fig. 1b), thereby increasing their total number up to 351 (13% of the immunopeptidome). However, at this discovery stage, we chose to conduct further analyses using only the 168 cryptic MAPs identified using our most stringent criteria (6.5% of the immunopeptidome).

The cryptic MAPs' repertoire is linked to the HLA genotype. Various human leukocyte antigen (HLA) allotypes have different peptide-binding motifs and therefore present different MAP repertoires. Accordingly, if a peptide eluted from cells of subject 1 is a genuine MAP, its presence on cells from other individuals should depend on the presence of the HLA allotype, presenting this peptide on cells from subject 1. In other words, the presence

of authentic MAPs should be 'HLA-restricted'. The restriction should be strong but it does not need to be perfect because there are some overlap in the MAP repertoires presented by various allotypes[25]. On the contrary, no HLA restriction should be seen between the HLA genotype and the presence of MHC-unrelated peptides. Therefore, to test whether our cryptic MAPs were HLA-restricted, we analysed the immunopeptidome of three other subjects who shared four, two or no HLA allotypes with subject 1 (Supplementary Fig. 4). For both conventional and cryptic MAPs, we found a very strong positive dependence between peptide detection in subjects 2–4 and the presence of the corresponding HLA-A or -B allotype (two-sided Fisher's exact test, $P<2.2\times10^{-16}$; Fig. 2a,b). The degree of HLA allotype restriction was similar for conventional and cryptic MAPs. Moreover, most of the MAPs detected in the absence of the relevant HLA allele were predicted to be promiscuous binders (Fig. 2c). These data further validate that cryptic peptides detected with our proteogenomic approach are genuine MAPs.

Cryptic MAPs derive from both coding and noncoding RNAs. Next, we analysed the origin of cryptic MAPs. A notable finding was that 20.2% of cryptic MAPs unambiguously allocated to one gene could be assigned exclusively to non-annotated antisense transcripts (transcribed from non-template DNA strand; Fig. 3a). This suggests that, although antisense transcripts are generally assumed to be noncoding[26], their translation can generate substrates for the MHC class I antigen presentation pathway. Next, we focused our efforts on sense cryptic MAPs, as annotations were available for their respective gene source, and made two observations. First, by using the gene biotype nomenclature that classifies genes according to their biological relevance[27], we observed that 86.6% of sense cryptic MAPs derived from protein-coding genes, 9% from genes assumed to be noncoding such as pseudogenes, annotated antisenses, long intergenic noncoding RNAs or processed transcripts and finally 4.5% from unannotated intergenic regions (Fig. 3b). Second, by analysing the location of sense cryptic MAPs within their respective gene source, we observed that 48.5% of them were produced by out-of-frame translation of exonic sequences. The remaining 51.5% originated from translation of allegedly noncoding sequences (Fig. 3c). Among those, cryptic MAPs predominantly derived from the translation of 5′UTRs as opposed to 3′UTRs (24.6% versus 7.5%). This observation is coherent with the reinitiation model for translation initiation, which implies that the probability of translation initiation

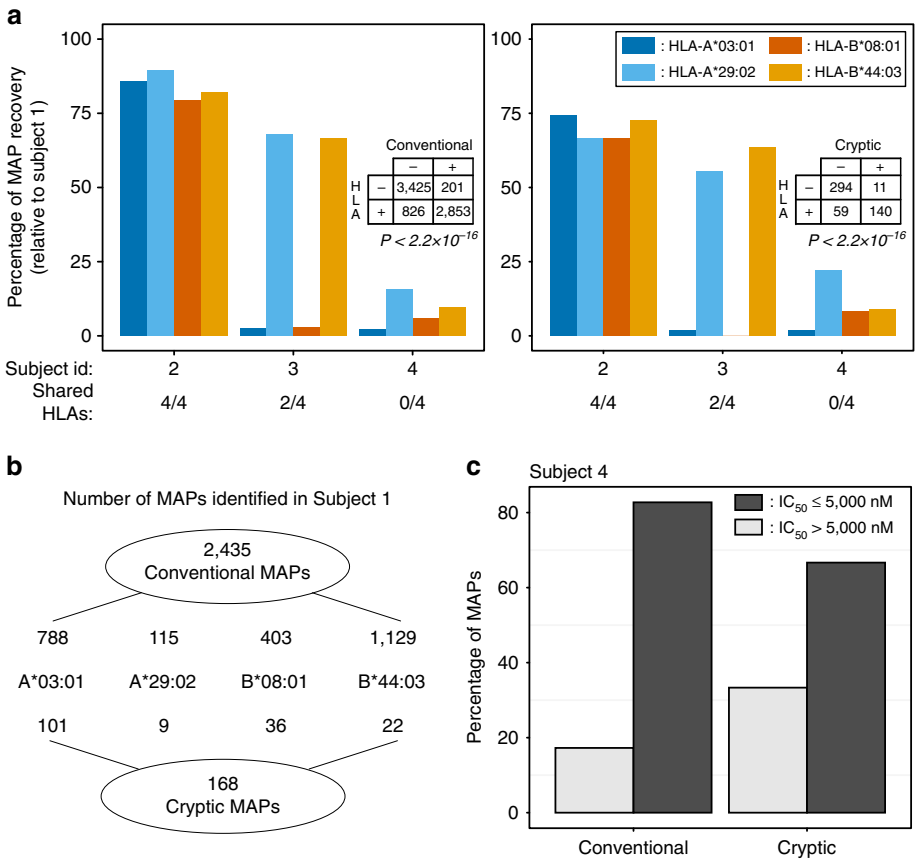


Figure 2 | Detection of Crypt. and Conv. MAPs is HLA-dependent. (**a**) Relationship between MAP detection and HLA genotype. We sequenced MAPs on B-LCLs from three subjects who shared four, two or no HLA alleles with subject 1. We then determined the number of Conv. (left) and Crypt. (right) MAPs found in subject 1 that were shared by subjects 2–4. Each bar represents one HLA allotype. A detailed schematic of the analysis can be found in Supplementary Fig. 4. MAP detection in subjects 2–4 correlated with presence of the HLA allotype presenting the MAPs in subject 1: $P<2.2\times10^{-16}$ for Conv. and Crypt. MAPs (two-sided Fisher's exact test). (**b**) Schematic detailing of the numbers of Conv. and Crypt. MAPs identified in subject 1 for the considered HLA alleles. (**c**) Most MAPs detected in subject 4 are promiscuous binders. Overall, 168 Conv. and 9 Crypt. MAPs detected in subject 1 were also detected in subject 4, even though the two subjects did not share any HLA alleles. Using NetMHCcons, we computed the predicted binding affinity (IC_{50}) of those MAPs for the four HLA-A and -B allotypes of subject 4, and we kept the lowest of the four IC_{50} values (corresponding to the highest MHC-binding affinity). The bar chart depicts the percentage of Conv. and Crypt. MAPs having an $IC_{50}\leq$ or $>5{,}000$ nM. Peptides with an $IC_{50}\leq5{,}000$ nM for the HLA-A/B allotypes of subject 4 were assumed to be promiscuous binders, that is, to bind subject 4 allotypes in addition to subject 1 allotypes.

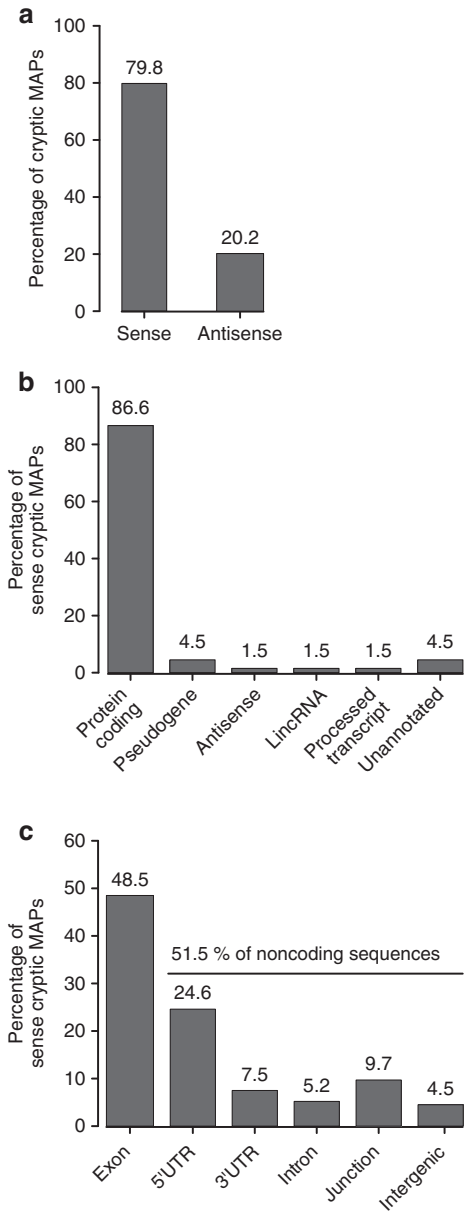


Figure 3 | Crypt. MAPs derive from both coding and noncoding transcripts. (**a**) Some Crypt. MAPs derive from novel antisense transcripts. Bar plot showing the percentages of Crypt. MAPs derived from sense and antisense transcriptions. (**b**,**c**) For Crypt. MAPs derived from sense transcription, we determined the percentage of each gene biotype in MAP source genes (**b**) and the proportion of Crypt. MAPs generated by six types of genomic regions (**c**). The 'exon' class refers to out-of-frame Crypt. MAPs, while the 'junction' category corresponds to peptides encoded by intron–exon or UTR–exon junction. LincRNA, long intergenic noncoding RNAs.

decreases along the transcript[28]. A small proportion of peptides (5.2%) derived from intronic sequences, a finding consistent with a report showing that a construct coding for the model SIINFEKL peptide, could generate MAPs after insertion into an intronic sequence[29]. Finally, we observed that 9.7% of cryptic MAPs derived from UTR–exon or intron–exon junctions and thus corresponded to translation products of overlapping short ORFs or retained intron transcripts, respectively. Overall, these results highlight the complexity of the immunopeptidome by showing that the landscape of cryptic MAPs includes both sense as well as antisense coding and noncoding RNAs.

Cryptic MAPs derive from ORFs with a 5′ end positional bias. We next sought to determine whether specific types of genes would preferentially generate cryptic as opposed to conventional MAPs. We first noted that very few genes generated both conventional and cryptic MAPs: (i) among the 121 cryptic MAP source genes, only 17 (that is, 14%) also gave rise to conventional MAPs and (ii) only 1% of the 1,731 conventional MAP source genes generated cryptic MAPs (Fig. 4a). The small overlap between genes coding cryptic versus conventional MAP suggests that these two gene sets possess some intrinsic differential feature(s). Further analyses highlighted two conspicuous differences between genes coding conventional versus cryptic MAPs. First, cryptic PCRs were located much closer to the 5′ end of their source transcript than conventional PCRs (Fig. 4b). This shift in PCR location was observed not only for cryptic MAPs coded by 5′UTRs and 5′UTR/exons but also for the entire set of exonic cryptic MAPs (Supplementary Fig. 5a). Second, the expression level of genes coding cryptic and conventional MAPs was different. Conventional MAPs have been shown to derive preferentially from abundant transcripts[30,31], and we observed that this was also the case for cryptic MAPs. However, the expression of cryptic MAP-coding genes was slightly but significantly inferior to that of conventional MAP source genes (Fig. 4c).

MAPs derive primarily from rapidly degraded proteins, and evidence suggests that the nonsense mediated decay (NMD) pathway plays a significant role in this process via translation-dependent degradation[32,33]. NMD targets messenger RNAs (mRNAs) containing a premature termination codon or normal mRNAs containing upstream ORFs[33,34]. Premature termination is predicted to result in more MAPs originating from the 5′ end of the transcript[35], as we observed for cryptic but not conventional MAPs (Fig. 4b). In addition, we found that the proportion of MAP-coding transcripts that harboured at least one upstream ORF was significantly higher for cryptic than for conventional MAPs (30% versus 13%), while transcripts generating both types of MAPs showed an intermediate percentage (20%, Fig. 4d). Since transcripts with an upstream ORF generated cryptic MAPs from 5′UTRs but also from exons and 3′UTRs (Supplementary Fig. 5b), NMD appears to be involved in the generation of all types of cryptic MAPs. Moreover, NMD was also reported to target transcripts bearing long 3′UTRs or 3′UTRs containing intronic sequences. While the transcript source of conventional and cryptic MAPs displayed the same frequency of 3′UTR introns (Supplementary Fig. 5c), cryptic MAP source transcripts had longer 3′UTRs than conventional MAP source transcripts (1,100 versus 687 nt, Fig. 4e). Taken together, these observations suggest that NMD contributes to the generation of cryptic MAPs while lowering the abundance of cryptic MAP source transcripts relative to conventional ones (Fig. 4c) because NMD reduces the steady-state levels of its target RNAs. Besides NMD, mRNA stability is also regulated by *cis*-regulatory elements that are located in 3′UTRs and interact with RNA-binding proteins[36]. In line with this, relative to conventional MAP source transcripts, the 3′UTRs of cryptic MAP source transcripts contained similar numbers of stabilizing elements but an increased number of destabilizing elements (Fig. 4f). In other words, cryptic MAP source transcripts display longer 3′UTRs with a selective enrichment in destabilizing elements. Taken together, our data suggest that cryptic MAPs derive from unstable transcripts targeted by NMD or 3′UTR-destabilizing elements.

Cryptic MAPs derive from precursors with atypical C termini. To gain further insights into the mechanisms responsible for the generation of cryptic MAPs, we analysed the nucleotide sequence of MAP source transcripts to predict their translation start and

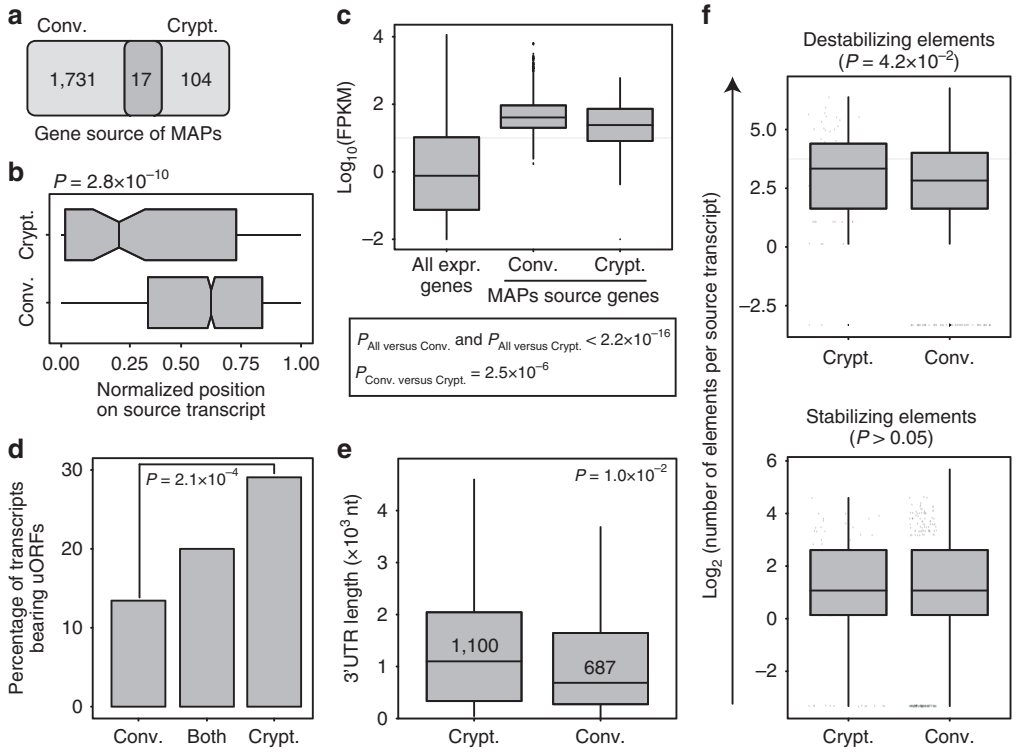


Figure 4 | Crypt. MAPs preferentially derive from unstable mRNAs. (**a**) Venn diagram showing minimal overlap between the gene source of Conv. and Crypt. MAPs. (**b**) Crypt. MAPs preferentially derive from the 5′ end of their source transcript. The length of each source transcript was normalized to 1, and the start of each MAP was then positioned on a 0–1 scale (*x* axis), where 0 represents the 5′ end of the source transcript. Crypt. MAPs deriving from intergenic and intronic regions were excluded from this analysis. See also Supplementary Fig. 5a. (**c**) Log_{10} expression values, in FPKM, of all genes expressed in B-LCL versus the subset of the gene source of Conv. and Crypt. MAPs. (**d**) Crypt. source transcripts preferentially bear upstream ORFs (uORFs). For each MAP source transcript, we predicted the 5′UTR and 5′UTR–exon ORF initiating at an AUG embedded in an optimal or strong Kozak context. The bar graph shows the proportion of source transcripts bearing at least one uORF and generating a Conv. MAP, a Crypt. MAP or both. See also Supplementary Fig. 5b. (**e**) Crypt. source transcripts display long 3′UTRs. Using pyGeno, we retrieved the 3′UTR of MAP source transcripts (when available) and computed their length in nucleotide (nt). The boxplot displays the resulting 3′UTR length distribution for Crypt. and Conv. MAP source transcripts excluding the upper outliers that represented 6 and 107 values out of 97 and 1,770 transcripts, respectively. (**f**) 3′UTRs of Crypt. but not Conv. MAP source transcripts are enriched in destabilizing elements. We looked for destabilizing and stabilizing elements identified in ref. 36 in the 3′UTR of Crypt. and Conv. MAP source transcripts. For each source transcript, we computed the number of destabilizing and stabilizing elements contained in its sequence. The resulting distributions are plotted for Crypt. and Conv. MAP source transcripts as the log_2 number of destabilizing (top panel) or stabilizing elements (bottom panel) per transcript. See also Supplementary Fig. 5c. Statistical significance was assessed with a two-sided (**b**,**c**,**e**) or one-sided (**f**) Wilcoxon rank sum test, or a two-sided Fisher's exact test (**d**). On box plots, boxes represent second and third quartiles, whiskers ±1.5 the interquartile range, and dots the outliers.

stop sites. Notably, we observed that translation initiation occurred at a known initiation codon for 69% of cryptic MAPs: AUG was used more often than near-cognate start codons, which differ from AUG by a single nucleotide (62% versus 7%). This suggests that, even for those atypical proteins, AUG is the preferential translation initiation codon (Fig. 5a). Among near-cognate start codons, CUG was the most commonly observed (Fig. 5b). This observation is in agreement with several reports demonstrating that CUG is the most efficient near-cognate start codon to initiate translation[18,37,38]. Other near-cognate start codons that were used more than one time included ACG and GUG, which were both shown to be enriched at translation initiation sites by ribosome profiling[38]. Finally, 31% of cryptic MAPs did not display any of the known translation initiation codons upstream of their respective PCR (Fig. 5a). In accordance with similar observations based on analyses of ribosome-profiling data[38], these data suggest that translation can be initiated at other codons than the classical AUG or near-cognate start codons.

The median length of conventional proteins is ~400 amino acids and, simply by virtue of their size, longer proteins generate more MAPs than shorter proteins[14]. Accordingly, the median length of conventional MAP source proteins in our data set was 523 amino acids. In stark contrast, the median length of cryptic MAP source proteins was 39 amino acids, and 75% of them had less than 62 amino acids (Fig. 5c). The shortest predicted cryptic proteins (3 out of 168) had a length of 10 amino acids and generated cryptic MAPs of 9 amino acids; MHC processing of these cryptic MAPs only required trimming of the N-terminal methionine. The generation of conventional MAPs is initiated by proteasomal cleavage followed in general by exopeptidase trimming of the N terminus but not the C terminus[39–41]. Therefore, with few exceptions, the C terminus created by the proteasome remains intact in conventional MAPs[42,43]. Given the remarkably short size of cryptic MAP source proteins, we hypothesized that many cryptic MAPs may not need proteasomal degradation before entering the MHC class I antigen presentation pathway. We reasoned that, if cryptic MAPs were proteasome-independent, their C terminus might be different from that of (proteasome-dependent) conventional MAPs. To test this hypothesis, we analysed amino-acid usage at the four C-terminal amino acids of individual MAPs and the four amino acids downstream of the C terminus (in the source protein) for

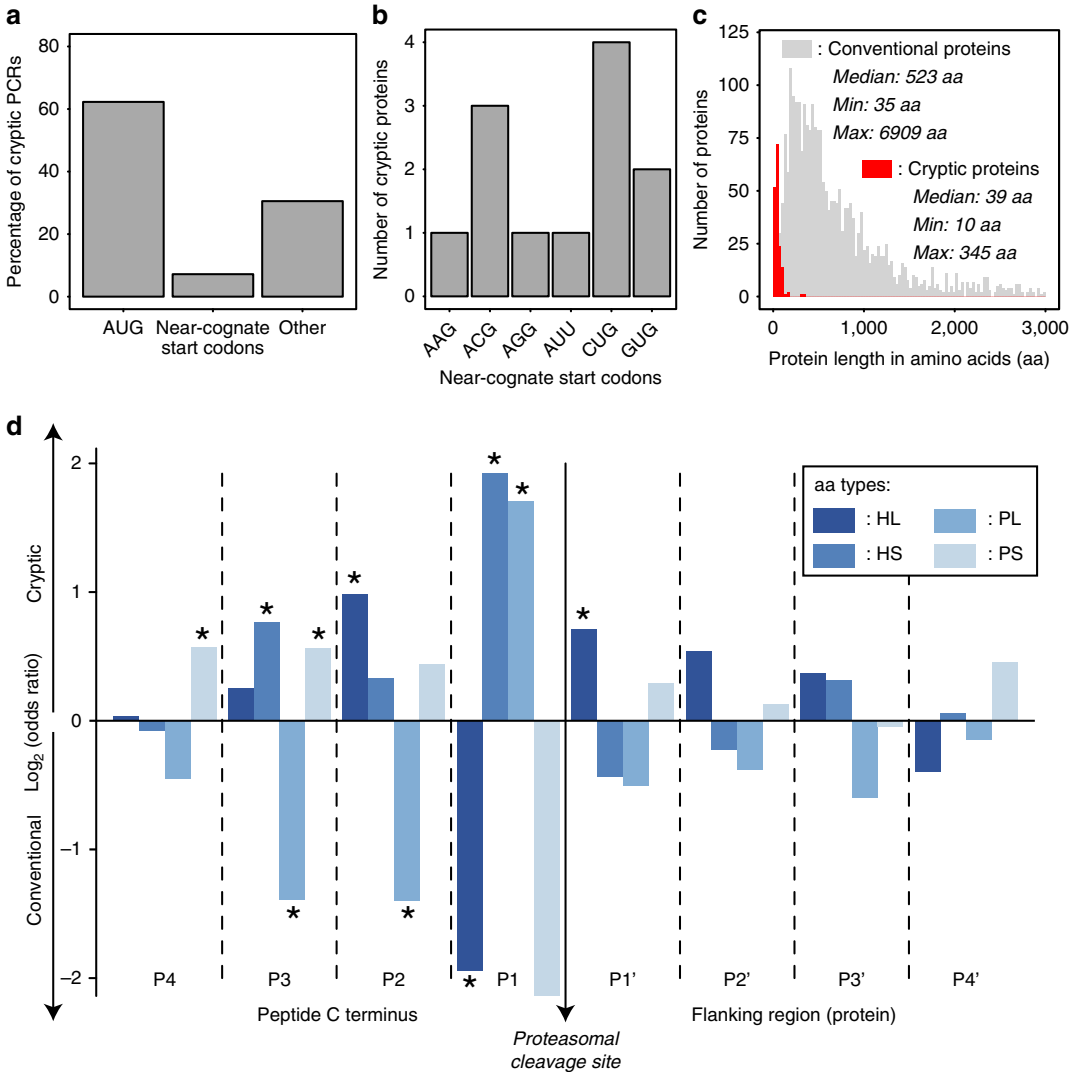


Figure 5 | Features of ORFs coding Crypt. MAPs. (**a**) Most Crypt. PCRs are in-frame with an upstream start codon. To predict the probable start codon of each Crypt. PCR, we sequentially applied the following rules: (i) presence of an upstream AUG within an optimal (GCC[R]CCstartG[V]), strong ([R]NNstartG[V]) or weak (anything else) Kozak context, (ii) presence of an upstream near-cognate start codon within an optimal or strong Kozak context, (iii) any other codon downstream of the first upstream stop codon. Bars represent the percentage of Crypt. PCRs displaying an upstream in-frame AUG, near-cognate start codon or any other codon as a probable initiation codon. (**b**) Bar plot showing near-cognate start codon usage at putative translational start sites of 12 Crypt. source proteins. (**c**) Length distribution of Conv. and predicted Crypt. proteins. Median, minimum (Min) and maximum (Max) observed lengths are indicated on the graph for both types of proteins. Conv. proteins having a length >3,000 amino acids are not displayed on the graph. (**d**) Crypt. and Conv. MAPs do not have the same amino-acid composition at their C termini. Amino acids (aa) were classified in four categories: Hydrophobic/Large (HL), Hydrophobic/Small-Medium (HS), Polar/Large (PL) and Polar/Small-Medium (PS)[44]. For the MAP C terminus (positions P4 to P1) and its C-terminal flanking region (positions P1′ to P4′), we compared the usage of those four aa categories at each position between Crypt. and Conv. MAPs. The graph displays the $\log_2$(odds ratio) and significant differences are marked with an asterisk (*$P<0.05$; two-sided Fisher's exact test).

conventional versus cryptic MAPs. The 20 amino-acid residues were grouped into four categories based on their bulkiness and hydrophobicity[44], and we analysed these data to determine which categories were enriched or depleted at each position for the two types of MAPs. We found that, out of the eight considered positions, five displayed significant differential amino-acid class usage between cryptic and conventional MAPs (Fig. 5d). Together, the facts that cryptic MAPs originate from very short proteins and that amino-acid usage around their C termini is different from that of conventional MAPs suggest that processing of cryptic MAPs may be proteasome-independent.

Cryptic MAPs display distinct features and are immunogenic. We next evaluated relevant structural and functional features of cryptic MAPs *per se*. Relative to conventional MAPs, we found that cryptic MAPs exhibited three distinctive characteristics: they were shorter, had different allotype-binding preferences and harboured more genomic polymorphisms (Fig. 6). The length distribution of cryptic MAPs revealed a significant enrichment in 8-mers and depletion in 10–11-mers when compared with conventional MAPs (Fig. 6a). This further supports the idea that cryptic and conventional MAPs are processed differently by peptidases. Unexpectedly, we found that cryptic MAPs were preferentially presented by HLA-A*03:01, while conventional MAPs were preferentially presented by HLA-B*44:03 in subject 1 (Fig. 6b). Proteogenomic studies of MAPs presented by other HLA allotypes will be required to assess whether differential allotype preferences of cryptic and conventional MAPs can be generalized. If it were the case, one implication would be that the HLA genotype dictates the breadth of the cryptic

immunopeptidome presented at the cell surface. No bias in favour or against ns-SNPs was found in conventional MAP PCRs[22]. However, we found that cryptic MAP PCRs contained a significantly higher frequency of ns-SNPs than conventional MAP PCRs (Fig. 6c; $P = 5.625 \times 10^{-3}$). In other words, cryptic MAPs derive from genomic sequences that are more polymorphic at the population level than conventional protein-coding sequences.

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

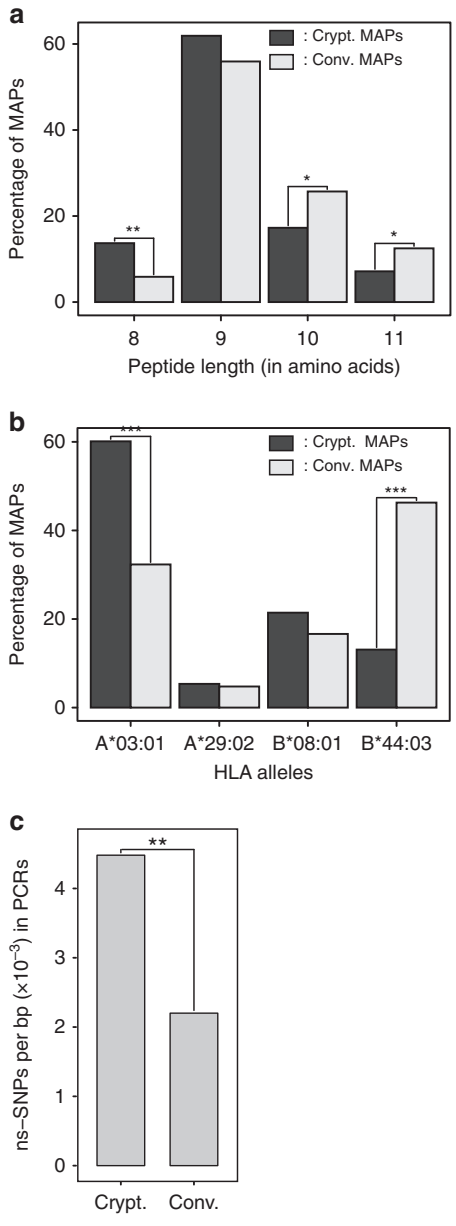


Figure 6 | Crypt. and Conv. MAPs display different features. (**a**–**c**) Bar plots showing that Crypt. and Conv. MAPs from subject 1 have different (**a**) length distribution, (**b**) allotype distribution and that (**c**) their PCRs exhibit different ns-SNP frequencies (from dbSNP138). In all cases, statistical significance was assessed using a two-sided Fisher's exact test: $^*P \leq 0.05$, $^{**}P \leq 0.006$, $^{***}P \leq 1.10^{-11}$ in the bar plots.

Discussion

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

About 50% of cryptic MAPs result from out-of-frame translation and the other half from translation of allegedly noncoding sequences. The ultimate biological role of cryptic translation remains elusive. However, it might be unwise to assume that this phenomenon merely represents 'translational noise'. Protein synthesis is demanding: it is the most

Table 1 | Features of polymorphic cryptic MAPs presented in Fig. 7.

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

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

Table 2 | Features of non-polymorphic cryptic MAPs presented in Fig. 7.

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

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

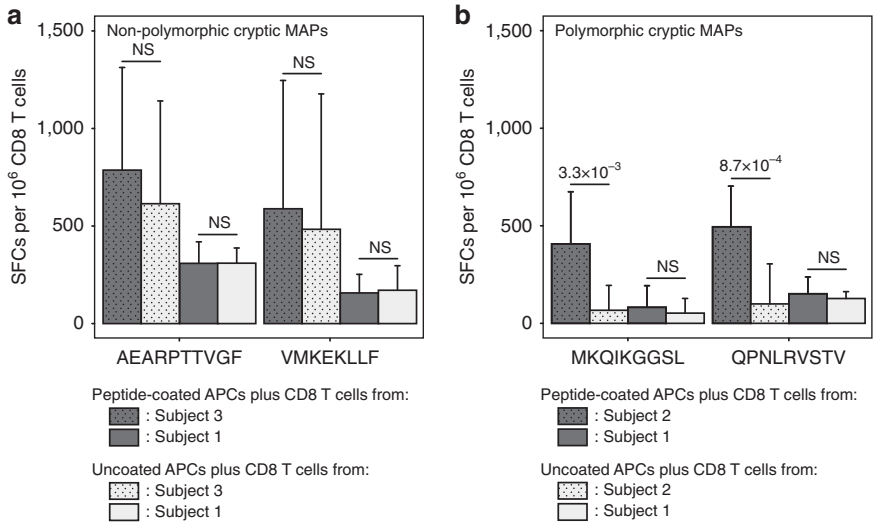


Figure 7 | Immunogenicity of Crypt. MAPs. (**a**,**b**) Only polymorphic Crypt. MAPs are immunogenic. IFN-γ Elispot counts showing the number of spot-forming cells (SFCs) per million CD8 T cells for two non-polymorphic (**a**) and two polymorphic (**b**) Crypt. MAPs. Final counts were obtained following the subtraction of background spots (peptide-coated APCs alone) from the spots obtained when CD8 T cells were exposed to peptide-coated or uncoated APCs. The experiment was performed in biological triplicates (each with three technical replicates), error bars represent s.d. and statistical significance was assessed using a two-tailed Student's *t*-test (NS: not significant, $P > 0.05$). Features of the four tested Crypt. MAPs are detailed in Tables 1 and 2.

energy-consuming process in the cell as it monopolizes 45% of cellular ATP supplies[47]. Furthermore, any RNA sequence subject to translation will experience selection against encoding a protein with detrimental impact on cell function[6]. In any case, noncoding RNAs are vital, and our demonstration that several noncoding RNAs generate MAPs means that CD8 T cells have an opportunity to scrutinize these transcripts.

The gene source of conventional MAPs are enriched in microRNA-binding elements, suggesting that mRNA destabilization favours MAP generation[30]. By comparing transcripts coding conventional and cryptic MAPs, we obtained meaningful evidence suggesting that cryptic MAPs derive from particularly unstable transcripts targeted by NMD or 3′UTR destabilizing elements: (i) cryptic MAP transcripts were enriched in upstream ORFs and their PCRs showed a strong 5′ end positional bias (suggestive of premature termination) and (ii) cryptic MAP transcripts displayed longer 3′UTR enriched in destabilizing but not stabilizing elements when compared with conventional source transcripts. Together with previous work by us and others, these data allow for the development of an emerging model in which mRNA instability is instrumental in the genesis of all types of MAPs. This model is an extension of the idea that most MAPs derive from defective ribosomal products[32,48,49]: unstable RNAs targeted by NMD, microRNAs or other 3′UTR-destabilizing elements would generate more defective ribosomal products and therefore more MAPs. The validity of this model can be submitted to high-throughput experimental validation: if it is correct, mRNA half-life should be negatively correlated to MAP generation. We do not exclude that translation efficiency, which partly depends on codon usage[50], might also regulate MAP generation. Indeed, although we did not find evidence for a codon bias in conventional source transcripts versus cryptic MAP source ORFs ($P = 0.34$, odds ratio = 1.02), we observed that MAP source transcripts or ORFs in general use rare codons slightly more frequently than transcripts that do not generate MAPs ($P < 2.2 \times 10^{-16}$, odds ratio = 1.14; Supplementary Tables 2 and 3). Therefore, it might be interesting to further investigate the impact of codon bias on MAP generation.

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

One area where cryptic MAPs may be most relevant is cancer immunology. Although the vast majority of cancer mutations involve non-exomic regions, searches for tumour-specific antigens (TSAs) have focused on exomic mutations[31,52–54]. Nonetheless, since numerous noncoding transcripts are expressed only in cancer cells[55,56], a number of cryptic MAPs may be genuine TSAs. Furthermore, we demonstrated that (i) cryptic MAP PCRs displayed a higher frequency of germline polymorphisms (ns-SNPs) than the conventional exome (Fig. 6c) and that (ii) polymorphic cryptic MAPs discovered by proteogenomics were immunogenic (Fig. 7b). Hence, it is reasonable to expect that cryptic MAPs bearing somatic mutations (that is, TSAs) should also be immunogenic. Accordingly, in melanoma and renal cell carcinoma, pioneering

studies using more traditional approaches have uncovered unique immunogenic cryptic TSAs derived from noncoding regions[19,21]. Assuming that cryptic MAPs may be a rich source of heretofore overlooked TSAs, it is imperative to directly explore the presence of cryptic TSAs using systems-level approaches. Expanding the repertoire of TSAs would be highly beneficial because the low number of immunogenic exome-derived TSAs is a major hurdle for cancer immunotherapy[57–59].

Methods

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

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

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

Generation of the control and all-frames databases. We generated two customized databases based on the RNA-seq data of subject 1. To generate the control database, we applied a workflow similar to the one of Granados *et al.*[22]: ns-SNPs identified in subject 1 were integrated at their correct position in the reference genome (GRCh37.75) to build a personalized genome. Using the Ensembl gtf file, we extracted all known transcripts and further *in silico* translated them in their canonical reading frame to obtain the canonical proteome of subject 1. To generate the all-frames database, we used all reads passing the Illumina quality filters and *in silico* translated them in the six possible reading frames using a sliding window of 33 bp to obtain all theoretical peptides having a length between 8 and 11 amino acids. For each peptide, we computed an *S*-value, that is, the number of times it was seen following the *in silico* translation process. Only peptides having an *S*-value $\geq$10 as well as a length between 8 and 11 amino acids were included in the predicted peptidome of subject 1. Both the canonical proteome and the predicted peptidome of subject 1 were compiled in fasta files to obtain the control and the all-frames database, respectively. Both databases were then concatenated with their respective decoy counterpart and submitted to the Mascot database search engine along with subject 1's immunopeptidomic data.

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

Control and all-frames database searches. The Mascot search engine (Matrix Science) was used in combination with the control or the all-frames database concatenated to their reverse database to identify peptides present in the immunopeptidome of subject 1. Mass tolerances on precursor and fragment ions were set to 5 p.p.m and 0.02 Da, respectively. Searches were performed without enzyme specificity, and cysteinylation, phosphorylation (on Ser, Thr and Tyr), oxidation (Met) and deamidation (Asn, Gln) were used as variable modifications. Following each database search, we converted raw files to peptide maps containing *m*/*z* values, charge state, retention time and intensity above detection threshold ($\geq$8,000) using ProteoProfile (http://proteomics.iric.ca/tools/ProteoProfile/)[62]. The peptide maps were used to extract the abundance of the identified peptides across the four replicates.

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

Identification of cryptic and conventional MAPs. Peptides identified with both the control and the all-frames databases were considered as conventional MAPs. Peptides solely identified by the all-frames database were considered as putative cryptic MAPs. To validate whether they were genuine cryptic MAPs, we mapped the subset of peptide-encoding reads using TopHat to discard peptides coming from multiple locations in the genome. The remaining cryptic MAP candidates were assigned to their respective source gene and their MS/MS spectra were manually validated. To determine the type of sequence (within the source gene) generating each cryptic MAP, we used the intersect function of the BEDTools suite on the bed file of our cryptic candidates as well as Ensembl gtf file. Peptides assigned to a gene source in the opposite orientation were classified as antisense cryptic MAPs, those deriving from noncoding RNAs, 5′UTR, intronic, 3′UTR or intergenic sequences were classified as sense noncoding cryptic MAPs. Peptides deriving from exons of protein-coding genes were subjected to a reading frame validation: only peptides produced by non-canonical reading frames were classified as sense coding cryptic MAPs. For sense cryptic MAPs (except intergenic ones), we retrieved the gene biotype of their respective gene source from Ensembl annotations (when available) using pyGeno. Finally, since MAPs derive preferentially from highly abundant transcript[30,31,66], we assumed that the conventional and sense cryptic MAPs passing all of our filtering steps were generated by the most highly expressed isoform of their respective source gene. A complete list of identified conventional and cryptic MAPs can be found in Supplementary Data 1 and Supplementary Data 2, respectively.

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

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

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

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

Prediction of cryptic source proteins. To predict the probable start codon of each cryptic PCR, we sequentially applied the following rules: (i) presence of an upstream AUG within an optimal (GCC[R]CCstartG[V]), strong

([R]NNstartG[V]) or weak (anything else) Kozak context, (ii) presence of an upstream near-cognate AUG within an optimal or strong Kozak context, (iii) any other codon downstream of the first upstream stop codon. The probable stop codon was assumed to be the first in-frame stop codon downstream of the PCR. This analysis was performed on personalized mRNA sequences of cryptic source transcripts for most sense cryptic MAPs. Since no gene structures were known for antisense, intronic and intergenic cryptic MAPs, we simply extracted the personalized genomic sequences flanking the PCR (750 bp long) and performed the same analysis.

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

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

Rare codon usage analysis. Codons were classified as rare and common if their observed usage frequency (http://www.genscript.com/cgi-bin/tools/codon_freq_table)[67] was lower and greater than their expected usage frequency (1/number of codons encoding a given amino acid), respectively. Out of 64 codons, 30 were classified as rare and 34 as common. Using an in-house python script, we computed the number of occurrence for each codon to further derive the number of rare and common codons used by each class of transcripts across (1) conventional source transcripts versus cryptic source ORFs and (2) MAP source transcripts or ORFs versus all the other transcripts for which a cDNA sequence was defined. Statistical significance was assessed using a two-sided Fisher's exact test.

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

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

References

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
2. Kim, M. S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
3. Weekes, M. P. *et al.* Quantitative temporal viromics: an approach to investigate host-pathogen interaction. *Cell* **157,** 1460–1472 (2014).
4. Alfaro, J. A., Sinha, A., Kislinger, T. & Boutros, P. C. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nat. Methods* **11,** 1107–1113 (2014).
5. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11,** 1114–1125 (2014).
6. Ingolia, N. T. *et al.* Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **8,** 1365–1379 (2014).
7. Branca, R. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nat. Methods* **11,** 59–62 (2014).
8. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–587 (2014).
9. Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513,** 382–387 (2014).
10. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).
11. 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).
12. Hassan, C. *et al.* The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell Proteomics* **12,** 1829–1843 (2013).
13. 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).
14. 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).
15. Lev, A. *et al.* Compartmentalized MHC class I antigen processing enhances immunosurveillance by circumventing the law of mass action. *Proc. Natl Acad. Sci. USA* **107,** 6964–6969 (2010).
16. de Verteuil, D., Granados, D. P., Thibault, P. & Perreault, C. Origin and plasticity of MHC I-associated self peptides. *Autoimmun. Rev.* **11,** 627–635 (2012).
17. Malarkannan, S., Afkarian, M. & Shastri, N. A rare cryptic translation product is presented by Kb major histocompatibility complex class I molecule to alloreactive T cells. *J Exp. Med.* **182,** 1739–1750 (1995).
18. 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).
19. Starck, S. R. & Shastri, N. Non-conventional sources of peptides presented by MHC class I. *Cell Mol. Life Sci.* **68,** 1471–1479 (2011).
20. Goodenough, E. *et al.* Cryptic MHC class I-binding peptides are revealed by aminoglycoside-induced stop codon read-through into the 3'UTR. *Proc. Natl Acad. Sci. USA* **111,** 5670–5675 (2014).
21. Weinzierl, A. O. *et al.* A cryptic vascular endothelial growth factor T-cell epitope: identification and characterization by mass spectrometry and T-cell assays. *Cancer Res.* **68,** 2447–2454 (2008).
22. Granados, D. P. *et al.* Impact of genomic polymorphism on the repertoire of human MHC class I-associated peptides. *Nat. Commun.* **5,** 3600 (2014).
23. Daouda, T. pyGeno: a Python package for precision medicine https://github.com/tariqdaouda/pyGeno (2015).
24. 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).
25. Sidney, J., Southwood, S., Pasquetto, V. & Sette, A. Simultaneous prediction of binding capacity for multiple molecules of the HLA B44 supertype. *J. Immunol.* **171,** 5964–5974 (2003).
26. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14,** 880–893 (2013).
27. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
28. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15,** 193–204 (2014).
29. Apcher, S. *et al.* Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. *Proc. Natl Acad. Sci. USA* **110,** 17951–17956 (2013).
30. Granados, D. P. *et al.* MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements. *Blood* **119,** e181–e191 (2012).
31. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515,** 572–576 (2014).
32. Yewdell, J. W. DRiPs solidify: progress in understanding endogenous MHC class I antigen processing. *Trends Immunol.* **32,** 548–558 (2011).
33. Apcher, S. *et al.* Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *Proc. Natl Acad. Sci. USA* **108,** 11572–11577 (2011).
34. Smith, J. E. *et al.* Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep.* **7,** 1858–1866 (2014).
35. Kim, Y., Yewdell, J. W., Sette, A. & Peters, B. Positional bias of MHC class I restricted T-cell epitopes in viral antigens is likely due to a bias in conservation. *PLoS Comput. Biol.* **9,** e1002884 (2013).
36. Zhao, W. *et al.* Massively parallel functional annotation of 3' untranslated regions. *Nat. Biotechnol.* **32,** 387–391 (2014).
37. 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 USA* **107,** 18056–18060 (2010).
38. 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).
39. 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).

40. Blum, J. S., Wearsch, P. A. & Cresswell, P. Pathways of antigen processing. *Annu. Rev. Immunol.* **31,** 443–473 (2013).
41. Weimershaus, M., Evnouchidou, I., Saveanu, L. & van Endert, P. Peptidases trimming MHC class I ligands. *Curr. Opin. Immunol.* **25,** 90–96 (2013).
42. de Verteuil, D. *et al.* Deletion of immunoproteasome subunits imprints on the transcriptome and has a broad impact on peptides presented by major histocompatibility complex I molecules. *Mol. Cell Proteomics* **9,** 2034–2047 (2010).
43. Neefjes, J., Jongsma, M. L. M., 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).
44. Mishto, M. *et al.* Proteasome isoforms exhibit only quantitative differences in cleavage and epitope generation. *Eur. J. Immunol.* **44,** 3508–3521 (2014).
45. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14,** 496–506 (2013).
46. David, A. *et al.* Nuclear translation visualized by ribosome-bound nascent chain puromycylation. *J. Cell Biol.* **197,** 45–57 (2012).
47. Princiotta, M. F. *et al.* Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity* **18,** 343–354 (2003).
48. Apcher, S., Manoury, B. & Fahraeus, R. The role of mRNA translation in direct MHC class I antigen presentation. *Curr. Opin. Immunol.* **24,** 71–76 (2012).
49. 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. USA* **111,** E1591–E1599 (2014).
50. Quax, T. E., Claassens, N. J., Soll, D. & van der Oost, J. Codon bias as a means to fine-tune gene expression. *Mol. Cell* **59,** 149–161 (2015).
51. Boon, T. & Van Pel, A. T cell-recognized antigenic peptides derived from the cellular genome are not protein degradation products but can be generated directly by transcription and translation of short subgenic regions. A hypothesis. *Immunogenetics* **29,** 75–79 (1989).
52. van Rooij, N. *et al.* Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* **31,** e439–e442 (2013).
53. Robbins, P. F. *et al.* Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19,** 747–752 (2013).
54. Gubin, M. M. *et al.* Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515,** 577–581 (2014).
55. White, N. M. *et al.* Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. *Genome Biol.* **15,** 429 (2014).
56. Trimarchi, T. *et al.* Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell* **158,** 593–606 (2014).
57. 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).
58. Hinrichs, C. S. & Restifo, N. P. Reassessing target antigens for adoptive T-cell therapy. *Nat. Biotechnol.* **31,** 999–1008 (2013).
59. Heemskerk, B., Kvistborg, P. & Schumacher, T. N. The cancer antigenome. *EMBO J.* **32,** 194–203 (2013).
60. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
61. 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).
62. Thibault, P. ProteoProfile https://proteomics. iric. ca/tools/ProteoProfile/ (2015).
63. Sennels, L., Bukowski-Wills, J. C. & Rappsilber, J. Improved results in proteomics by use of local and peptide-class specific false discovery rates. *BMC Bioinformatics* **10,** 179 (2009).
64. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64,** 177–186 (2012).
65. 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).
66. Fortier, M. H. *et al.* The MHC class I peptide repertoire is molded by the transcriptome. *J. Exp. Med.* **205,** 595–610 (2008).
67. GenScript bioinformatic tools. GenScript codon usage frequence table tool *http://www.genscript. com/cgi-bin/tools/codon_freq_table* (2015).
68. Bollard, C. M. *et al.* Complete responses of relapsed lymphoma following genetic modification of tumor-antigen presenting cells and T-lymphocyte transfer. *Blood* **110,** 2838–2845 (2007).
69. Wolfl, M. & Greenberg, P. D. Antigen-specific activation and cytokine-facilitated expansion of naive, human CD8+ T cells. *Nat. Protoc.* **9,** 950–966 (2014).
70. Vincent, K. *et al.* Rejection of leukemic cells requires antigen-specific T cells with high functional avidity. *Biol. Blood Marrow Transplant.* **20,** 37–45 (2014).

Acknowledgements

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

Author contributions

C.M.L., T.D. and J.-P.L. designed the study. C.M.L., T.D., J.-P.L., E.B., O.C.-L., M.-P.H., D.P.G. and C.D. performed experiments; C.M.L. and C.P. wrote the first draft of the manuscript; all authors analysed the data and edited the final manuscript.

Additional information

Accession codes: RNA-seq data for the four subjects are available in the Gene Expression Omnibus database under accession code GSE67174 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=upkvweysnxabzkr&acc=GSE67174). MS data are available in PeptideAtlas for subjects 1 and 2 (http://www.peptideatlas.org/PASS/PASS00270); data from subjects 3 and 4 have been submitted to the ProteomeXchange Consortium[68] via the PRIDE partner repository with the data set identifiers PXD001898 (project accession) and 10.6019/PXD001898 (project DOI). In addition, the entire list of MAPs identified in subject 1 has been deposited into the Immune Epitope Database (http://www.iedb.org) under accession code 1028836.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: 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 doi: 10.1038/ncomms10238 (2016).