Université de Montréal

Studies of MHC class I antigen presentation & the origins of the immunopeptidome

par Hillary Pearson

Programme de Biologie Moléculaire
Faculté de Médecine

Mémoire présenté
en vue de l'obtention du grade de Maître ès Sciences
en Biologie Moléculaire
option Générale

Avril 2016

© Hillary Pearson, 2016

Résumé

La présentation d'antigène par les molécules d'histocompatibilité majeure de classe I (CMHI) permet au système immunitaire adaptatif de détecter et éliminer les agents pathogènes intracellulaires et des cellules anormales. La surveillance immunitaire est effectuée par les lymphocytes T CD8 qui interagissent avec le répertoire de peptides associés au CMHI présentés à la surface de toutes cellules nucléées.

Les principaux gènes humains de CMHI, HLA-A et HLA-B, sont très polymorphes et par conséquent montrent des différences dans la présentation des antigènes. Nous avons étudié les différences qualitatives et quantitatives dans l'expression et la liaison peptidique de plusieurs allotypes HLA. Utilisant la technique de cytométrie de flux quantitative nous avons établi une hiérarchie d'expression pour les quatre HLA-A, B allotypes enquête. Nos résultats sont compatibles avec une corrélation inverse entre l'expression allotypique et la diversité des peptides bien que d'autres études soient nécessaires pour consolider cette hypothèse.

Les origines mondiales du répertoire de peptides associés au CMHI restent une question centrale à la fois fondamentalement et dans la recherche de cibles immunothérapeutiques. Utilisant des techniques protéogénomiques, nous avons identifié et analysé 25,172 peptides CMHI isolées à partir des lymphocytes B de 18 personnes qui exprime collectivement 27 allotypes HLA-A,B. Alors que 58% des gènes ont été la source de 1-64 peptides CMHI par gène, 42% des gènes ne sont pas représentés dans l'immunopeptidome. Dans l'ensemble, l'immunopeptidome présenté par 27 allotypes HLA-A,B ne couvrent que 17% des séquences exomiques exprimées dans les cellules des sujets. Nous avons identifié plusieurs caractéristiques des transcrits et des protéines qui améliorent la production des peptides CMHI. Avec ces données, nous avons construit un modèle de régression logistique qui prédit avec une grande précision si un gène de notre ensemble de données ou à partir d'ensembles de données indépendants génèrerait des peptides CMHI. Nos résultats montrent la sélection préférentielle des peptides CMHI à partir d'un répertoire limité de produits de gènes avec des caractéristiques distinctes. L'idée que le système immunitaire

peut surveiller des peptides CMHI couvrant seulement une fraction du génome codant des protéines a des implications profondes dans l'auto-immunité et l'immunologie du cancer.

Mots-clés: complexe majeur d'histocompatibilité (CMH) de classe I, antigènes d'histocompatibilité humains (HLA), immunopeptidome, expression quantitative, spectrométrie de masse, régression logistique, modélisation

Abstract

Antigen presentation by major histocompatibility complex class I (MHCI) molecules allows the adaptive immune system to detect and eliminate intracellular pathogens or abnormal cells. Immune surveillance is executed by CD8+ T cells that monitor the repertoire of MHCI-associated peptides (MAPs) presented at the surface of all nucleated cells.

The primary human MHCI genes, HLA-A and HLA-B, are highly polymorphic and consequentially demonstrate differences in antigen presentation. We investigated qualitative and quantitative differences in expression and peptide binding. Using quantitative flow cytometry we establish a clear hierarchy of expression for the four HLA-A,B allotypes investigated. Our results are consistent with an inverse correlation between expression and peptide diversity although further work is necessary to solidify this hypothesis.

The global origins of the MAP repertoire remains a central question both fundamentally and in the search for immunotherapeutic targets. Using proteogenomics, we identified and analyzed 25,172 MAPs isolated from B lymphocytes of 18 individuals who collectively expressed 27 HLA-A,B allotypes. While 58% of genes were the source of 1-64 MAPs per gene, 42% of genes were not represented in the immunopeptidome. Overall, we estimate the immunopeptidome presented by 27 HLA-A,B allotypes covered only 17% of exomic sequences expressed in subjects' cells. We identified several features of transcripts and proteins that enhance MAP production. From these data we built a logistic regression model that predicts with high accuracy whether a gene from our dataset or from independent datasets would generate MAPs. Our results show preferential selection of MAPs from a limited repertoire of gene products with distinct features. The notion that the immune system can monitor MAPs covering only a fraction of the protein coding genome has profound implications in autoimmunity and cancer immunology.

Keywords: major histocompatibility complex (MHC) class I , human leukocyte antigen (HLA), immunopeptidome, quantitative expression, mass spectrometry, logistic regression, modeling

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Acronyms

3'UTR	3' untranslated region
5'UTR	5' untranslated region
7-AAD	7-aminoactinomycin D
aa	Amino acid residue
ABC	Antibody binding capacity
ACT	Adoptive cell therapy
AHCT	Allogeneic hematopoietic cell transplantation
AIRE	Autoimmune regulator
ANOVA	Analysis of variance
APC	Antigen presenting cell
AU	Adenosine & uridine ribonucleic acids
AUC	Area under the curve
β_2m	β_2-micoglobulin
BCR	B cell receptor
B-LCL	B lymphoblastoid cell line
bp	Base pairs
CMH	Complexe majeur d'histocompatibilité
CMHI	Complexe majeur d'histocompatibilité de classe I
CRT	Calreticulin
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
DRiP	Defective ribosomal products
EBV	Epstein-Barr virus
EDTA	Ethyenediamineteraacetic acid
ER	Endoplasmic reticulum
ERAP 1/2	Endoplasmic reticulum aminopeptidase associated with antigen processing
ERp57	Endoplasmic reticulum resident protein 57
FACS	Fluorescence-activated cell sorting
FBS	Fetal bovine serum
FC	Flow cytometry
FDR	False discovery rate
FPKM	Fragments per kilobase of transcript per million mapped reads

FSC	Forward scatter
GC	Guanosine & cytidine ribonucleic acids
GO	Gene ontology
GPCR	G-protein coupled receptor
GSEA	Gene set enrichment analysis
GWAS	Genome wide association studies
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HPLC	High performance liquid chromatography
IQR	Interquartile range
MAE	Mild acid elution
MAP	MHC class I associated peptide
MFI	Mean fluorescence intensity
MHC	Major histocompatibility complex
MHCI	Major histocompatibility complex class I
MHCII	Major histocompatibility complex class II
MiHA	Minor histocompatibility antigen
mRNA	Messenger RNA
MS	Mass spectrometry
neo-MAP	neoantigen MAPs
NMD	Nonsense mediated decay
nsSNP	Non-synonymous single nucleotide polymorphism
uORF	Upstream open reading frame
PBMC	Peripheral blood mononuclear cell
PBS/BSA	Phosphate buffered saline, 1% bovine serum albumin
PLC	Peptide loading complex
pMHCI	peptide-MHCI complex
QIKIFIT	Quantitative indirect immunofluorescence kit, © Dako
RDP	Rapidly degraded protein
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
ROC	Receiver operating characteristic
SABC	Specific antibody binding capacity

SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SSC	Side scatter
TAP	Transporter associated with antigen processing
TCGA	The Cancer Genome Atlas
TCR	T cell receptor
TIL	Tumour infiltrating lymphocyte
TS	TargetScan 7.0

Acknowledgements

This work was made possible by many talented and generous people. The progression of my studies was shaped by discussions over coffee, in lab meetings, and in the sofas overlooking the Saint-Joseph Oratory. If I have discovered one thing in my journey from learning how to use a pipette to coding genome wide data analyses, it is that research is truly a collaborative effort.

First and foremost, I would like to thank my research director Dr. Claude Perreault for the opportunity to join team immunopeptidome. I feel fortunate to have found a mentor who encouraged me to take on new challenges while reminding me to focus judiciously and whose passion and wisdom in unparalleled.

Thank you to Jean-Baptiste for joining me in this adventure, to my parents for their unwavering support, and to my family for inspiring me every step of the way. Thank you to everyone in the Perreault lab for their enthusiasm and helpfulness regarding my research. Our lunchtimes, chocolate breaks and shared love of coffee were indispensible research fuel. I am very lucky to have worked with such a brilliant and welcoming group of people. A special thank you to Diana Granados for being a patient and thoughtful supervisor. Thank you to Tariq Daouda and Sébastien Lemieux for their guidance in the realm of statistics and bioinformatics. Thank you to the coauthors of the article presented in chapter 3 for their tremendous efforts assembling the immunopeptidome. Thank you to Dr. Brian Wilhelm and Dr. Luis Barreiro for their time and effort in revising my master's thesis.

I would also like to acknowledge the efforts of everyone at the IRIC to provide opportunities for budding scientists and build a fruitful research environment. Finally, this work would not have been possible without a community that values science and the generous support of the Quebec Breast Cancer Foundation.

Overview

The adaptive immune system is charged with surveillance and elimination of threats to host survival. This is achieved through two pillars of adaptive immunity: self / non-self discrimination and diversity. Self versus non-self discrimination describes the ability of effector cells to recognize foreign bodies by becoming tolerant of the self. Diversity within the tools of the adaptive immune repertoire is essential for detection and recognition of diverse threats to survival.

Allelic polymorphism of the major histocompatibility complex class I (MHCI) genes is one mechanism of diversity. In antigen presentation, MHCI bind intracellular protein fragments produced by cytosolic degradation and present these peptides at the cell surface. Each MHCI allele produces an allotype, a polymorphic protein, capable of presenting different repertoires of MHCI associated peptides (MAPs) defined by the variable shape of the peptide binding groove. Collectively, the repertoire of MAPs, the immunopeptidome, represents a vision of the self from the perspective of the adaptive immune system. The antigen presentation pathway also captures peptides derived from pathogens or abnormal proteins that will be recognized as non-self and may initiate an immune response.

In humans, the MHCI locus is polygenic as well as polymorphic and contains the major genes: HLA-A, HLA-B and HLA-C. At the population level, HLA polymorphism confers different fitness for diseases ranging from infection to autoimmunity to cancer; the presence of some HLA allotypes may be protective while others increase susceptibility. The mechanism of disease association for different HLA allotypes has yet to be described convincingly. Hypotheses include that certain allotypes present specific immunodominant peptides which mediate responses in infection or autoimmunity. Another possibility is the inherent differences in global peptide repertoire and antigen presentation by different HLA allotypes alter the T cell repertoire and subsequent immune responses.

The global dynamics of antigen production, presentation and recognition are central to effective immunosurveillance. Therefore, we studied how allelic diversity impacts expression

and antigen binding properties of different HLA allotypes. We also characterized the immunopeptidomes of 18 individuals presenting 27 HLA-A,B allotypes to elucidate the genetic origins of MAPs.

This master's thesis is presented in 4 chapters and 4 appendices. Chapter 1 introduces the role of antigen presentation in the adaptive immune system and outlines the research questions. Chapter 2 presents the central results from studies of differences in HLA allotype expression and peptide binding. Chapter 3 presents an article in preparation entitled 'the immunopeptidome presents selected portions of human genome with distinct features to CD8+ T cells'. Chapter 4 discusses and offers perspectives on results presented in this work. Appendix 1 through 3 include optimized protocols corresponding to results presented in chapter 2; appendix 4 contains contributions to separate article.[1]

Our findings highlight fundamental differences in absolute HLA expression and invite complete elucidation of the HLA allotype specific expression cycle to reveal different functional properties. We explore in detail the genetic origins of MAPs across 27 allotypes. Our results show that MAPs derive from a select portion of the transcribed exome (< 17%) since only 58% of genes generate MAPs and MAPs derive preferentially from adjacent regions. We annotate MAP source and non-source genes and used features to predict with good accuracy whether a given gene will generate MAPs. The notion that the immune system can monitor MAPs covering only a fraction of the protein coding genome has profound implications in autoimmunity and cancer immunology.

Chapter 1 - Introduction

1.1 The adaptive immune system

The adaptive immune system of jawed vertebrates has evolved with the central purpose of eliminating threats to the host at a cellular level. Two challenges are inherent to this goal: how to identify diverse threats at a molecular level and how to monitor a complex system with many hiding places. To address the first, genetic recombination and somatic hypermutation alter the DNA sequence of linear loci to generate receptors that recognize diverse targets. Diversity is also inherent in the allelic polymorphism of molecules that bind antigens (protein fragments) to present to these receptors. Host-wide surveillance is achieved through continuous presentation of antigens derived from intracellular and extracellular compartments to diversified receptors. Antigens derived from foreign proteins, for example in infection, pregnancy or transplantation, or aberrant proteins, in neoplastic or stressed cells, can initiate immune responses. When foreign antigens are detected by receptors, a process of clonal selection - proliferation of the cell expressing the recognisant antigen binding receptor - engages a host-wide response to identify and eliminate the specific threat.[2–4]

1.1.1 Key components of adaptive immunity

The precision of the adaptive immune system depends on complex interactions between many subtypes of haematological cells which may be classified in two major branches: B cells and T cells. Each of these make use of clonally distributed antigen binding receptors, the B cell receptor (BCR) and T cell receptor (TCR), and rely on crosstalk for activation and survival. B cells are responsible for antibody mediated immune responses that identify structurally diverse targets in the extracellular environment. TCRs are restricted to recognizing protein fragments presented at the surface of antigen presenting cells (APCs). Professional APCs initiate an immune response by presenting peptides derived from foreign or aberrant proteins to T cells and by providing essential costimulatory signals to guide the expansion of appropriate subpopulations. A final cornerstone of adaptive immunity is immunological

memory. Once an immune response has been mounted, a subset of antigen detecting cells will differentiate into memory cells. Upon re-challenge by the same antigen, a swift protective immune response driven by memory cells will eliminate the threat - such is the principle of vaccination to stave off infection. The mechanisms of activation, cross-talk and memory in each arm of adaptive immunity have been reviewed extensively.[5–11]

T cell recognition is contingent of the participation of normal cells, APCs, in immune surveillance. Genes within the major histocompatibility complex (MHC) region of the genome bind and present peptide antigens derived from the extracellular and intracellular environments for MHC class II (MHCII) and MHC class I (MHCI) molecules respectively. MHCII molecules interact with T cells bearing the CD4 costimulatory receptor whereas MHCI are recognized by CD8+ T cells. MHCII expression is restricted to professional antigen presenting cells whereas MHCI is expressed on all nucleated cells. Functionally, MHCII antigen presentation stimulates 'helper' T cells to coordinate the immune response since MHCII antigens reflect the extracellular environment. Conversely, MHCI antigen presentation initiates a cytotoxic response from CD8+ T cells to eliminate cells harbouring pathogen derived or abnormal proteins. Over the course of an immune response, the activation of CD8 and CD4 T cells is coordinated by professional APCs while effector functions operate interdependently.[3,12–14]

1.1.2 Tolerance & discrimination by T cells

The T cell branch of the adaptive immune system uses somatic recombination to generate an incredibly diverse repertoire of T cells capable of recognizing unseen targets. T cells must discriminate between peptide antigens derived from host proteins and peptides that reflect a threat to the host. Self / non-self discrimination is achieved by T cell education in the thymus. Among developing T cells, only a minority survive the process of thymic selection which ensures an immunocompetent and self-tolerant T cell repertoire. Positive selection provides survival signals in the thymic cortex to T cells bearing receptors that recognize MHCI and MHCII molecules. Without these signals, T cells that do not interact with MHCI or MHCII on cortical cells will perish. In the thymic medulla, negative selection

eliminates autoreactive T cells with high affinity for self antigens bound to MHCI or MHCII. Medullary thymic epithelial cells display promiscuous expression of tissue specific genes to ensure T cells are tolerized to a comprehensive repertoire of self peptides. Thymic education therefore produces a repertoire of MHC-restricted T cells capable of discriminating between the self and the non-self. [15–17]

1.1.3 MHC I genomics & evolution

Finally, we arrive at the focal point of this work and a keystone of adaptive immunity: antigen presentation by MHC class I molecules. The MHCI α-chain is a transmembrane protein with a luminal peptide binding groove consisting of a basal beta sheet and lined on each side by α-helices (Figure 1). The binding groove can accommodate a variety of peptides derived from products of intracellular degradation. The MHCI gene is both polygenic and polymorphic. In humans, the three main gene loci are called HLA-A, HLA-B, and HLA-C. All exhibit exceptional levels of polymorphism, currently there are 3,356 HLA-A, 4,179 HLA-B and 2,902 HLA-C alleles documented.[18] Polymorphisms are essentially localized to the peptide binding grove.[19] The allele-specific structure of the groove translates to presentation of peptides with different binding motif generally defined by the electrochemical properties of anchor residues the P2 and PΩ (terminal) sites. HLA alleles can be organized into superfamilies that present peptides with similar binding properties based on the evolution of polymorphic loci.[20,21]

From an evolutionary point of view, MHCI and the surrounding regions are unique. The evolution MHC I is one of few examples of diversifying selection. In all likelihood, selection was driven by herd immunity and differential fitness in the face of a plethora infectious agents over time. By virtue of the fact that it contains many polymorphisms exhibiting linkage disequilibrium, the MHC locus has revealed much about ancestry, migration and selection in population genetics.[23–25]

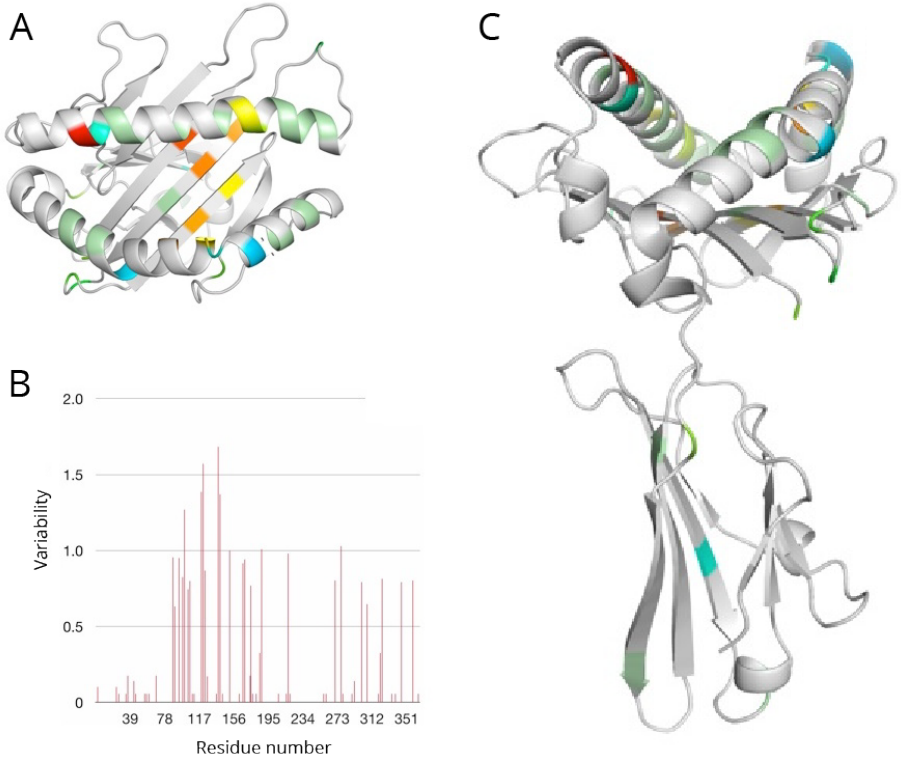

Figure 1. The structure and polymorphism of MHC class I molecules. (A) Top view of an HLA-A α-chain binding pocket, highly variable residues are coloured. (B) Variability of HLA-A residues within the binding pocket. (C) Side view of the HLA-A α-chain. Adapted from Gherardi.[22]

MHCI molecules are expressed on nucleated cells in normal tissues in a constitutive fashion with ~10^5 peptide-MHCI complexes at the surface of each cell although this varies based on cell type. HLA-C tend to be expressed at ~10% of HLA-A and HLA-B.[26–29] Secondary functions of MHCI include promoting neuronal plasticity, maternal-fetal interaction and olfaction.[23]

1.2 MHC I antigen processing & presentation

The pathways of MHCI antigen processing and presentation have been a major focus of research in immunology for the past three decades.[12,13,30] The pathway of class I processing

and presentation is well established.[13] However, the specific origins of MAPs and the relative contribution of different sources remains contentious.[30] The global dynamics of antigen production, presentation and recognition are central to effective immunosurveillance.

1.2.1 Classical antigen processing of endogenous peptides

The classical pathway of antigen processing (visualized in Figure 2) begins with the 20S proteasome, an enzyme structurally and functionally homologous to a food disposal unit one might attach to their kitchen sink. The barrel-shaped set of stacked multi-subunit rings is responsible for the degradation of a significant portion of cytosolic proteins. Proteins targeted for degradation by ubiquitination are first recognized by the proteasome cap.[13] Proteins are deubiquitinated, unfolded and fed into the barrel where proteolytic reactions produces fragments of roughly 3 to 20 amino acids.[31] Interestingly, the incorporation of alternate subunits in the caps and barrel of the proteasome dramatically alters the repertoire of MHCI presented peptide antigens.[32] Degradation products may be further trimmed by cytosolic peptidases before translocation into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP).

Meanwhile, the polymorphic MHCI α-chain is translated into the ER and undergoes multi-step glycosylation, resulting primarily in complex *N*-glycans. Beta 2 microglobulin (β_2m) associates with the α-chain and confers stability to the nascent MHCI. TAP imports peptides into the immediate vicinity of the peptide loading complex (PLC) machinery. ER localized aminopeptidases (ERAP1/2) conduct further trimming of potential MHCI peptides. Peptide-MHCI binding is facilitated by components of the PLC: the chaperone calreticulin (CRT), the disulphide bond isomerase ERp57, and the bridging protein tapasin (Figure 2). The PLC helps stabilize and induce conformational changes that facilitate peptide binding in the MHCI groove. In addition to acting as a scaffold for PLC components, tapasin is also essential in the process of peptide editing, the exchange of peptides in favour of those with higher affinity. High affinity peptide-MHCI-β_2m complexes (pMHCI) are released from the PLC and exported via the Golgi secretory pathway to the cell surface.[13,33]

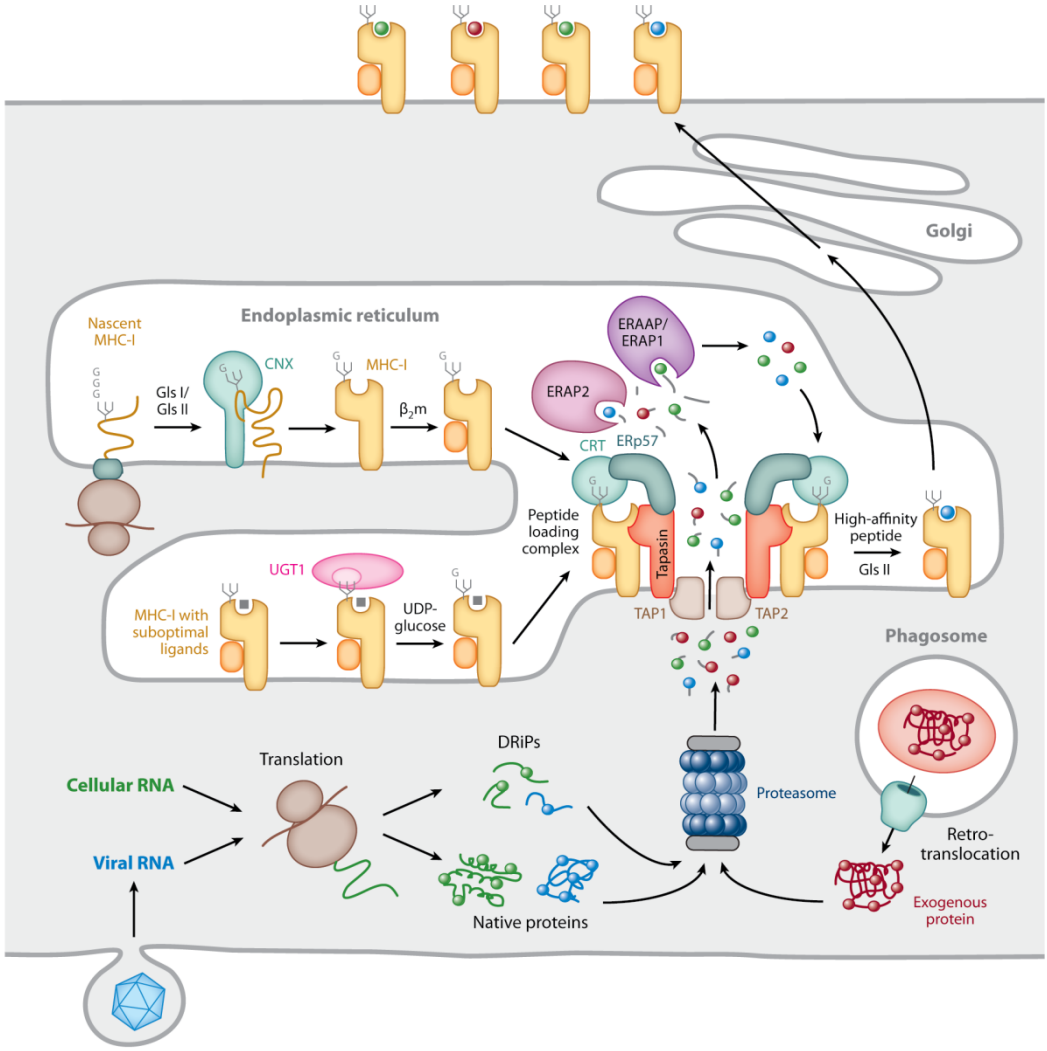

Figure 2. Pathways of MHCI processing and presentation. Adapted from Blum et al.[13]

Peptide-MHCI expression is dependent on the binding affinity of the particular peptide for the MHCI binding pocket; when the peptide dissociates, the MHCI-β_2m complex becomes less stable and is internalized for recycling or degradation.[12,34–36] The past 15 years have seen elucidation of the mechanisms and roles of PLC components. The salient finding is that the PLC is essential in shaping immune responses through the process of peptide selection.[33]

1.2.2 Cross-presentation of exogenous peptides

An alternate pathway for MHCI presentation of exogenous peptides, cross-presentation, resembles the MHCII pathway and makes use of MHCII machinery.[23] Cross-presentation may occur via the cytosolic or vacuolar pathways. The cytosolic pathway by definition is the presentation of exogenous peptides that is sensitive to proteasome inhibition. Processing occurs within the cytosol while peptide loading may occur via the classical ER pathway or within the phagosome (Figure 2). The vacuolar pathway permits exogenous antigen presentation via phagocytosis of extracellular proteins and phagosomal degradation. MHCI molecules may arrive in the phagosome through recycling or trafficking of newly synthesized complexes. The contribution of cross-presentation to the MHCI peptide repertoire is minimal in normal tissues. Cross-presentation is particularly effective in some professional APCs such as dendritic cells and plays an important role in the early stages of immune activation by priming naive CD8+ T cells.[13,37,38]

1.2.3 Noncanonical pathways of antigen generation

The major source of MAPs is generally though to be rapidly degraded proteins (RDPs) from many contexts. RDPs may include excess subunits form multiprotein complexes, pioneer translation products and defective ribosomal products (DRiPs).[15] One study showed mRNAs carrying premature stop codons will undergo nonsense-mediated decay yet effectively produce MAPs, presumably during the pioneer round of translation.[39] Recently, MAPs deriving from introns, out-of-frame translation and antisense transcripts further implicated immature mRNAs in antigen generation.[40] The discovery of MAPs derived from traditionally non-coding regions is one example of how immunology intersects and informs our understanding of fundamental biological processes.[13]

Another pathway first proposed by Yewdell, the DRiP hypothesis, postulates that nascent prematurely terminated and misfolded proteins as well as defective mRNA are the major source of MAPs.[41] Evidence supporting DRiPs focus on the kinetics of antigen presentation: MAP are efficiently generation from stable viral proteins and MHCI presentation is swiftly abrogated upon translation inhibition. Selectively presenting newly

synthesized peptides may allow a cell to preferentially include non-self antigens in the MAP repertoire prior to viral interference with the canonical antigen presentation pathway.[39,40,42–44] While the DRiP hypothesis remains controversial,[45,46] it is becoming increasingly clear the immunopeptidome not merely a reflection of the proteome.[15]

1.2.4 The role of MHCI in activation of the CD8+ T cell response

All these roads lead to antigen presentation and potential for T cell recognition. The dynamics of recognition remain puzzling: T cells can recognize single agonists presented in a sea of self peptides.[47] The number of copies of each unique peptide-MHCI complex is estimated between 1 to 10^4 per cell and is MHCI allotype dependent.[15] Furthermore, each TCR can recognize upwards of a million different peptides bound to MHCI.[48] How specific recognition is achieved in these conditions remains difficult to explain. Once an immune response has been initiated, cytokines including interferons can upregulate HLA gene expression to facilitate recognition.[49] The liaison of TCR to peptide to MHCI describes the immunological synapse between CD8+ T cells and APCs. Signalling and activation following recognition at the synapse is defined by CD8 costimulation and the local immune environment.[47]

While many potential non-self peptides may be presented upon intracellular infection, the T cell response tends to be focused on a few immunodominant peptides. Immunodominance is shaped by antigen processing, MHCI loading, T cell specificity and pMHCI surface expression time. Expression time is intrinsically related to the stability of the complex and binding affinity to the MAP; higher affinity interactions promote stability and may induce more effective T cell responses.[33,50] Once an immune response is mounted, T cells demonstrate sticking sensitivity and can recognize even a single pMHCI target.[13] High affinity recognition by the TCR of a CD8+ T cell initiates a signalling cascade that induces cytotoxic killing of the target cell.[5]

1.3 Studying the immunopeptidome

Immunologists have long hunted MHC class I associated peptides (MAPs) to understand how the self is defined for T cells and explore potential therapeutic applications. Despite extensive knowledge of antigen processing, it is impossible to predict the composition of the MHCI peptide repertoire.[12] Early studies identified single peptides; as tens and hundreds of peptides were discovered so were alleles specific binding motifs.[27] The study presented in Chapter 3 identified 25,172 unique HLA-A,B bound peptides; the immune epitope database, a repository for discovered MAPs, describes 219,463 peptide epitopes identified to date.[51] This progression reflects an improvement in techniques employed to identify MAPs. Since the advent of high throughput genomics and proteomics, many groups have studied immunopeptidomes of various MHCI allotypes in various hosts.[14,40,52–56] Their contribution to our knowledge of the cellular origins of MHCI presented peptides is outlined in the introduction of Chapter 3. Methods to identify MAPs and other focal points of these studies including characteristics of MAP sequences and genetic origins are summarized in this section.

1.3.1 Diverse methods identify MAPs

Current methods to identify MAPs can be broadly classed into two categories: high throughput proteomics and *in silico* screening. The first relies on experimental evidence of MAPs via isolating and sequencing peptides using mass spectrometry (MS) while the latter predicts and subsequently validates potential MAPs.

The techniques and challenges involved in experimental characterization of the immunopeptidome are expertly reviewed by Caron et al.[57] Isolation of MAPs from APCs may be achieved through mild acid elution to release peptides from MHCI on the surface of live cells or through immunoprecipitation of pMHCI complexes from cell lysate. While the first method offers higher sample throughput and yield, the latter has greater specificity and sample flexibility. To sequence isolated MAPs, classical data-dependent MS uses known protein coding regions matched to MS spectra for large scale identification. Alternatively, targeted

data acquisition allows specific identification and quantification of predefined sets of MAPs. An emerging technique deemed data-independent acquisition (DIA) combines these two strategies but requires detailed libraries of peptide retention time and fragmentation patterns. MAPs present a unique challenge to MS because of variable sequences (compared to samples treated with proteases with specific cleavage sites) and intra-laboratory differences in acquisition that result in limited reproducibility. Only ~10% of spectra in a given experiment are confidently assigned to a MAP sequence; these assignments offer a certain if incomplete picture of the immunopeptidome. With each new generation of mass spectrometry instruments and techniques, the limits of detection are pushed to identify lowly expressed peptides and move towards a complete picture of the immunopeptidome.[56,57]

Alternatively, *in silico* MAP identification uses binding affinity predictions along protein coding sequences of a particular organism to predict a high affinity immunopeptidome. Binding affinity predictions are complex as they must be allele and peptide specific. Fundamentally, binding motifs are governed by electrochemical and structural rules which have made this problem quite manageable for artificial neural networks such as NetMHC.[58,59] One study using this approach estimated interallelic differences in binding affinity and diversity for several HLA-A and HLA-B allotypes.[50] Several studies have identified MAPs spanning de novo mutations in tumour cells using this method.[60] Certainly, the major limitation of *in silico* approaches is lack of information about antigen processing. As a result, the rate of false positives is estimated at ~90%.[54] The central question remains: will peptides predicted to bind MHCI actually undergo appropriate processing and be presented at the cell surface? To address this, such studies must move to relatively low throughput experimental validation such as T cell reactivity assays,[60,61] or pMHCI multimer staining of T cells with flow cytometry.[62]

The selection of methods is largely shaped by the goals and setting of each study. Those seeking to identify few targets can afford to whittle down many potential targets post-identification while those seeking a comprehensive picture of the immunopeptidome are limited to confident identifications. Each method demands different resources, requires specific samples and has limited detection thresholds. To achieve an ultimate goal, the large

scale identification of MAPs in clinical tissue samples, will perhaps require convergence of these methods. One could imaging a pipeline that marries genomic and bioinformatic profiling of a sample with a comprehensive database of experimentally defined MAPs followed by high throughput validation via flow cytometry or mass spectrometry. Such an approach will require a collaborative effort but has realistic potential to shape personalized therapies in a clinical setting (see section 1.4.2 for further details).

1.3.2 Structural features of MAPs

Studies of the sequences of MHCI presented peptides have produced two salient findings: MAPs are constrained in terms of binding motif and length. As described previously, the extensive polymorphism of MHCI HLA-A and HLA-B alleles is localized to the peptide binding groove which functionally translates to alleles-specific binding motifs (Figure 3). Motifs are generally defined at the second residue, P2, and the C terminal residue, PΩ, although exceptions exist such as HLA-B*08:01 which has a P5 anchor. Typically the P2 anchor relies on charged interactions,[25] while the PΩ anchor tends to be more hydrophobic including small aliphatic residues or aromatic residues.

In our study we identified peptides corresponding to conventional motifs for each of the 27 alleles studied; motifs of each set of nonamer are shown in Figure 3. Since we predicted binding affinities with NetMHC to assign peptides among potential HLA alleles and applied a strict predicted affinity threshold of <1250nM, our results can only reinforce known motifs.[58] Evidently, it is difficult to estimate the potential contribution of non-canonical binding motifs in multiallelic systems. MAPs that do not conform to the rules of prediction algorithms are likely underrepresented in the current literature.

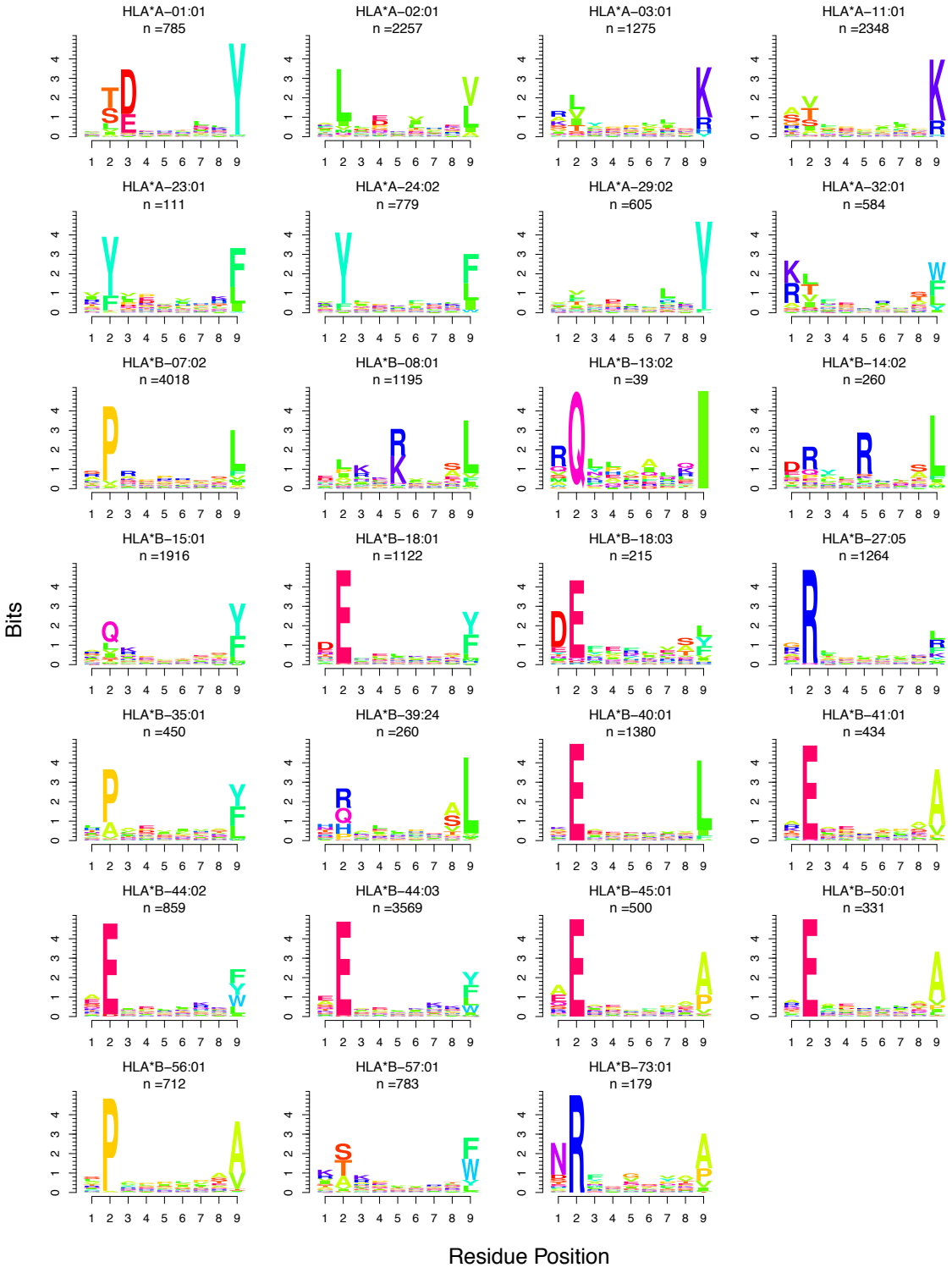

Figure 3. Binding motifs of nonamer peptides for 27 HLA-A & HLA-B alleles studied in Chapter 3. The number of peptides used to determine the motif

is indicated. Binding affinities predicted with NetMHC 3.0 and NetMHCcons 1.1.[63,64] Plotted with *motifStack* in R.[65,66]

The second constraint introduced by antigen processing machinery and HLA binding is on the length of MAPs. The canonical length of MHCI peptides is 8-11 amino acids although peptides up to 15 amino acids have been identified.[67] Ours and previous work reflect the dominance of nonamer peptides independent of the allele under consideration (Figure 4), although preference for other lengths appears to be allele dependent.[53,67] Structural studies of pMHCI have identified two modes of binding that allow the MHCI binding groove to accommodate longer peptides. Central bulging of the peptide allows the P2 and PΩ residues to fit in the same binding pockets as shorter peptides.[68,69] Alternatively, N or C terminal extensions of the peptide beyond the binding pocket are also possible.[70]

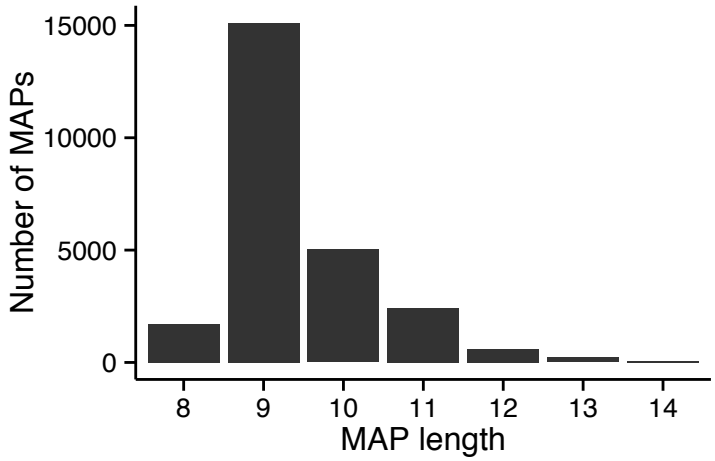

Figure 4. The length distribution of MAPs presented by 27 HLA-A & HLA-B allotypes studied in Chapter 3.

For all of these constraints, an incredible diversity of MAP sequences persists. Of course, this allows individuals expressing different allotypes to capture a representative array of peptides from self and non-self sources.

1.3.3 Genomic origins of MAPs

MAPs can be segregated in terms of their genomic origins: conventional antigens, cryptic antigens, minor histocompatibility antigens (MiHAs), and mutation-derived antigens (neo-MAPs). While the majority of studies, including the one presented in Chapter 3, focus on identification of conventional peptides, it is becoming increasingly clear that the immunopeptidome captures diverse genomic events (Figure 5).

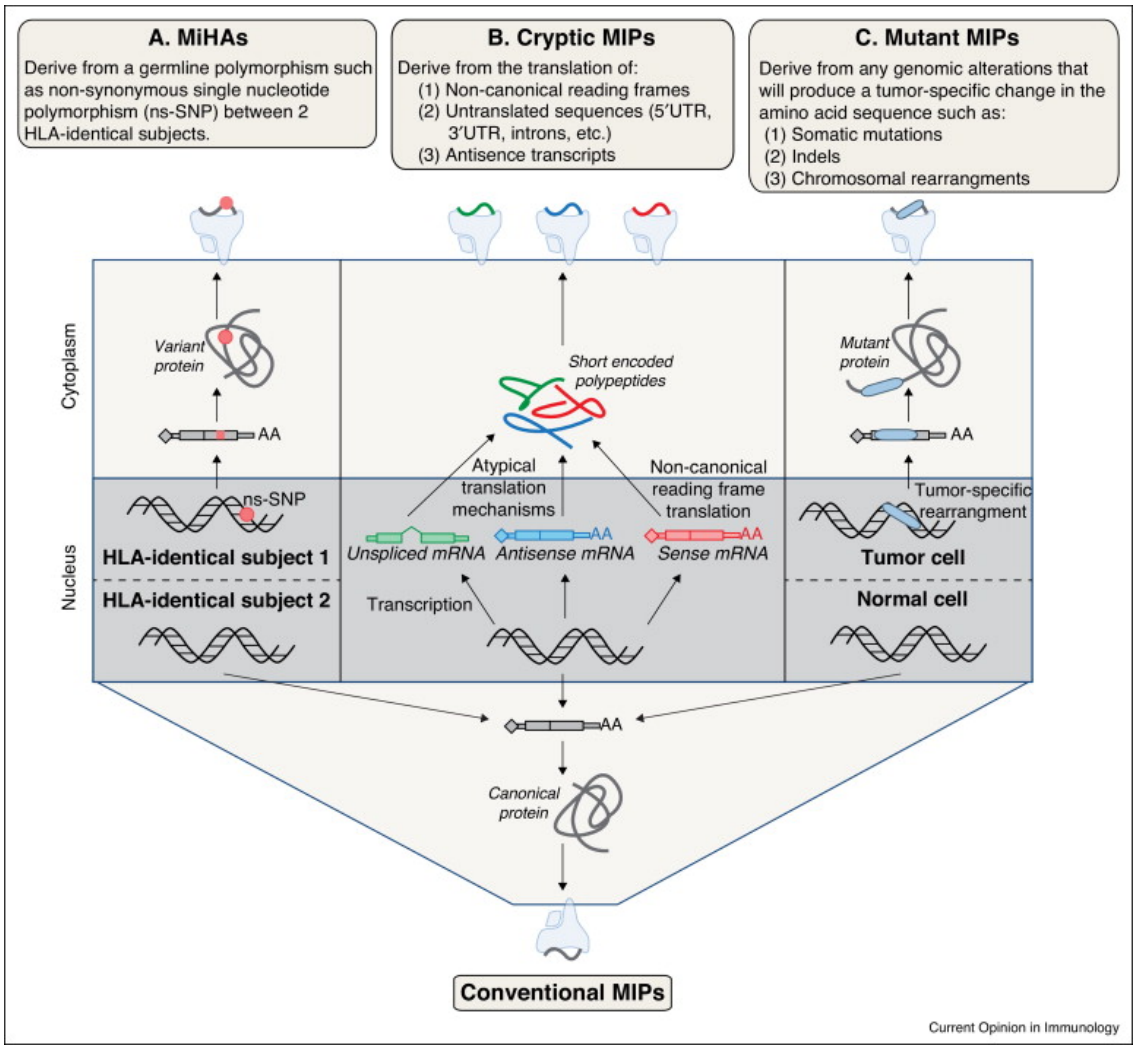

Figure 5. MAPs derive from diverse genomic origins. Origins include (A) MiHAs, (B) cryptic MAPs, (C) mutant neo-MAPs and conventional MAPs. Adapted from Granados et al.[15]

Conventional MAPs derive from consensus protein coding sequences, that is translation products of known protein coding genes that are shuttled into the antigen processing pathway. In contrast, cryptic antigens derive from non-canonical protein coding regions such as antisense transcripts, introns, UTRs, long non-coding RNA, and alternative reading frames. One study estimates the proportion of cryptic MAPs around 10%.[40] Potential sources of these antigens include i) pioneer translation products in the nucleus, ii) short open reading frames, iii) out-of-frame translation of mature transcripts, iv) translation of traditionally non-coding RNA and v) unstable transcripts undergoing nonsense mediated decay.[13,39,40,43]

The immunopeptidome also captures genomic alterations such as non-synonymous variants, mutations, and rearrangements with peptides spanning these transformations. MiHAs are peptides with non-synonymous genetic polymorphisms (nsSNPs) contained in their sequence.[1] CD8+ T cells from individuals with different alleles at a MiHA loci recognize the MiHA generated from the alternate allele as non-self. MiHA recognition contributes to graft versus host disease in bone marrow transplants between HLA matched donor-recipient pairs.[72] A final class of peptides, neo-MAPs, are derived from mutations or genomic rearrangements. From a clinical point of view, these antigens are of particular interest as they may be uniquely expressed on a particular subpopulation, such as neoplastic cells, and could be used to identify and target harmful cells for destruction (see Section 1.4.2).[15]

Some MAPs lack precise genomic origins or are incompletely described by nucleotide sequences. Studies of the proteasome have revealed a capacity to splice together peptide fragments and highlight the inclusion of splicing products in the immunopeptidome.[73] Peptide antigens may also contain diverse post-translational modifications.[25] Therefore, from a T cell perspective, a given MAP encoding sequence may generate structurally distinct epitopes.

The discovery of various classes of MAPs demonstrates that the antigen presentation pathway captures the genomic and translational complexity of each cell. For immunologists concerned with identifying MAPs this has important implications. Techniques that rely on a reference genome will exclude the discovery of cryptic, polymorphic and mutant MAPs.

Presently, techniques relying on six frame translation of personalized RNA sequencing data alone may discover such antigens.[40] It is essential to be aware of these gaps in our knowledge and, with advances in proteogenomics, aim to develop techniques that produce an increasingly complete picture of the immunopeptidome.

1.4 The immunopeptidome in disease

Of any region in the genome, the MHC locus is associated with the most diseases.[23] Within the locus, MHC class I and II are responsible for the majority of associations due to their diversity and central role in modulating immune responses. The MHCI immunopeptidome projects fragments from the intracellular environment that reflect the metabolic events within the cell. From the perspective of a T cell or a biologist trying to identify subpopulations of cells, MAPs offer a wealth of targets that that are accessible at the cell surface and specific to the intracellular events of each cell. Recognition of non-self MAPs by T cells first and foremost allows the elimination of intracellular infection. However, when the distinction between self and non-self is confused, autoimmune disease results. CD8+ T cells naturally recognize mutated targets on neoplastic cells and may effectively stave off cancer for years;[74] cancer immunotherapy aims to adapt this highly effective target elimination system to enhance anti-tumoral responses.

1.4.1 MHCI in the pathogenicity of infection & autoimmunity

The advent of genome wide association studies (GWAS) has validated associations between the MHC region and disease phenotypes. However, the MHC locus on chromosome 6 contains such an exceptional density of polymorphism, epistasis, and functionally related genes that it has been difficult to tease out exact mechanisms. CD8+ T cell activation mediates both infection and autoimmune disease, potentially through recognition of non-self and self peptides respectively.[23]

The importance of MHCI presentation in combating infection can be illustrated by viral genomes, which face strong selection and must be highly economical yet have developed a

plethora of mechanisms to impede antigen presentation.[75] Conversely, infection is considered a major selective pressure driving polymorphism in MHCI. As viruses mutate and evolve, different alleles may be selected for their ability to present immunogenic viral epitopes and mediate elimination of infections.[23] Alternatively, inherent differences in antigen presentation by different HLA allotypes will shape the T cell repertoire and subsequent immune responses. One well studied example is the differential ability of MHCI allotypes to control HIV progression conferring protection or susceptibility. For example, HLA-B*57 is associated with lower viral load and slower decline in the number of CD4+ T cells. A SNP linked to HLA-C expression levels has also independently been tied to HIV control.[76]

The strongest genetic risk factors for autoimmune diseases are consistently ns-SNP loci within the class I and class II genes. The subset of diseases characterized by autoantibodies are strongly linked with MHCII while other diseases tend to be linked with MHCI.[77] One of the most potent associations ties susceptibility to ankylosing spondylitis to residues within the binding grooves of HLA-B*27 and polymorphism of the PLC aminopeptidase, ERAP1.[78] A mechanistic hypothesis to explain these associations is that a peptide uniquely processed by ERAP1, in complex with HLA-B*27 is structurally homologous to a non-self antigen. This 'arthritogenic peptide' becomes a target for cross-reactive T cells and autoimmune attack ensues. Studies of the immunopeptidomes of linked and unlinked HLA-B*27 allotypes have had modest success identifying the elusive arthritogenic peptide.[79,80]

Elucidating the role of MHCI in infection and autoimmunity has followed a similar progression. GWAS have been indispensible in implicating different MHCI alleles in disease although the precise mechanisms remain elusive due in part to the aforementioned challenges of studying the immunopeptidome.

1.4.2 Cancer immunotherapy

The potential for antigens within the immunopeptidome to specifically identify and mediate targeted destruction of tumour cells is of significant interest for our work. Several lines of evidence support immune recognition of neoplastic cells through neo-MAPs. First, meta analyses have shown a correlation between the mutational load and immune infiltration

of a tumour. Second, checkpoint blockade antibodies such as anti-PD1 and anti-CTLA4 - which essentially release the breaks on T cell activation - have effectively cured some metastatic cancers.[74,81,82]

Targeted methods in immunotherapy aim to enhance the precise interaction mediating tumour elimination and reduce off-target side effects. For example in adoptive cell therapy (ACT), autologous tumour infiltrating lymphocytes or genetically engineered T cells that recognize neo-MAPs are cultured and selected for tumour recognition *ex vivo*. Reactive T cells are administered to the patient following lymphodepletion to favour a focused anti-tumoral response. Evidently, the complex protocol of ACT has required challenging optimization every step of the way but recent successes, particularly in cases of metastatic melanoma, are promising.[83] A similar approach harnesses native anti-tumour response by administering a peptide vaccine with suitable adjuvants to prime and activate T cells recognizing neo-MAPs.[81,84]

A major limiting step for large scale clinical implementation of both methods is the identification of suitable targets. The search for targets has revealed one major finding: effective neo-MAPs must be considered non-self from the host T cell perspective. The mechanism of negative selection (elimination of autoreactive T cells) limits the immunogenicity of self peptides. Furthermore, if self peptides are immunogenic, they can induce autoimmune-like toxicity towards other tissues.[83] Ideal neo-MAP targets therefore derive from alterations specific to neoplastic cells such as nonsynonymous mutations, frameshift mutations or rearrangements. If neo-MAPs deriving from common driver mutations in cancer exist and could be identified, a regularized therapy could be implemented for cohorts of patient. Alternatively, novel methods that efficiently predict or experimentally identify neo-MAPs derived from unique mutations offer a solution with broader applications.[85] Unfortunately, such MAPs are difficult to identify using current DDA MS techniques which rely on reference databases of non mutated proteins.[57] An ideal therapy might employ a multi-target approach to match intra-tumoral heterogeneity and thwart immune escape. A better understanding of origins of the immunopeptidome and continual

improvements in proteogenomics should facilitate neo-MAPs identification through predictive or experimental means.

1.5 Research context

1.5.1 Research objectives

MHCI is the centerpiece of adaptive immune surveillance and shapes the progression of numerous diseases. Questions revolving around MHCI antigen presentation are therefore of both fundamental and clinical importance. A fundamental question facing immunologists is the extent of differences in expression and peptide presentation brought about by HLA polymorphism: what is the impact of HLA allelic diversity on expression and binding of peptide repertoires? We were also driven by the question of the genetic origins of the self peptide repertoire presented by MHCI: from a T cell perspective, what is the self? To answer these questions, we make use of cutting-edge proteomics, genomics, and informatics to discover and analyse the immunopeptidome.

We hypothesized that variation in MHCI expression would be both allotype and subject dependent. We also explored linear relationships between MHCI expression and other variables including peptide diversity or binding affinity. Regarding the genetic origins of the immunopeptidome, we hypothesized that MAPs would derive from a distinct subset of genes and gene products with common features that may be related to antigen processing.

This work has 6 primary experimental objectives addressed in chapter 2 (objectives 1-3), and chapter 3 (objectives 4-6).

1. To determine absolute abundance of MHCI expression on B-LCLs.
2. To compare inter-individual and inter-allotype differences in MHCI expression and peptide presentation.
3. To devise and overall estimate of binding affinity and diversity for MHCI peptides.
4. To identify MAPs from a broad population of HLA allotypes and different subjects.

5. To asses the extent of MAP generation from the entire set of protein coding genes.
6. To determine whether specific features influence the ability of discrete genes to generate MAPs.

1.5.2 Model cell lines

We chose to study MHCI expression and antigen presentation on human B cells transformed by Epstein-Barr virus (EBV) infection. The resulting immortalized B lymphoblastoid cell lines (B-LCLs) are quite amenable to *in vitro* culture and offer the following advantages:

i. B-LCLs closely resemble primary B cells;[86]
ii. B cells in PBMCs from most individuals are easily EBV transformed, therefore we were able to study multiple individuals bearing many HLA allotypes;
iii. B-LCLs grow in suspension and therefore do not require protease mediated digestion for analysis, which would otherwise cleave surface MHCI;
iv. B-LCLs express relatively high levels of MHCI, therefore fewer cells are required for the high-throughput proteogenomic pipeline.

Chapter 2 - Studies of MHCI expression & peptide presentation

HLA polymorphism shapes the selection and development of CD8+ T cells by presenting immunopeptidomes with different structures and diversity. The impact of HLA polymorphism on peptide binding is well described: each allotype bind peptides with particular residues in the appropriate anchor positions and demonstrates a predictable binding motif. However, the impact of HLA polymorphism on surface expression has yet to be comprehensively described. MHCI antigen presentation shapes each step in the development and responses of CD8+ T cells; recent work has highlighted the potential influence of HLA expression in thymic selection and disease phenotypes.[29,87,88]

With newly available quantitative flow cytometry techniques, we set out to determine the absolute abundance of HLA molecules and compare inter-allotype differences in MHCI expression. We hypothesized differences in the overall binding affinity and diversity of the peptide repertoire would be related to differences in HLA expression. In the context of large scale elutions studies, we studied mild acid elution (MAE) efficacy for different HLA allotypes. Additionally, we present preliminary results of allotype specific expression recovery following MAE or proteolytic cleavage. Finally, we investigated an important subset of MAPs, minor histocompatibility antigens, as part of a larger study.[1]

2.1 Methods

The study protocol was approved by the Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont. Written informed consent was obtained from donors. B cells from 18 donors PBMCs were transformed with Epstein-Barr virus to generate immortalized B lymphoblastoid cell lines (B-LCLs) as previously described.[56] B-LCLs were maintained in RPMI 1640 supplemented with 10% FBS, 1% penicillin/streptomycin and 1%

L-glutamine at 37°C. Flow cytometry analysis was conducted on a BD FACSCANTO II. Statistical analyses were carried out in R version 3.1.3, data visualization was achieved with the *ggplot2* package.[65,71] Detailed protocols for each analysis are included in the appendices for reference:

Annexe 1. Protocol for QIFIKIT quantitation of MHCI expression on B-LCLs.

Annexe 2. Protocol for mild acid elution of surface MHCI peptides on B-LCLs.

Annexe 3. Protocol for papain digestion of surface MHCI on B-LCLs.

2.2 Quantitative analysis of MHCI expression

Routine flow cytometry (FC) is an immensely powerful technique to comparatively assess expression of molecular markers on 10^4-10^7 cells in a single experiment. It employs fluorescently labeled antibodies with strong affinity and high specificity for their targets to measure protein expression. Due to sensitivity to changes in instrument parameters and experimental conditions, traditional FC results are limited to relative comparisons. To quantify FC fluorescence in absolute terms, the QIFIKIT relies on the fundamentally linear relationship between fluorescence and number of antibody-bound fluorescent molecules. Control populations of beads with a known number of receptors are used to generate a calibration curve that translates MFI of a secondary antibody into specific antibody binding capacity (SABC) or the approximate number of molecules per cell. The challenges of exact quantitation via FC include i) untangling the contributions of non-specific staining, ii) mitigating limiting factors such as antibody concentration, and iii) limiting inter-experiment variation. To maximize the precision of our analyses, we examined the influence of total cells, antibody concentration, blocking Fc receptors, washing protocol and instrument parameters. The optimized parameters are noted in the protocol in appendix 1. Results show the average of 3 independent experiments.

First, we quantified MHCI expression using a pan HLA-A,B,C antibody. Globally, MHC class I expression was 1.1 x 10^6 molecules per cell with ± 16% inter-individual variation

(Figure 6A). Using the same technique, Berlin et al. found between 50,000 and 300,000 MHCI molecules expressed on acute myeloid leukemia cells and benign leukocytes.[89] We conclude that B-LCLs exhibit particularly high MHCI expression. Next, we quantified expression of 4 common HLA allotypes: HLA-A*02:01, HLA-A*03:01, HLA-A*11:01, and HLA-B*07:02. We found relatively consistent expression of each allotype across cell lines (Figure 6B). A hierarchy of allotype expression emerged: A*02:01 > A*03:01 > B*07:02 ≈ A*11:01. Even when normalized to global HLA-A,B,C expression in each cell line, the relative contribution of each allotype to surface expressed was consistent. Finally, to estimate the variance attributable to environmental vs. genetic factors, we compared the complete profiles of HLA-A,B expression on B-LCLs derived from monozygotic twins (subjects 8 and 9, Figure 6D). We found very similar profiles of expression between these two subjects in global and allotype-specific quantifications. We noted that differences in the process of EBV transformation may influence MHCI expression.[90] This analysis also confirms that HLA-A,B expression makes up for the majority of HLA-A,B,C levels.

One hypothesis to explain these results is that allotypes forming more stable pMHCI complexes persist on the surface longer and exhibit greater surface expression. The fact that different alleles demonstrate variable increases in expression in homozygous conditions may support this hypothesis. For example, we could imagine in the case of HLA-A*02:01, peptides efficiently form highly stable complexes allowing homozygous cell lines to almost double expression. In contrast, if peptide-HLA-B*07:02 complexes are less stable, additional proteins in homozygous cell lines may be retained intracellularly during quality control steps, or recycled from the surface at a greater rate. Notably, transcript level expression for HLA-A was ~50% of HLA-B genes and was similar in all cell lines. Other factors shaping MHCI expression may include protein abundance, peptide supply, the restrictiveness of binding motifs,[50] or allotypic differences in tapasin and TAP interaction.[35,91,92] Globally, a combination of these factors may define inter-allotypic variation in MHC expression. We conclude surface expression is an intrinsic feature of each allotype.

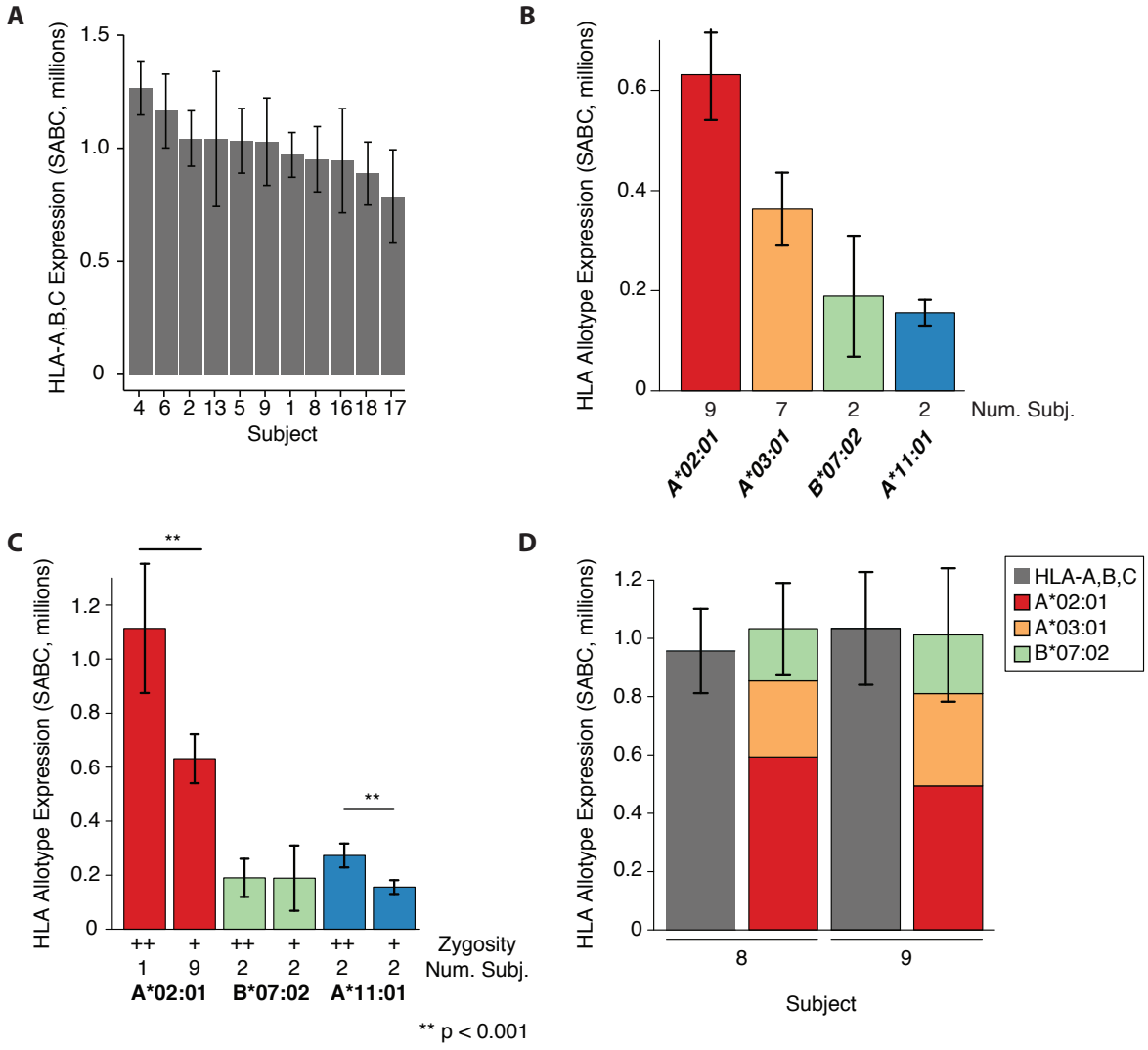

Figure 6. Absolute global and allotype specific HLA expression on B-LCLs. (A) Absolute expression of HLA-A, B and C molecules on 11 B-LCL cell lines. (B) Absolute expression of 4 HLA allotypes in heterozygous cell lines. The number of subjects is indicated. (C) A comparison of absolute expression of 3 HLA allotypes in B-LCLs with homozygous and heterozygous genotypes. The number of subjects and zygosity are indicated. (D) Complete expression profiles of HLA-A,B allotypes in monozygotic twin subjects. Results are the average of 3 experiments. Surface expression is measured in Specific Antibody Binding Capacity (SABC), details in appendix 1. The HLA alleles expressed by each subject are indicated in Table I.

2.3 The efficiency of mild acid elution is HLA allotype dependent

In large scale proteogenomics studies of the immunopeptidome, MAPs can be isolated from surface MHCI molecules using mild acid elution (MAE) to release peptides from MHCI binding pockets. With this technique, studies have found differences in the diversity of the peptide repertoires detected for different allotypes.[42,55,56,93] One hypothesis among many is that MAE efficacy differs by allotype and contributes to differences in the number of identifications. We investigated this question with small scale MAE studies following the protocol in appendix 2. Results are representative of at least one experiment on 2 cell lines per allotype in technical triplicate.

Preliminary results suggest MAE liberates peptides in an allotype dependent manner (Figure 7). Relative to untreated cells, we saw a global decrease in expression for all alleles following MAE, reflecting the instability of the MHCI α-chain without bound peptide.[94] We initially remarked poor efficacy of elution for A*03:01 and A*11:01, which belong to the same superfamily and have a similar binding motif of T/L/V at P2 and K dominating PΩ (Figure 3). MAE was considerably more efficient for B*07:02 which has a P2: P, PΩ: L motif. Finally, MAE was most effective for A*02:01 which has a P2: L and PΩ: V/L. Interestingly the hierarchy of elution efficacy could be related to the types and strengths of interactions mediating peptide binding for each allotype. It appears MAE more efficiently elutes peptides relying on hydrophobic interactions such as L and V anchors than polar or charged anchors such as T, K, and P. Further investigation is warranted to confirm this hierarchy across cell lines and allotypes.

Importantly, these experiments were preformed with ~6 x 10^5 cells whereas an immunopeptidome MAE experiment uses ~5 x 10^8 cells to have enough material for MS analyses. The applicability of these findings on a larger scale is not yet clear. This question could be answered simply by incorporating allele specific antibodies and flow cytometry analysis following a large scale elution study.

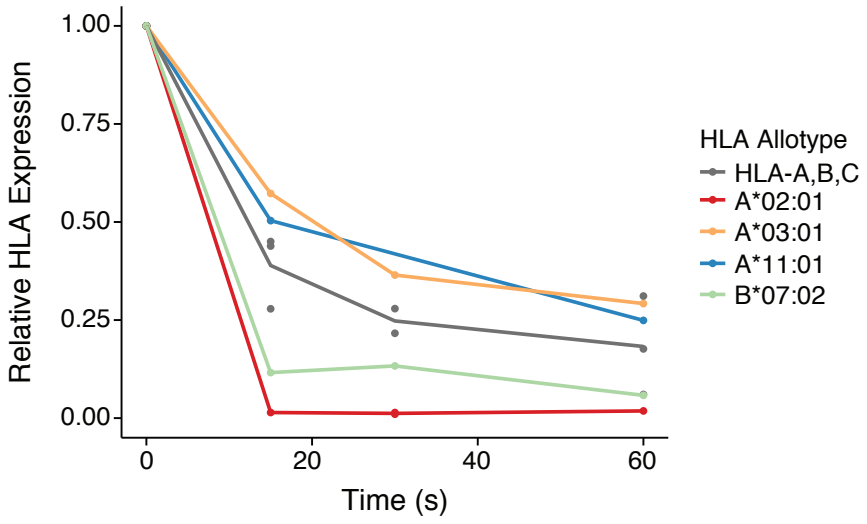

Figure 7. Relative MHCI expression of 4 HLA allotypes and global HLA expression at during mild acid elution lasting 15, 30 or 60 seconds. Each point represents a single experiment on one cell line expressing each allele heterozygously, analyses were performed in technical triplicate.

To limit the potential bias in the MAP repertoire introduced by allotype specific MAE efficacy, a longer elution time could be used. Residual MHCI expression after 5 minutes of MAE is negligible (< 5%) independent of allotypes although longer elution periods may lead to cell death.

2.4 Recovery of HLA expression

We subsequently investigated inter-allotype differences in recovery of MHCI expression following either MAE or proteolytic cleavage. The papain protease, naturally produced by papayas, has been used in immunopeptidome studies to cleave surface MHCI molecules.[95,96] We expected some variation in recovery given established inter-allotype differences in tapasin dependence and peptide editing,[97] in the occurrence of the allotypic binding motifs in the proteome,[50] and in absolute allotype expression. We hypothesized allotypes demonstrating more efficient peptide loading would recover expression more quickly.

Using the MAE and papain protocols detailed in appendices 2 & 3, we carried out a single experiment using triplicate analysis of 2 cell lines for each allele studied.

First, our findings indicate B-LCLs recover HLA-A,B expression following MAE and protease treatment at different rates (Figure 8). Uniquely in the case of MAE, recycled MHCI may contribute to recovery. However, recovery in the MAE condition is less efficient which invites the possibility that this treatment may influence cellular metabolism. A similar hierarchy is found in both treatments, A*02:01, A*03:01 and B*07:02 demonstrate comparable recovery whereas A*11:01 is re-expressed more slowly. Altogether, these findings suggest that steady state MHCI expression is minimally influenced by the export of pMHC complexes for 3 of 4 alleles studied. Consequentially, we hypothesize the rate of internalization of MHCI - linked to the net stability of allotype-peptide complexes - may be a determinate factor in absolute expression, as Meyadera et al. have shown for MHCII.[98]

To move forward with these results, more alleles must be studied. Importantly, the current literature would classify all 4 of these allotypes as independent or only moderately dependent on tapasin, which influences peptide editing.[92,97] To assess whether tapasin alters the speed of peptide loading one could compare the tapasin dependent B*44:02 or B*08:01 allotypes.

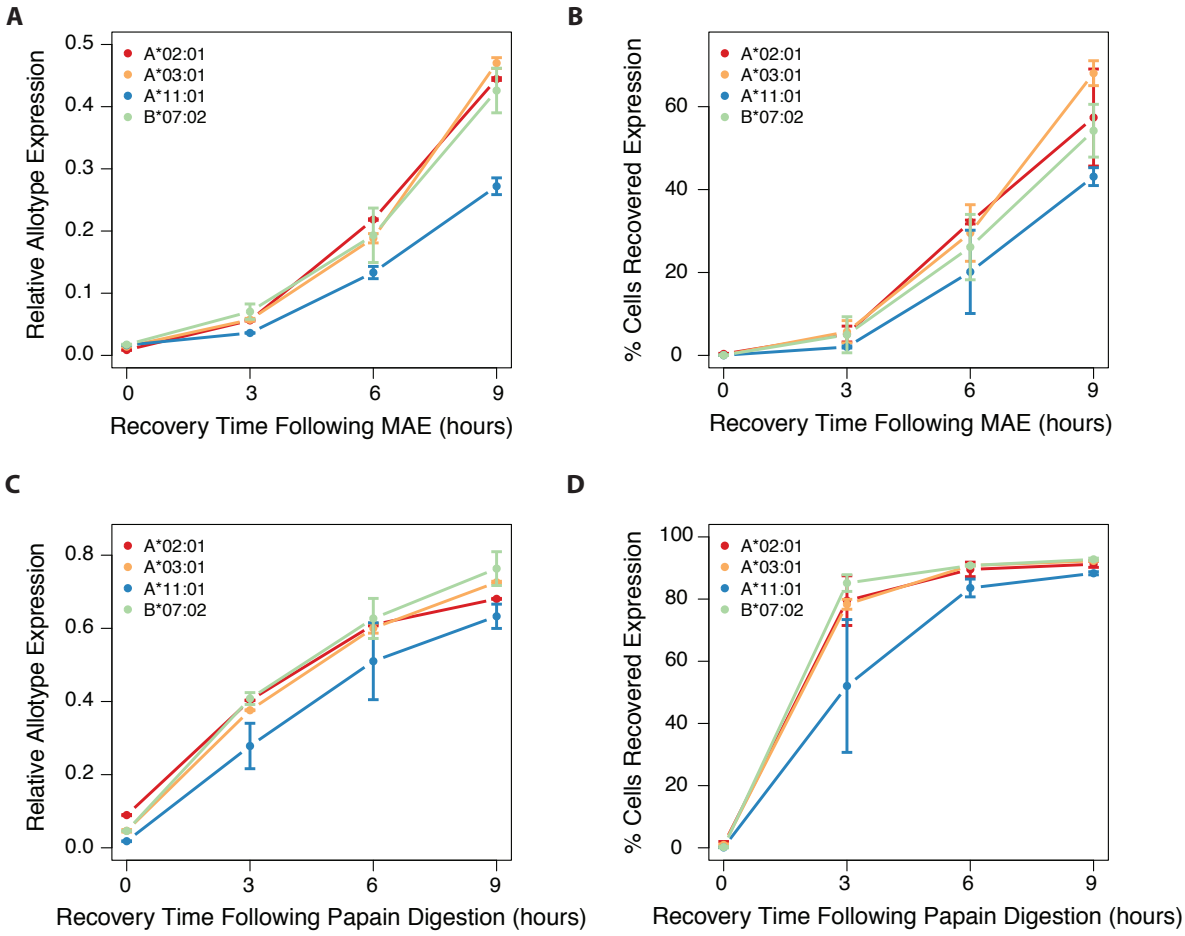

Figure 8. Recovery of MHCI expression over 9 hours following MAE or papain digestion. Recovery following MAE (A) or papain protease digestion (C) relative to untreated control. Proportion of cells with full recovery relative to untreated control following MAE (B) or papain protease digestion (D). Results of a single experiment of triplicate analysis of 2 cell lines heterozygously expressing each allotype studied.

2.5 Estimating the diversity and binding affinity of HLA allotype peptide repertoires

An emerging hypothesis postulates that HLA allotype expression is inversely correlated with the diversity of allotype-specific peptide repertoires.[50,93,88] Chappell et al. have proposed that some alleles are undergo selection as 'generalists' to present diverse repertoires with relatively low expression while other alleles are 'specialists' presenting a fewer peptides with high expression. Generalists may manage presentation of common pathogens while specialists may be selected for swift presentation of particularly virulent emerging pathogens.[29] We explored this hypothesis using experimental data describing the immunopeptidomes of 18 subjects (see Materials & Methods, Chapter 3.5). Predictions of peptide-allotype binding affinity by NetMHC 3.0 were also incorporated as a rough estimate of the stability of MHCI complexes.[63]

A range of peptide diversity and binding affinity are seen across the 21 allotypes studied (Figures 9A & B). A single allotype on a given cell line could present from 99 to 2,674 peptides. The allele under consideration had significantly more influence on diversity than the subject (Two way ANOVA, $p = 5.5 \times 10^{-6}$ for allele, $p = 0.02$ for subject), leading us to believe diversity is truly an allotype-specific phenomena. Similarly, the geometric mean of predicted binding affinities was shaped primarily by allotype (Two way ANOVA, $p < 6.5 \times 10^{-20}$ for allele, $p = 0.04$ for subject). The correlation between binding affinity and diversity is negligible (Spearman's correlation coefficient $\rho = 0.02$). Our results show limited concordance with another study of *in silico* predicted immunopeptidomes of multiple HLA-A,B allotypes based on binding affinity.[50] This discrepancy reflects, at least in part, the selection of potential MAPs during antigen processing steps and illustrates the importance of *in vitro* MAP identification.

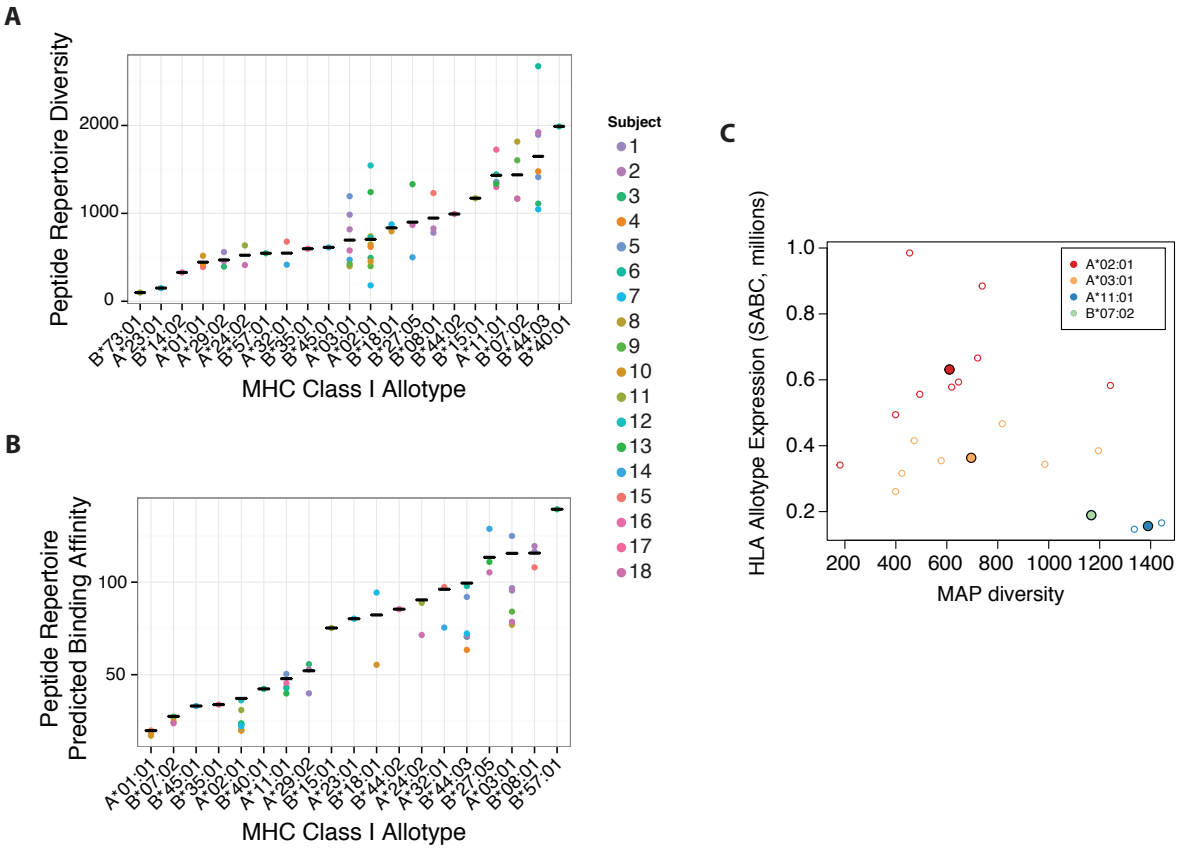

Figure 9. Binding affinity, diversity, and expression of MHCI allotypes. (A) Peptide repertoire diversity of 21 HLA-A,B allotypes. (B) Geometric mean of predicted binding affinities for the peptide repertoires of 21 HLA-A,B allotypes. Results for each cell line shown with points, mean per allotype indicated with a bar. (C) Diversity and surface expression of 4 HLA-A,B allotypes. Grand means shown with filled in points, individual cell lines shown in outlined points. Expression averages are shown for 3 independent experiments.

We examined the correlation of expression and diversity for the 4 allotypes with quantitation data (Figure 9C). Our results appear consistent with an inverse relationship between diversity and HLA allotype expression although we see significant inter-individual heterogeneity (Pearson correlation r = -0.89 for grand means, r = -0.39 for individual cell lines). The principles of the specialist-generalist model are intriguing however our results - especially in terms of allotype diversity - suggest a spectrum rather than a dichotomy of HLA allotype grouping.

2.6 Features of minor histocompatibility antigens

The MiHA subset of the immunopeptidome, those peptides that span nsSNP, are enticing targets in the context of allogeneic hematopoietic cell transplantation (AHCT) to treat leukemia. A MAP containing a nsSNP present uniquely in the recipient may be recognized by donor T cells that have not been tolerized to this particular peptide. AHCT can therefore mount an immune response against neoplastic host cells expressing MiHAs in the graft-vs-leukemia effect.[99] In collaboration with several groups, we conducted a large scale analysis of MiHAs presented by B-LCLs.[1] This authors' contributions included assistance in experimental procedures, annotation of MAPs, and a study of the promiscuity of MiHA binding.

We first asked in what contexts have the 100 discovered MiHAs been previously described. By conducting a literature review as well as consulting the SYFPEITHI database[100] and the Immune Epitope Database,[51] we found the majority of MiHAs were novel (62%) while the rest were documented as either MHCI binders (16%) or MiHAs (22%). We annotated the genetic origins of MiHAs using pyGeno[101] and predicted the binding affinities of each MiHA with NetMHC 3.4[63] for the two HLA allotypes studied: HLA-A*02:01 and HLA-B*44:03 (results presented in appendix 4).

Next, we investigated the extent of promiscuous binding - the capacity to bind more than one HLA allotype - of all MiHAs. We predicted the binding affinity for each MiHA and the top 1% most frequent HLA-A,B allotypes in the USA European Caucasian cohort.[102] Predictions were obtained with NetMHC 3.4 or NetMHC cons 1.0 for 35 allotypes.[63,64] Only 42% of discovered MiHAs had the highest affinity for the allotype on which they were discovered; other alleles, notably HLA-A*32:01 for HLA-A*02:01 peptides and HLA-B*18:01 or HLA-B*40:01 for HLA-B*44:03 peptides, could bind the same peptides with greater affinity. Out of the 35 most frequent HLA-A,B allotypes, a given MiHA could bind upwards of 12 different allotypes with and affinity < 5,000 nM (Figure 10).

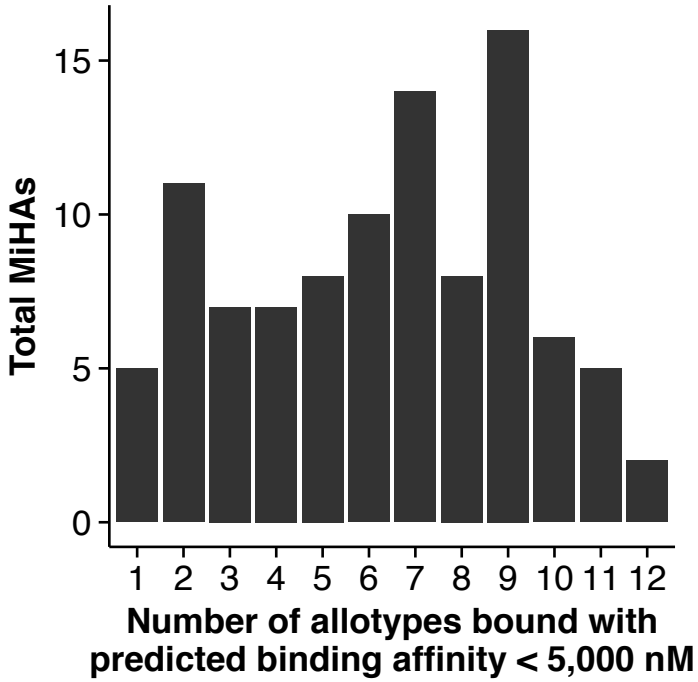

Figure 10. MiHA promiscuity: the number of allotypes predicted to bind each MiHA with an affinity < 5,000 nM out of the top 35 most frequently occurring HLA-A,B allotypes.

Assuming similar modes of antigen processing between individuals, these results suggest that MiHAs identified for HLA-A*02:01 and HLA-B*44:03 could well be presented by other HLA allotypes. These findings are especially relevant for AHCT treatment of leukemia as the therapeutic applications of this pool of MiHAs could be extended to more individuals bearing other HLA alleles.

Chapter 3 - The immunopeptidome presents selected portions of the human genome with distinct features to CD8+ T cells

Hillary Pearson[1,2], Diana Paola Granados[1,2], Tariq Daouda[1,3], 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[1,4,6], Pierre Thibault[1,3,5,6*], Claude Perreault[1,2,6,7*]

[1]Institute for Research in Immunology and Cancer, [2]Department of Medicine, [3]Department of Biochemistry, [4]Department of Computer Science and Operations Research, [5]Department of Chemistry, [6]Canadian National Transplant Research Program, Université de Montréal, Québec, Canada H3T 1J4
[7]Division of Hematology-Oncology, Hôpital Maisonneuve-Rosemont, Montreal, Quebec, Canada H1T 2M4

This article is in preparation as of April 2016.

3.1 Abstract

Using proteogenomics, we identified 25,172 major histocompatibility complex class I-associated peptides (MAPs) isolated from B lymphocytes of 18 individuals who collectively expressed 27 HLA-A,B allotypes. While 58% of genes were the source of 1-64 MAPs per gene, 42% of genes were not represented in the immunopeptidome. Overall, we estimate the entire MAP repertoire presented by 27 HLA-A,B allotypes covered at most 17% of exomic sequences expressed in B lymphocytes. We identified several features of transcripts and proteins that enhance MAP production. From these data we built a logistic regression model that predicts with high accuracy whether a gene from our dataset or from independent datasets would generate MAPs. Our results show preferential selection of MAPs from a limited repertoire of proteins with distinct features. The notion that the immune system can monitor MAPs covering only a fraction of the protein coding genome has profound implications in autoimmunity and cancer immunology.

3.2 Introduction

Major histocompatibility complex class I molecules (MHCI) present thousands of peptides at the cell surface of nucleated somatic cells (Granados et al., 2015). These MHCI-associated peptides (MAPs), collectively referred to as the immunopeptidome, regulate each step in the development and function of CD8+ T cells (Govern et al., 2010; Chakraborty and Weiss, 2014). 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 (Butler et al., 2013; Caron et al., 2011; Vrisekoop et al., 2014). Genesis of the immunopeptidome can be broadly divided into two events: i) the processing of MAPs and ii) their binding to MHCI molecules (Yewdell et al., 2003; Hammer et al., 2007). 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 (Rammensee et al., 1999; Kim et al., 2014b).

However, the first event, processing of MAPs, is a complex multi-step process whose overall outcome cannot be predicted (Granados et al., 2015). Some proteins appear to generate more MAPs than others, but the mechanistic underpinning for these discrepancies remains elusive (de Verteuil et al., 2012).

Classic biochemical studies have shown that MAP processing is initiated by proteasomal degradation of cellular proteins, followed by further trimming by cytosolic peptidases, transport in the endoplasmic reticulum (ER) and final trimming by ER peptidases (Eisenlohr et al., 2007; Hammer et al., 2007; Vigneron and Van den Eynde, 2012; Rock et al., 2014; Blum et al., 2013). Other provocative studies suggest that MAPs preferentially originate from defective ribosomal products (DRiPs) and can be created by nonsense mediated decay, mRNA destabilization or noncanonical translation in the cytosol or the nucleus (Goodenough et al., 2014; Anton and Yewdell, 2014; Apcher et al., 2013; Granados et al., 2012; Laumont et al., 2016). Large-scale mass spectrometry (MS) offers the sole direct approach to analyze 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 (Caron et al., 2015a; Hickman et al., 2004; Mester et al., 2011; Hassan et al., 2013). However, the rules of MAP processing cannot be figured out by studying the immunopeptidome presented by a single HLA allotype because the peptide binding motif can bias the ability of different allotypes to present peptides coded by different genes (Hoof et al., 2012; Paul et al., 2013).

The goal of our study was 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 therefore used a well validated high-throughput proteogenomic approach in order 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 25,172 non-redundant MAPs, which derived from 6,231 out of the 10,677 genes expressed in B-LCLs. Hence, while 58% of genes were the source of 1-64 MAPs per gene, 42% of genes were not represented in the immunopeptidome. Overall, we estimate the immunopeptidome presented by 27 alleles covered at most 17% of exomic sequences expressed

in B-LCLs. We then used a series of bioinformatic tools to understand how 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 receiver operating characteristic area under the curve of 0.82. Our results show that the immunopeptidome is forged from a limited repertoire of gene products with distinct features influencing transcription, translation and proteasomal degradation.

3.3 Results

3.3.1 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 (Granados et al., 2014; Laumont et al., 2016; Granados et al., 2016). Transcriptome and exome sequencing data were used to build personalized protein databases for B-LCLs of 18 subjects using pyGeno (Daouda et al., 2016). 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: i) a peptide length of 8–14 amino acids, ii) a 1% false discovery rate based on searches against concatenated target/decoy databases (Elias and Gygi, 2007), and iii) a predicted MHCI-binding affinity <1,250 nM according to NetMHC or NetMHCcons algorithms (Lundegaard et al., 2008; Karosiene et al., 2012) (Figure 18A).

We identified 25,172 non-redundant MAPs which derived from 6,231 genes (Figure 11A, Table I). MAP source genes produced up to 64 individual MAPs and 68% of these genes produced more than one MAP (Figure 11B). To estimate the depth of a multi-allelic 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 11C). We

counted the number of unique identifications when a given number of randomly selected allotypes were considered. For MAPs, the nearly linear nature of this relationship demonstrates little redundancy in the peptides presented by different allotypes (Figure 11C, left panel). Conversely, the redundancy of the genes generating MAPs across all 27 HLA allotypes is much greater (Figure 11C, right panel). As more allotypes are considered, a diminishing number of unique genes are 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 (Figure 18B). Most MAPs (89%) were presented by a single HLA allotype (Figure 11D, left panel). The few promiscuous binders were presented by MHCI allotypes with similar peptide binding motifs (i.e., same "superfamily"), such as A*03:01 and A*11:01 (Sidney et al., 2008). In contrast, the majority of MAP source genes (67%) produced MAPs for multiple allotypes, some for up to 24 of the 27 allotypes studied (Figure 11D, right panel).

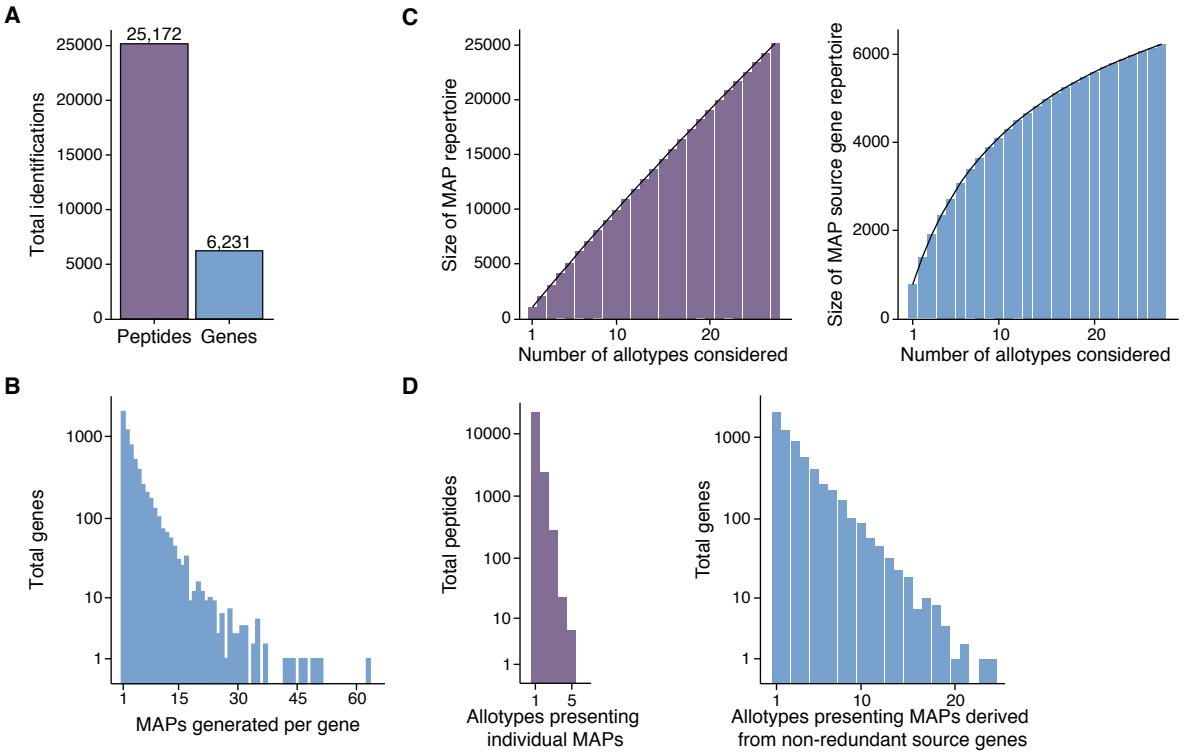

Figure 11. The depth and breadth of the multi-allelic immunopeptidome presented by 27 HLA allotypes. (A) Total number of non-redundant MAPs and their source genes in the immunopeptidome of 18 B-LCLs. (B) Histogram

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 as each additional randomly selected HLA allotype was considered. Results show the average of 1000 simulations. (D) The promiscuity of HLA presentation for MAPs (left panel) and their source genes (right panel). Histograms show the number of allotypes associated with each peptide or gene.

Two major points can be made from these data: i) a distinct subset of genes produce most MAPs and ii) our method captured the majority of MAP source genes (Figure 11C, right panel). As a corollary, these results suggest a model whereby a common pool of source proteins selectively enter the antigen processing pathway and can generate MAPs with suitable motifs for most MHCI allotypes.

3.3.2 Discrete protein regions are preferential sources of MAPs

We next asked whether there might be "hotspots" in MAP source genes, i.e., regions or domains that provide disproportionally high amounts of MAPs. To this end, we analyzed the spatial distribution of MAPs along proteins that generated multiple MAPs. We first identified 6,325 pairs of overlapping MAPs formed by 8,228 individual peptides (33% of our entire dataset). In a given pair, MAPs differed from each other at their N- and/or C-terminus (Figure 12A). These pairs may result from differential trimming of a common precursor by various peptidases in the cytosol and ER. Notably, 83% of MAP pairs bound different allotypes; of these, 48% bound allotypes from different superfamilies (Figure 12B). Hence, from the perspective of an MHCI allotype, generation of overlapping MAP pairs is 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 non-overlapping MAPs along proteins. For each protein, the distances between all MAPs were calculated. To exclude overlapping peptides, MAPs within 8 residues of each other were merged. A control distribution was generated by randomly placing the same number of MAPs along the same protein length. We

found that MAPs colocalized more than expected in both absolute and relative terms ($p = 6 \times 10^{-6}$ and $p = 4 \times 10^{-8}$ respectively, Figure 12C). We surmise that colocalization must result from short-range effects because MAPs were found within a window of ~25 amino acids (Figure 12C). The fact that no MAPs could be assigned to 42% of genes and that MAP coding sequences are clustered in source genes suggest that the immunopeptidome covers a limited portion of the whole exome. To estimate global exome coverage, i) we moved a walking window of 150 base pairs (50 amino acids, the rough length of the short range effect) along the exome coding for the 10,677 genes expressed in our B-LCLs, and ii) we calculated the number of MAPs seen in each window. We found that 83% of windows generated no MAPs whereas 17% of windows covered 1-11 MAPs per window (Figure 12D). When we reduced the window size to 75 base pairs, only 10% of windows were source of MAPs (data not shown). From this, we conclude that the immunopeptidome presented by 27 HLA-A,B allotypes covers an unexpectedly small portion of the whole exome.

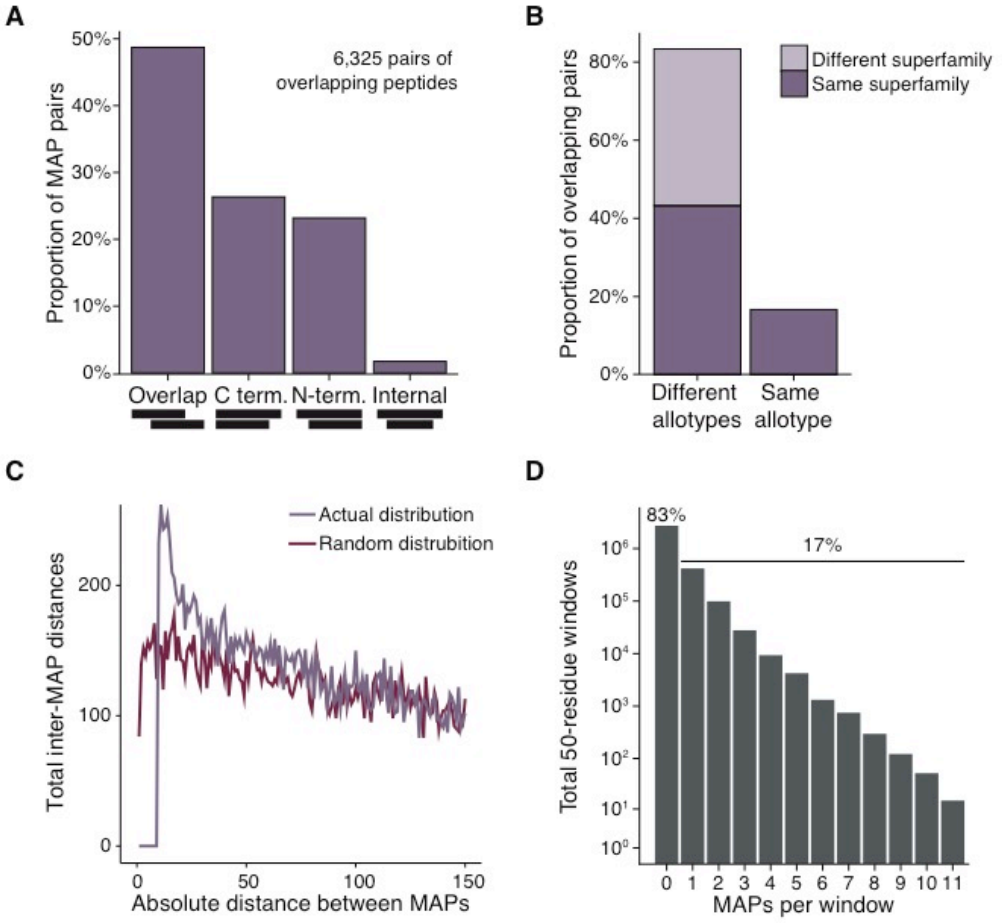

Figure 12. MAP distribution along source proteins. (A) Distribution of overlap types for 6,325 pairs of overlapping MAPs formed by 8,228 individual peptides: pairs with any overlapping residues and no common ends; pairs with a common C-terminus; pairs with a common N-terminus; and pairs with one peptide 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, the superfamily origins are indicated (Sidney et al., 2008). (C) Spatial distribution of MAPs along proteins generating more than one MAP compared to a random distribution matched to the length of source proteins. Absolute distances were computed and shown for distances up to 150 amino acids, beyond this random and actual distributions were largely the same. MAPs within 8 amino acids of each other, the length of canonical binding motifs, were merged in the actual distribution. Overall distances between MAPs are significantly less in absolute and relative comparisons with the matched random distribution ($p = 6 \times 10^{-6}$ and $p = 4 \times 10^{-8}$ respectively). (D) Exome coverage by the immunopeptidome. A window of 150 base pairs (50 amino acids) was moved along the transcribed exome of B-LCLs. Histogram shows the distribution of MAP number per window.

3.3.3 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. Based on RNA-sequencing data, we defined the B-LCL transcriptome as 10,677 expressed (FPKM > 1) and annotated protein-coding genes; 6,231 genes were a source of MAPs while 4,446 were not (Figure 13A, details in Materials & Methods). We then applied a variety of analyses and prediction algorithms to study the features of MAP source 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 9mer MHCI binders was no greater in source genes than non-source genes (Figure 13B). Since the difference between MAP source and non-source 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 according to previous studies based on smaller datasets. According to some reports, MAP derive preferentially from highly abundant mRNAs or proteins (Granados et al., 2012; Hoof et al., 2012; Bassani-Sternberg et al., 2015), but other reports cast some doubts on this contention (Weinzierl et al., 2007; Mester et al., 2011). 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 13C). However expression alone provided an incomplete portrait of antigen presentation: some highly expressed genes generated no MAPs and, more startlingly, some lowly expressed genes were capable of generating MAPs. Since the transcriptome is an imperfect mirror of the proteome (Jovanovic et al., 2015; Liu and Aebersold, 2016), we also analyzed the relationship between protein abundance in human B cells (Kim et al., 2014a) and MAP generation. MAP source proteins are more abundant than non-source proteins (Figure 13D), yet the fact that some proteins with similar expression belonged to source or non-source groups suggested that other factors were at play.

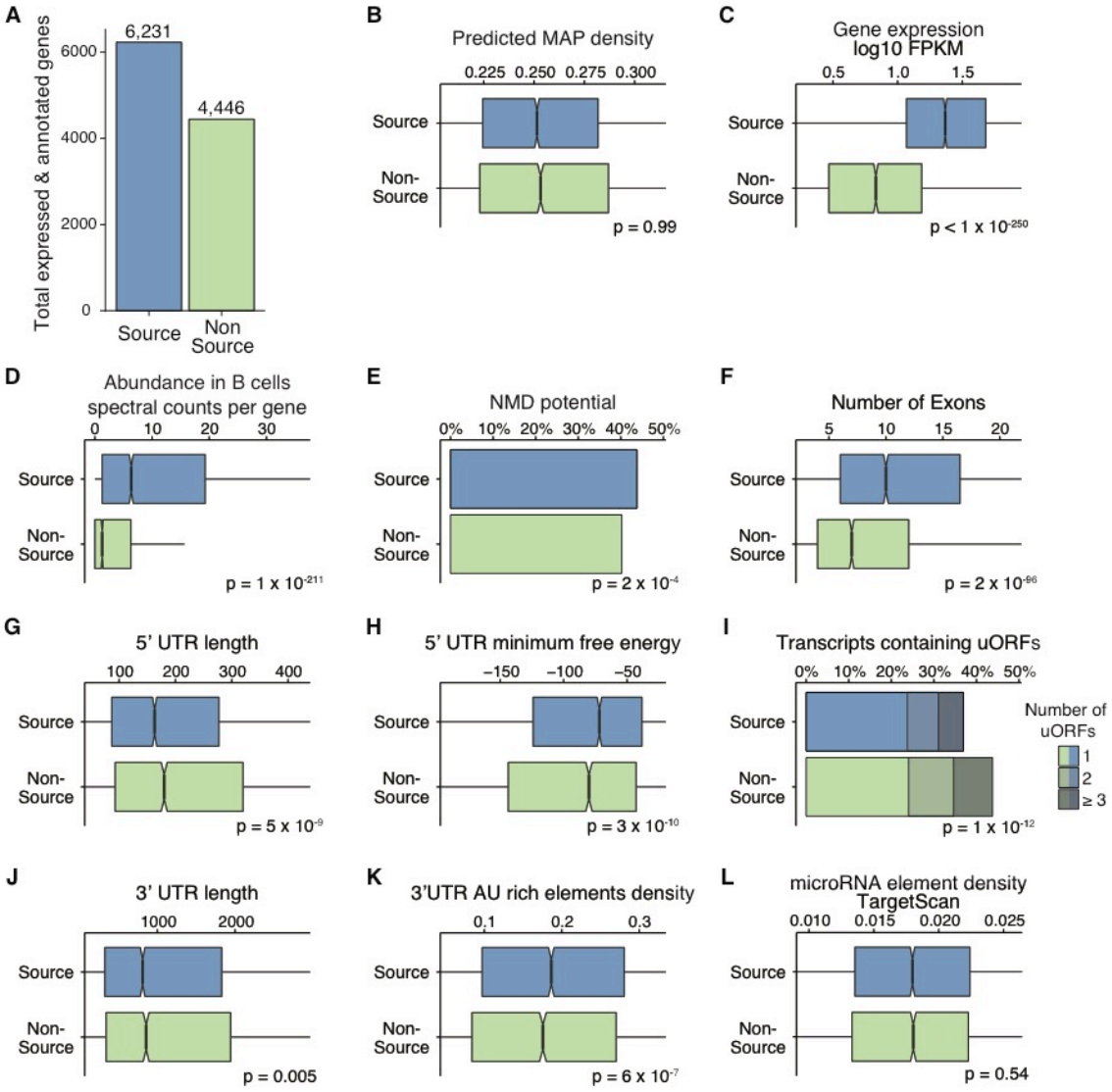

Figure 13. Features of MAP source genes and transcripts. (A) Total number of annotated source and non-source genes with mean expression >1 FPKM in B-LCLs. (B) Density of 9mer peptides predicted to bind any of the 27 HLA allotypes studied with an affinity ≤1250 nM. P value from a one-tailed Student's t-test. (C) Average expression of MAP source and non-source transcripts across 18 B-LCL cell lines. (D) Protein abundance in human B cells in spectral counts per gene; obtained from the Human Proteome Map (Kim et al., 2014a). (E) Proportion of genes with at least one transcript isoform undergoing nonsense-mediated decay according to Ensembl assembly 37. (F) Total number of exons per transcript. (G) Absolute length of the 5'UTR. (H) Minimum free energy of 5'UTR secondary structure predicted by RNAfold in the Vienna package (Lorenz et al., 2011). (I) Proportion of transcripts containing uORFs with

absolute counts indicated by shading. (J) Absolute length of the 3'UTR. (K) Density of AU rich elements along the 3'UTR. (L) Density of predicted microRNA target sites along the 3'UTR using TargetScan 7.0.

3.3.4 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 (Princiotta et al., 2003). 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 non-source transcripts. Coherent with the concept that nonsense mediated decay is a source of MAPs (Apcher et al., 2013), we observed that the proportion of genes with at least one transcript with an NMD biotype was higher in source relative to non-source genes (Figure 13E). Also, consistent with the positive correlation between the number of exons and translation efficiency, (Floor and Doudna, 2016), we found that MAPs derived from transcripts composed of more exons than non-source transcripts (Figure 13F), even when normalized for transcript length ($p = 2 \times 10^{-55}$).

We next examined features of the 5'UTR for evidence of translational regulation related to antigen processing. Upstream open reading frames (uORFs) tend to negatively influence translation by destabilizing transcripts and acting as a physical obstacle slowing ribosomal scanning (Calvo et al., 2009). The 5'UTRs of MAP source transcripts were significantly shorter and contained fewer uORFs (Figure 13G,I). In the same vein, the predicted secondary structure of the 5'UTR was less stable for MAP source transcripts (Figure 13H) although no definitive differences between the amount of pairing in this structure nor the GC content were found (Figure 19B,D).

The 3'UTR is a critical site of translational control containing regulatory elements such as AU rich elements and binding sites for microRNAs and RNA binding proteins (Szostak and Gebauer, 2013). We initially remarked that 3'UTRs were longer in non-source transcripts suggesting greater potential for regulation (Figure 13J). The density of AU rich elements was

greater in source 3'UTRs (Figure 13K) which may implicate transcripts in rapid decay or finer stability regulation (Schott and Stoecklin, 2010). Accordingly, slightly lower GC content was found in source 3'UTRs (Figure 19C). Stabilizing and destabilizing regulatory elements (Figure 19G,H) were queried in the 3'UTRs of all transcripts (Zhao et al., 2014) and revealed similar prevalence in source and non-source transcripts. Moreover, we were unable to confirm previous results that MAPs derive preferentially from transcripts with microRNA binding sites using two independent datasets (Granados et al., 2012) (Figure 13L and Figure 19E,F). However, our negative findings regarding binding sites for microRNAs and RNA binding proteins must be considered with some reservations. Firstly, because we used a more stringent p-value threshold of 0.001. Several differences would have been considered significant at a threshold of 0.05 (Figure 19). Secondly, microRNA regulation is highly cell-type specific while the methods used to predict microRNA involvement operate at an organism-wide level (Agarwal et al., 2015). Finally, since the effects of 3'UTR regulatory elements are heavily context-dependent (Szostak and Gebauer, 2013) the role of 3'UTR regulation in MAP generation in B-LCLs may be obscured by some lack of specificity.

Notably, features enriched in MAP source transcripts (Figure 13F-L and Figure19) had minimal correlations with protein abundance (absolute Spearman's ρ of 0.22 for number of exons and $\rho < 0.12$ for others, Figure 23). This led us to postulate that gene expression and transcript features may provide non-redundant information for the modeling of MAP generation.

3.3.5 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 (Hoof et al., 2012; Bassani-Sternberg et al., 2015) (Figure 14A). This may reflect that longer proteins, relative to shorter proteins, i) contain more appropriate MHCI binding sequences , ii) have a greater chance to form DRiPs, and iii) bind more ribosomes (Hoof et al., 2012; Floor and Doudna, 2016). MAP source proteins had lower hydropathy scores, indicating more polar

amino acid composition (Figure 14B). Furthermore, the predicted isoelectric point revealed greater acidic composition of source proteins (Figure 14C). At the next level of complexity, the predicted secondary structure of MAP source proteins showed distinct contribution of helix, turn and sheet motifs (Figure 14D-F). In particular, MAP source proteins showed a conspicuous enrichment in sheet motifs (Figure 14F).

The ubiquitin proteasome system is a key entry point for proteins into the MHCI processing pathway (Yewdell et al., 2003; de Verteuil et al., 2010). We first examined MAP proteins for proteasomal degradation motifs. We found that compared to non-source proteins, MAP source proteins contained higher frequencies of i) KEN-box and D-box motifs targeted by the anaphase promoting complex ubiquitin ligase (Liu et al., 2012) (Figure 14H,I), ii) PEST motifs which serve as proteolytic signals for the proteasome and other proteases (Rechsteiner and Rogers, 1996) (Figure 14H,I), and iii) canonical lysine ubiquitination sites (Chen et al., 2013) (Figure 14J).

Unstructured protein regions serve as initiation sites for proteasomal degradation (Prakash et al., 2004), and intrinsically disordered segments favour proteasome degradation (van der Lee et al., 2014). Therefore, to analyze the potential influence of protein disorder on MAP generation, we computed the disorder status of proteins in our dataset with the neural network predictor PONDR VLXT (Romero et al., 2001). Whether the average disorder of all residues, the proportion of disordered residues, the length of N-terminal disorder or the presence of internally disordered regions longer than 30 residues were considered, MAP source proteins consistently contained greater disorder compared to non-source proteins (Figure 14G). Similar results were obtained using two other disorder predictors: DISOPRED (Jones and Cozzetto, 2015) and IUPRED (Dosztanyi et al., 2005) (Figure 21B). We conclude that primary and secondary structure of proteins, and particularly those linked to proteasomal degradation, have a strong influence on MAP generation.

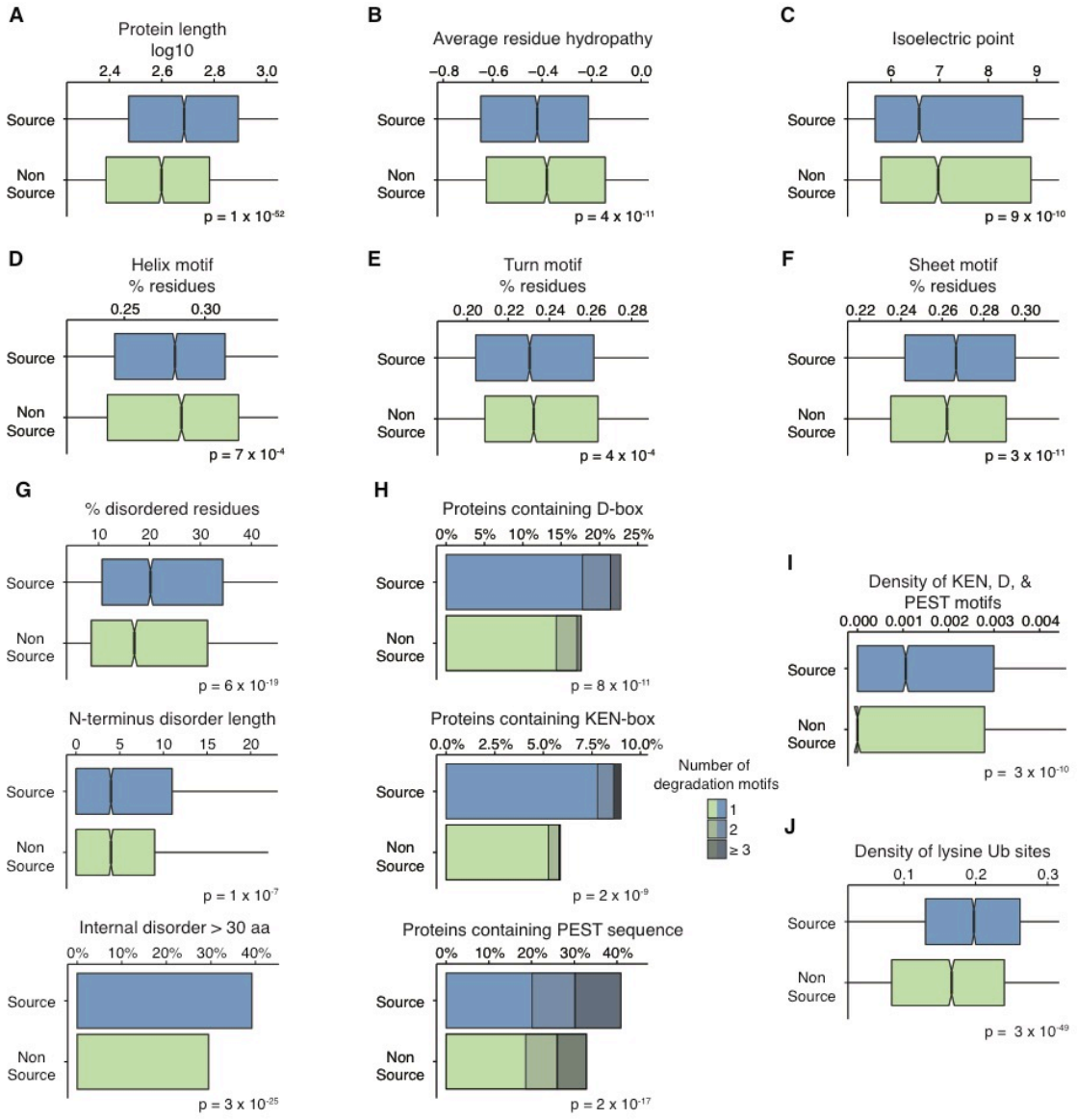

Figure 14. Features of MAP source proteins. (A) Protein length with a log10 transformation. (B-F) Metrics assessing amino acid content and secondary structure were predicted using ProtParam within BioPython (Gasteiger et al., 2005)¹⁰³¹⁰³. (G) Protein disorder was predicted using PONDR VXLT with a disorder cutoff of 0.7. (H) Proteasomal degradation motifs were predicted for all protein sequences using GPS-ARM and EMBOSS (Liu et al., 2012; Rice et al., 2000). (I) The total number of degradation motifs per protein normalized for protein length. (J) The proportion of lysine residues predicted by UbiProber to have a high probability of ubiquitination (Chen et al., 2013).

3.3.6 GO Terms analysis

We next compared the enrichment of gene ontology terms in MAP source and non-source genes using the topGO algorithm to eliminate redundancies (Alexa and Rahnenfuhrer, 2010). Our findings here confirm and extend reports based on smaller datasets (Hoof et al., 2012; Hickman et al., 2004; Granados et al., 2012). The source gene population was highly enriched in genes coding for intracellular proteins interacting with DNA, RNA and other proteins (Figure 15A). This may result from significantly higher expression of genes implicated in housekeeping functions such as poly(A) RNA binding, mitotic cell cycle, and mRNA processing. However, 16 of the top 100 GO terms enriched in source or non-source genes showed no difference in gene expression, suggesting GO annotation describes other factors in antigen processing. Non mutually exclusive hypotheses are that source 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 (Anton and Yewdell, 2014). Non source proteins were enriched in membrane components and related signalling processes, showing that proteins traversing the secretory pathway are poorly represented in the MHCI immunopeptidome (Figure 15B).

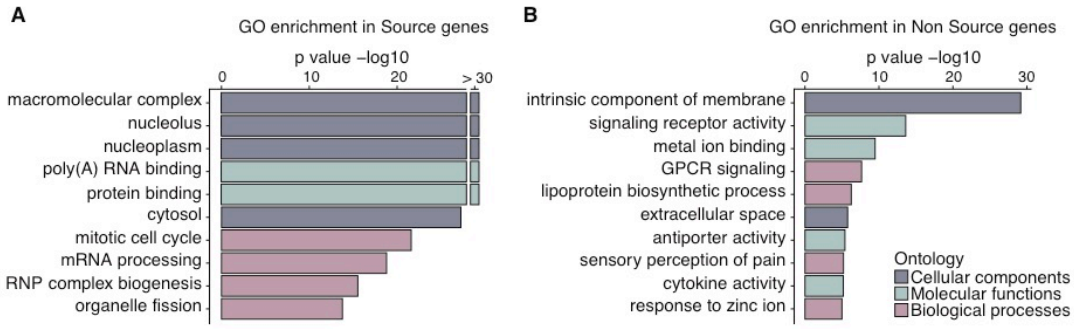

Figure 15. Gene ontology analysis of source and non-source genes. Enrichment in source (A) and non-source (B) groups was calculated on a background of both groups using the topGO algorithm to eliminate redundancies (Alexa and Rahnenfuhrer, 2010). The top 10 most enriched functions are shown for each group including all three ontology categories. RNP: ribonucleoprotein, GPCR: G-protein coupled receptor.

3.3.7 Modeling MAP generation

Having identified features that differentiate MAP source vs. non-source 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 Table II, we trained a logistic regression model on 80% of our dataset and tested its ability to discriminate source vs. non-source genes on the remaining 20% of our dataset. The process was repeated 1,000 times with randomly divided training and testing datasets. Prediction scores, falling between 0 and 1, demonstrated a considerable ability to correctly discriminate between MAP source and non-source genes (Figure 16A). 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 16B).

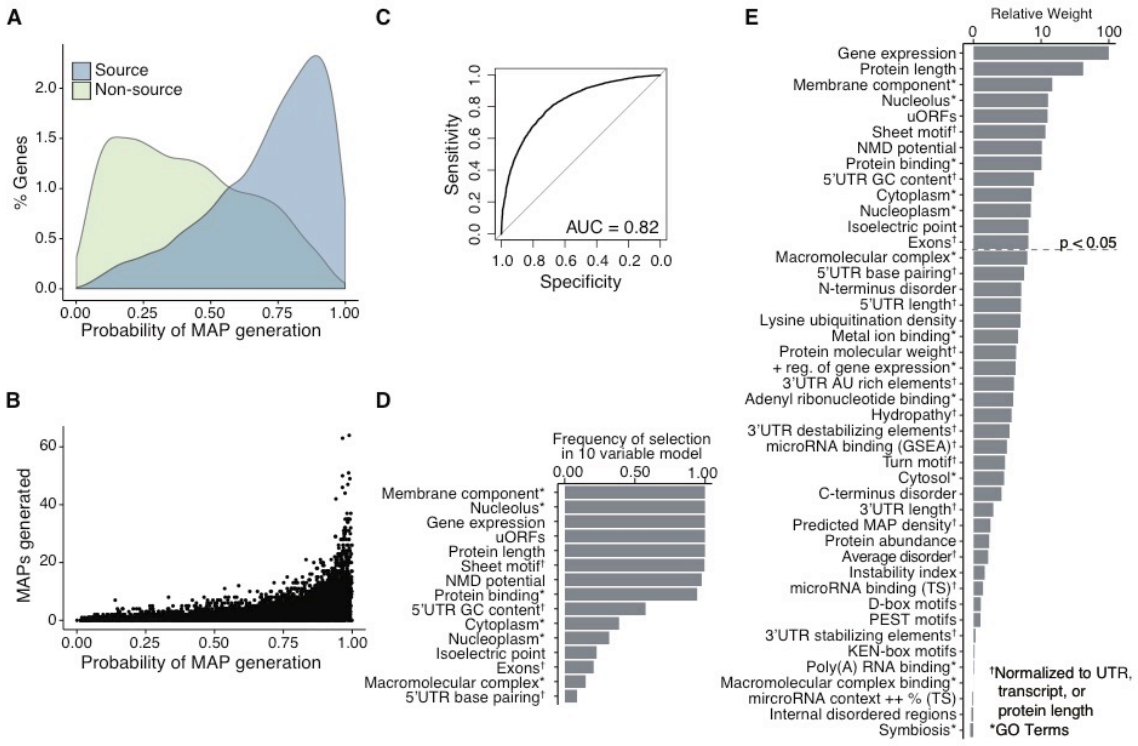

Figure 16. A logistic regression model to predict whether or not a gene will generate MAPs. (A) Prediction scores for each gene grouped by experimentally defined source classification. (B) Prediction scores for each gene and the number of MAPs generated. (C) Model performance measured by a ROC plot of sensitivity (the rate of true positives) as a function of specificity (the rate of true

negatives); the AUC is 0.82. (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 two class logistic regression model. Variables normalized by the length of the corresponding UTR, transcript or protein are denoted with †, GO terms denoted with *. TS: TargetScan, GSEA: Gene Set Enrichment Analysis database. All metrics are averaged over 1000 models (see Materials and Methods).

To assess the overall predictive power of the model, we constructed receiver operator characteristic plots (ROC) with averaged prediction scores and found an area under the curve (AUC) of 0.82 (Figure 16C). By examining the parameters of the model, we assessed the relative contribution of each feature to learning (Figure 16E). We found that gene expression was by far the most informative variable followed by protein length and protein abundance. Features of genes, transcripts and proteins were included in the group of relatively less important variables indicating 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 asses feature importance. We assessed the predictive capacity of a logistic regression model forced to select only the top 10 most informative features. Despite this constraint, the model achieved an average AUC of 0.81 (data not shown). The frequency with which features were selected in this model (Figure 16D) coincided with the relative weight when all input variables were considered (Figure 16E).

A two class distinction of MAP source and non-source genes does not to take into consideration that some source genes generate up to 64 non-redundant MAPs while other genes produce only one (Figure 11B). To integrate these findings we produced a nuanced version of the classification model that made predictions for three 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 0.87, while the low and none groups had AUCs of 0.65 and 0.82 respectively (Figure 22A). Clearly, the model had difficulty with the low group for which its predictions reached a maximum probability of 0.44 compared to 0.99 for the high and none categories (Figure 22C). Interestingly, when we compared the relative contribution of different input parameters between the two class and three class models we

found a very similar hierarchy (Figure 16E and Figure 22B). We conclude that no particular feature within the model distinguishes genes that generate few vs. numerous MAPs.

3.3.8 Model validation with independent datasets

The various strategies used for high-throughput MS analyses of the immunopeptidome present strengths and limitations (Caron et al., 2015b). 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 two other groups in the JY B-LCL cell line. MAPs in these two datasets were isolated by MHCI immunoprecipitation; one study used data-dependent MS (Bassani-Sternberg et al., 2015) and the other used data-independent MS (Caron et al., 2015a). While our dataset contained MAPs presented by 27 HLA-A,B allotypes, the two other datasets were limited to two HLA-A,B allotypes: HLA-A*02:01 and HLA-B*07:02. Notably, 86-87% of source genes for the two other datasets were included in our own dataset (Figure 17A). We extracted prediction scores for genes classified as source in each dataset. The salient finding was that the predictions for MAP source genes were at least as good for the two independent datasets as for our own (Figure 17B). We conclude that our prediction model is robust and that its accuracy is not biased by the method used for MAP isolation or identification.

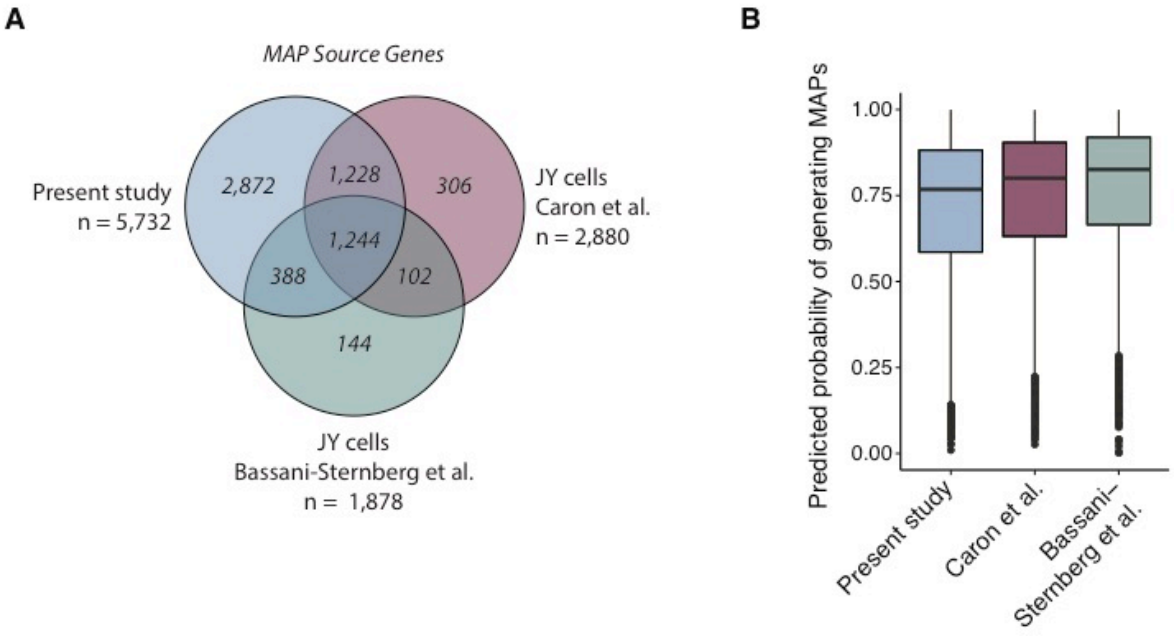

Figure 17. Evaluation of gene prediction scores with two independent datasets. (A) Overlap in source gene identifications between the present study and two independent studies of JY B-LCLs (Caron et al., 2015a; Bassani-Sternberg et al., 2015). 86-87% of MAP source genes identified in the two independent datasets were included in our MAP source genes. (B) Prediction scores for genes classified as source in each study were extracted from the two class logistic regression model.

3.4 Discussion

To the best of our knowledge, this study reports the largest dataset of MHCI-associated peptides to date. Several points can be made from our comprehensive analyses of 25,172 MAPs presented by 27 HLA-A,B allotypes which illustrate how there can be "strength in numbers" (Benoist et al., 2006). Indeed, while analyses of smaller datasets suggested that individual genes were represented in the immunopeptidome by only a single MAP (Hoof et al., 2012), we found that MAP source genes generated up to 64 non-redundant MAPs. Importantly, we found that MAPs presented by 27 MHCI allotypes altogether cover an unexpectedly small fraction of the protein-coding exome (10-17%) because

i) 42% of genes generate no MAPs, and ii) MAPs derive 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 (two to four HLA-A,B allotypes/individual) will originate from common genomic regions. Further studies are certainly warranted in order to explore whether, relative to the whole exome, MAP "hotspots" have distinctive features that would make their monitoring by T cells of special importance. For instance, are these hotspots preferential sites of somatic mutations in cancer cells or do they resemble viral genes?

Our report suggests that at the systems-level, MAP generation is regulated by numerous features of transcripts and proteins that 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 processing 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 (de Verteuil et al., 2010; Caron et al., 2011). 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 which collectively express almost all protein coding genes (Sansom et al., 2014; St-Pierre et al., 2015). 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 (Sansom et al., 2014; Brennecke et al., 2015). 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 models that predicts whether or not a given gene will produce MAPs with a ROC AUC of 0.82. The robustness of this model was validated by

testing on independent datasets. 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. Indeed, there are certain limitations to a rather coarse two 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 (Caron et al., 2015b). 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 (Caron et al., 2015a).

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 (Schumacher and Schreiber, 2015). However, discovery of cancer specific MAPs is currently fraught with major difficulties. Typically, neo-MAP discovery follows 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% (Robbins et al., 2013; Yadav et al., 2014; Blankenstein et al., 2015). Our contention is that this low success rate is simply due to the fact that few mutations are strategically located in MAP hotspots and that most mutations are in exomic sequences that are not covered by the immunopeptidome. We believe that progress in the field neo-MAP discovery would be greatly facilitated by large scale analyses of cancer cell immunopeptidomes.

3.5 Materials and Methods

3.5.1 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 mass spectrometry are outlined and described previously (Granados et al., 2014; Granados et al., 2016). 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 binding affinity less than 1,250 nM. When possible binding affinities were predicted with NetMHC 3.4 (21 allotypes), otherwise NetMHCcons 1.1 was applied (6 allotypes). Peptides were mapped to proteins in ENSEMBL assembly 37 using PyGeno (Daouda et al., 2016). 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 expressed transcript.

3.5.2 Simulations of the redundancy in MAP and MAP source gene repertoires

Allotypes were randomly ordered and either peptides or genes were considered. The number of non-redundant identifications was counted considering the repertoires of each subsequent allotype. The simulation was repeated 1000 times; average repertoire sizes are shown. The same simulation considering subjects instead of allotypes was also performed. We noted greater redundancy in this simulation due to some subjects sharing the same allotypes.

3.5.3 Spatial localization of MAPs along source proteins

Every pair of overlapping MAPs was extracted for each protein generating more than one 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 one amino acid 'Overlap'. Alleles presenting each peptide pair and their superfamilies were compared (Sidney et al., 2008). All distances between MAPs

on the same protein were computed for the actual distribution. MAPs start sites within 8 amino acids of each other were considered as one peptide with an averaged start site. For the random distribution, an equivalent number of MAPs were randomly placed within the same protein length and inter-MAP distances computed. For relative comparisons, distances were normalized for the length of each protein. To estimate exome coverage, a window of 150 base pairs was moved residue by residue along each of the 10,677 proteins expressed in our B-LCLs; the number of MAPs seen in each window was counted.

3.5.4 Evaluating features of transcripts and proteins

To ensure the quality and relevance of our source and non-source gene sets, we considered all genes expressed > 1 FPKM on average in all 18 B-LCLs using TopHat mapping to ENSEMBL human assembly 37. For each gene, the most expressed protein-generating transcript with complete HAVANA annotation and the corresponding protein were selected. For MAP source transcripts, the transcript had to generate at least one MAP. Feature assembly was executed in Python version 2.7.10, pyGeno was used to extract transcript and protein sequences (Daouda et al., 2016). Annotation translation was determined with the ENSEMBL BioMart extension (Zerbino et al., 2015).To calculate the predicted MAP density, NetMHC was used to predict the binding affinity of overlapping 9mers 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 ≥1250nM was calculated for each protein.

B cell protein abundance in average spectral counts per gene was extracted from the Human Proteome Map (Kim et al., 2014a). Genes with at least one transcript with an NMD biotype in ENSEMBL were considered to have NMD potential. uORFs were defined as non-overlapping 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 (Lorenz et al., 2011). The percentage of AU rich elements was defined as the fraction 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 and normalized for 3'UTR length (Zhao et al., 2014). TargetScan 7.0 was employed to predict microRNA binding sites within the 3'UTR (Agarwal et al., 2015). 3'UTRs were prepared by removing ORFs; the number of non-overlapping microRNA binding sites was computed for all families of microRNAs, the summed Context++ score and mean percentile of this score were extracted for each transcript. Gene Set Enrichment Analysis microRNA target motifs were downloaded 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 (Gasteiger et al., 2005).

3.5.5 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 (Liu et al., 2012). PEST motifs were predicted using the function *epestfind* within EMBOSS version 6.5.7 (Rice et al., 2000). Ubiquitination sites were predicted with UbiProber (Chen et al., 2013) with a stringency of 70%. Three disorder prediction softwares 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 a biophysical model based on local interaction energies (Dosztanyi et al., 2010). 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 (Figure 21A).

3.5.6 Data visualization

Boxplots were made in R version 3.1.3 using ggplot2 version 1.0.0 (Wickham, 2009). Notched boxplots show the median values of each population with boxes extending from the

first to the third quartile - the interquartile range (IQR). Whiskers extend from the lowest to highest values within 1.5 x IQR. Notches around the median show 1.58 x IQR / sqrt(number of samples), roughly a 95% confidence interval. The range of the upper axis was narrowed from minimum of the 15th to the maximum of the 85th percentile of either source or non-source populations. Outliers are not shown.

3.5.7 Gene ontology analysis

We compared either source or non-source genes on a background of both groups using the R package topGO (Alexa and Rahnenfuhrer, 2010). The Fisher weight algorithm was used to reduce redundancies and compute p-values.

3.5.8 Statistical analysis

Given that our data comparing source vs. non-source populations included roughly 10,000 genes, p-values were considered significant if they exceeded the threshold of 0.001. Unless otherwise noted, we employed two-sample Wilcoxon rank sum tests to compare continuous variables and Fisher's Exact Test to compare count data because of the robustness of these tests. All statistical analyses were performed in R version 3.2.2.

3.5.9 Logistic regression modeling

The variables listed in Table II were used as input variables for logistic regression models run with the R packages caret and MASS (Kuhn, 2016; Venables and Ripley, 2002). Genes without 5'UTRs were excluded bringing the total number of genes to 9,807. The top 50 most enriched GO terms from the source and non-source groups were included. To limit the extent of correlation in input variables which can obscure their relative weight, some variables were excluded. Input variables were also normalized by length of the appropriate UTR, transcript or protein. Spearman's rank correlation coefficient ρ was calculated for each pair of input parameters (Figure 23). Near-zero variance parameters were excluded, we noted that this excluded the majority of GO terms. The data was 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. Variable importance was computed based on the t statistic for all model parameters. An ordered logistic regression model with three class outcomes was built using the same protocol; categories were selected to optimize class balance (number of genes: 4,075 'none'; 2,957 'low'; 2,775 'high'). All metrics reported are averages of 1000 iterations of data division and model building. External datasets studying the immunopeptidome of the JY B-LCL cell line (Bassani-Sternberg et al., 2015; Caron et al., 2015a) were used to reclassify source genes; 16-20% of genes were excluded with the filters applied to define the expressed and annotated B-LCL gene set. Prediction scores for these groups of genes were extracted from the two class model using all features.

3.6 Supplementary figures & tables

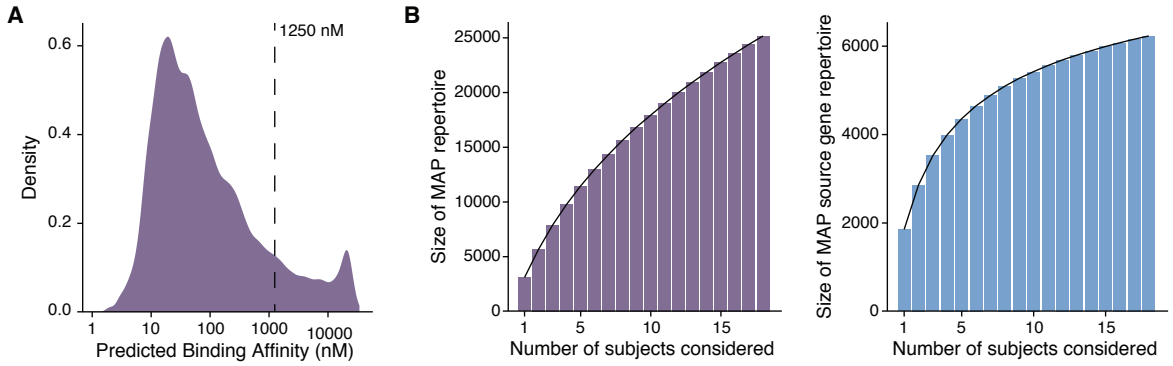

Figure 18. Supplementary characterization of MAP and MAP source gene repertoires. (A) The predicted binding affinities of identified MAPs prior to application of the <1250nM filter. (B) The number of unique identifications of MAPs (left panel) and MAP source genes (right panel) was counted as each additional randomly selected subject was considered. Results show the average of 1000 simulations. Note: common alleles between subjects increase the redundancy of peptides identified between subjects.

Table I. MAP identifications by subject and allele. Total number of MAPs discovered and assigned to each HLA-A,B allotype expressed in each subject. Peptides were assigned based on highest affinity with a global threshold of ≥1,250 nM.

	HLA-A		**HLA-B**	
B-LCL ID	*Number of MAPs identified*			
1	**03:01** *674*	**29:02** *370*	**08:01** *566*	**44:03** *1632*
2	**03:01** *584*	**29:02** *289*	**08:01** *619*	**44:03** *1655*
3	**02:01** *756*	**29:02** *411*	**44:03** *2007*	**57:01** *783*
4	**01:01** *561*	**02:01** *661*	**07:02** *2219*	**44:03** *1954*
5	**03:01** *203*	**11:01** *946*	**44:03** *1077*	**50:01** *331*
6	**02:01** *587*	**11:01** *1203*	**40:01** *1380*	**44:03** *1600*
7	**02:01** *144*	**23:01** *111*	**18:01** *355*	**44:03** *586*
8	**02:01** *767*	**03:01** *657*	**07:02** *3107*	
9	**02:01** *478*	**03:01** *639*	**07:02** *2615*	
10	**01:01** *544*	**02:01** *565*	**18:01** *1007*	**39:24** *260*
11	**02:01** *913*	**24:02** *719*	**15:01** *1916*	**73:01** *179*
12	**02:01** *1160*		**13:02** *39*	**41:01** *434*
13	**02:01** *1033*	**11:01** *1121*	**27:05** *1016*	**56:01** *712*
14	**03:01** *355*	**32:01** *308*	**27:05** *356*	**45:01** *500*
15	**01:01** *327*	**32:01** *491*	**08:01** *908*	
16	**11:01** *1087*		**14:02** *260*	**44:02** *859*
17	**11:01**		**18:03**	**35:01**

	1426		*215*	*450*
18	**03:01** *479*	**24:02** *334*	**07:02** *918*	**27:05** *669*

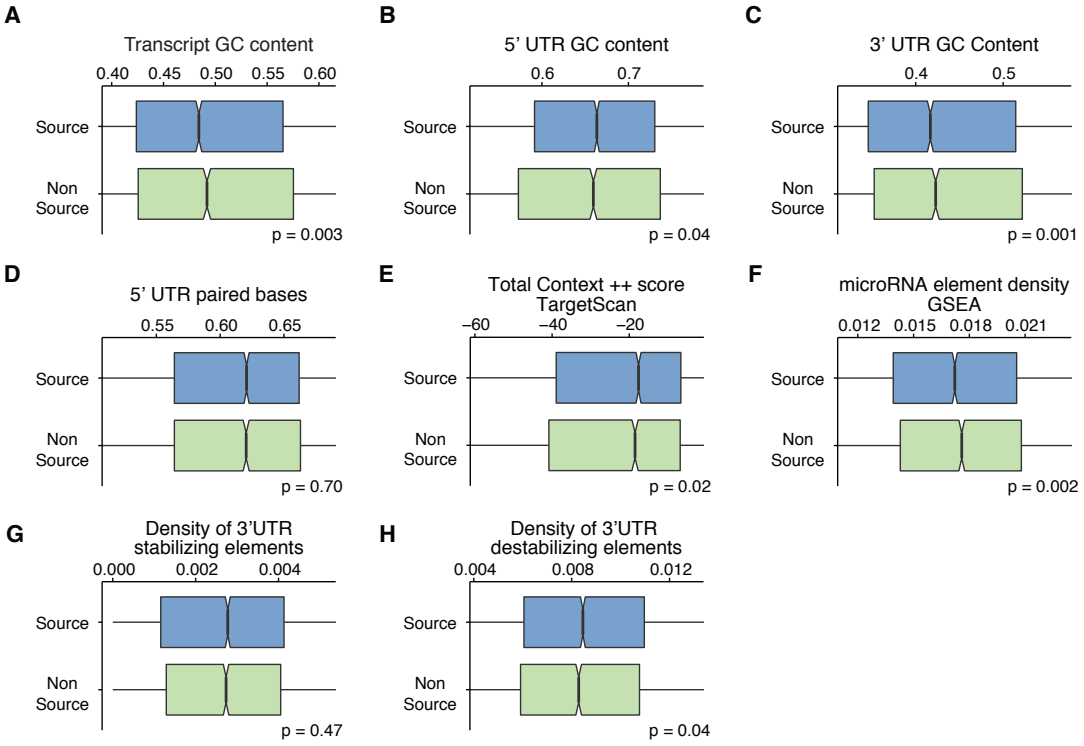

Figure 19. Supplementary features of MAP source transcripts. (A-C) Proportions of GC bases along transcripts, 5'UTRs and 3'UTR. (D) Proportion of bases paired in the most stable predicted secondary structure of the 5'UTR predicted by RNAfold in the Vienna package (Lorenz et al., 2011). (E) Cumulative gene context ++ score for microRNA binding sites in the 3'UTR predicted by TargetScan 7.0 (Agarwal et al., 2015). (F) Density of microRNA binding sites along the 3'UTR predicted by Gene Set Enrichment Analysis (GSEA) database (Subramanian et al., 2005). (G-H) Density of stabilizing and destabilizing elements in the 3'UTR (Zhao et al., 2014).

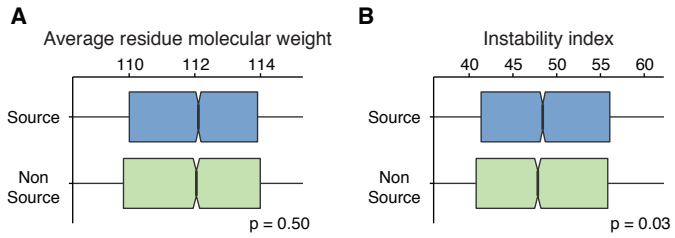

Figure 20. Supplementary features of MAP source proteins. (A) Average molecular weight of amino acid residues. (B) Predicted instability index using ProtParam within BioPython (Gasteiger et al., 2005).

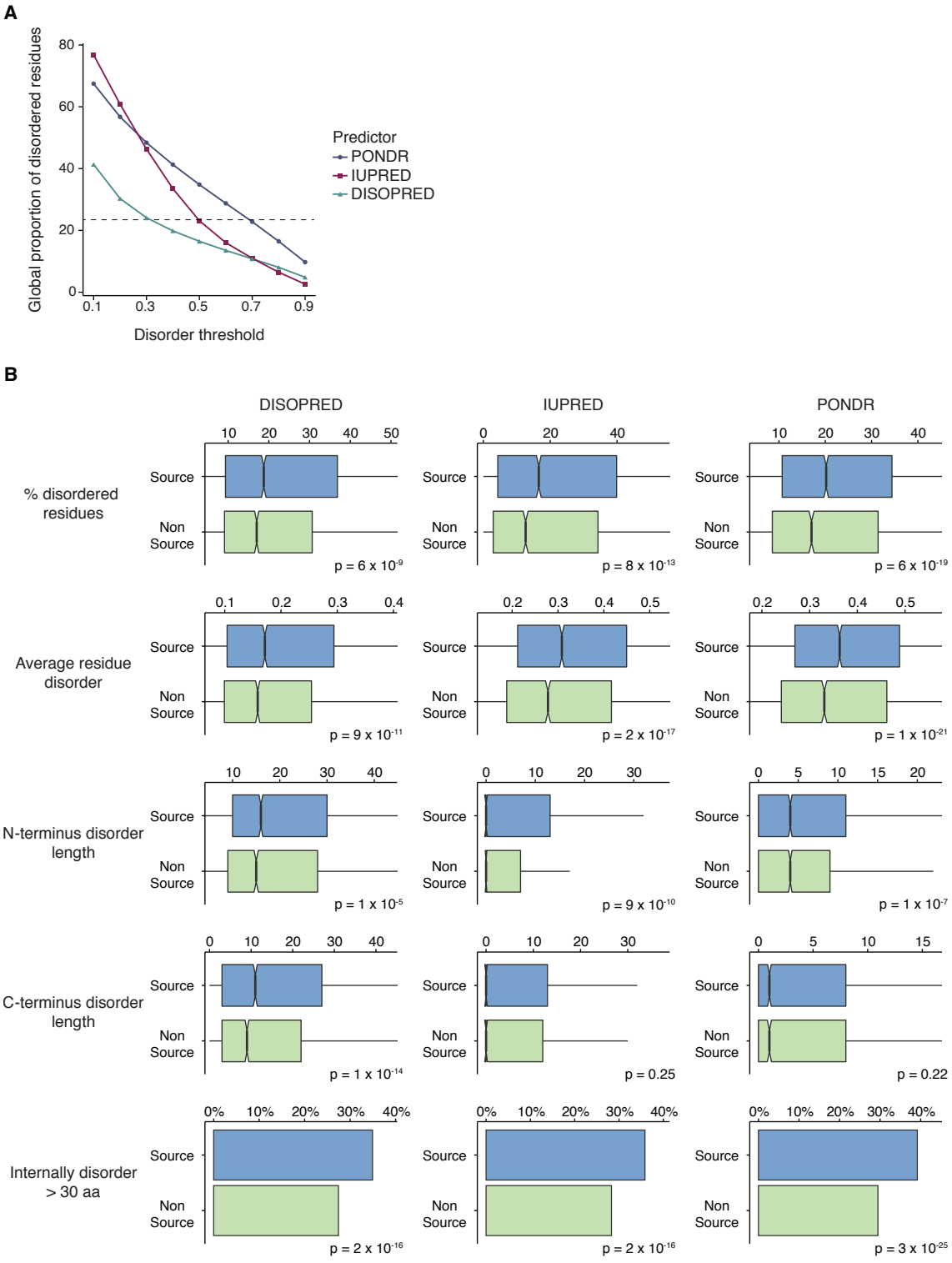

Figure 21. Protein disorder predicted by three complementary methods: PONDR VL-XT, DISOPRED and IUPRED. (A) Global proportion of disordered residues as a function of cutoff value above which a residue is

'disordered' for each predictor. Cutoff values for each predictor were chosen to roughly equate proteome wide disorder: 0.3, 0.5 and 0.7 for DISOPRED, IUPRED and PONDR VL-XT, respectively. (B) Prediction of 5 metrics of disorder for each predictor: the proportion of disordered residues, average residue disorder, the length of N and C terminus disorder and occurrence of internally disordered regions longer than 30 amino acids (Romero et al., 2001; Jones and Cozzetto, 2015; Dosztanyi et al., 2005).

Table II. Features used for predictive modeling of MAP source vs. non-source genes.

	Description	**Tool**	**Citation**
Gene Features			
Gene Expression	Frequency per kilobase of transcript per million mapped reads (FPKM)	TopHat	Trapnell et al., 2009
NMD Potential	≥1 'NMD' transcript biotype	Ensembl	Zerbino et al. 2015
Transcript Features			
Number of exons	Normalized for transcript length	PyGeno	Daouda et al., 2016
5'UTR Length	Proportion of transcript length	PyGeno	Daouda et al., 2016
5'UTR GC Content	% composition of 5'UTR	PyGeno	Daouda et al., 2016
uORF	Total canonical reading frames in 5'UTR	PyGeno	Daouda et al., 2016
5'UTR base pairing	% paired RNA residues	RNAfold	Lorenz et al., 2011
3'UTR Length	proportion of transcript length	PyGeno	Daouda et al., 2016
3'UTR destabilizing motifs	Motif density along 3'UTR	-	Zhao et al., 2014
3'UTR stabilizing motifs	Motif density along 3'UTR	-	Zhao et al., 2014
3'UTR AU rich elements	Proportion of 3'UTR A/U sequences ≥ 5 nucleotides	PyGeno	Daouda et al,. 2016
miR elements (GSEA)	Total non-overlapping binding sites	GSEA miR motifs	Subramanian et al., 2005
miR elements (TS)	Total non-overlapping binding sites	TargetScan	Agarwal et al., 2015
miR context ++ precentile (TS)	Context++ score percentile	TargetScan	Agarwal et al., 2015
Protein Features			
Protein Length	-	PyGeno	Daouda et al., 2016
B Cell Abundance	Spectral counts per gene per experiment	Human Proteome Map	Kim et al., 2014
Predicted MAP density	Proportion of 9mers binding any of 27 HLA-A,B allotypes >1,250 nM	NetMHC; NetMHCcons	Karosiene et al., 2012; Lundegaard et al., 2008

Residue MW	Average residue Molecular weight	PyGeno	Daouda et al., 2016
Residue Hydropathy	Average residue GRAVY Index	ProtParam	Gasteiger et al., 2005
Isoelectric Point	-	ProtParam	Gasteiger et al., 2005
Instability Index	-	ProtParam	Gasteiger et al., 2005
Helix	Proportion of residues in predicted motif	ProtParam	Gasteiger et al., 2005
Turn	Proportion of residues in predicted motif	ProtParam	Gasteiger et al., 2005
Sheet	Proportion of residues in predicted motif	ProtParam	Gasteiger et al., 2005
N-terminus disorder	Length of disordered residues at N-terminus	PONDR VLXT	Romero et al., 2001
C-terminus disorder	Length of disordered residues at C-terminus	PONDR VLXT	Romero et al., 2001
Internal Disorder	Total disordered regions >30 residues	PONDR VLXT	Romero et al., 2001
Average Disorder	Average residue disorder prediction	PONDR VLXT	Romero et al., 2001
PEST motifs	Total motifs	EMBOSS	Rice et al., 2000
KEN Box	Total motifs	GPS-ARM	Liu et al., 2012
D Box	Total motifs	GPS-ARM	Liu et al., 2012
All	Density of KEN, D, PEST motifs	GPS-ARM + EMBOSS	Liu et al., 2012; Rice et al., 2000
Lysine ubiquitination sites	Proportion of Lysine residues predicted to be ubiquitinated > 0.7	UbiProber	Chen et al., 2013
Gene Ontology			
GO:0005654 Nucleoplasm		topGO	Alexa and Rahnenfuhrer, 2010
GO:0032991 Macromolecular complex		topGO	Alexa and Rahnenfuhrer, 2010
GO:0005730 Nucleolus		topGO	Alexa and Rahnenfuhrer, 2010
GO:0044822 Poly(A) RNA binding		topGO	Alexa and Rahnenfuhrer, 2010
GO:0005515 Protein binding		topGO	Alexa and Rahnenfuhrer, 2010
GO:0005829 Cytosol		topGO	Alexa and Rahnenfuhrer, 2010
GO:0044877 Macromolecular complex binding		topGO	Alexa and Rahnenfuhrer, 2010
GO.0044403 Symbiosis		topGO	Alexa and Rahnenfuhrer, 2010
GO:0010628 Positive regulation of gene expression		topGO	Alexa and Rahnenfuhrer, 2010
GO:0032559 Adenyl ribonucleotide binding		topGO	Alexa and Rahnenfuhrer, 2010

GO:0005737 Cytoplasm	topGO	Alexa and Rahnenfuhrer, 2010
GO:0031224 Intrinsic component of membrane	topGO	Alexa and Rahnenfuhrer, 2010
GO:0046872 Metal ion binding	topGO	Alexa and Rahnenfuhrer, 2010

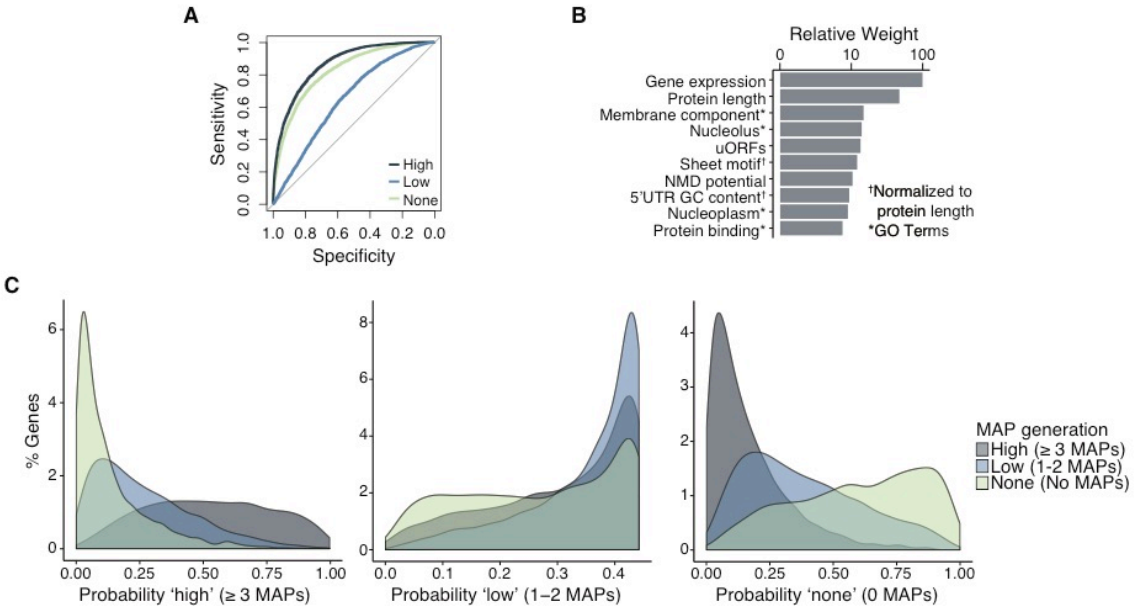

Figure 22. An ordered logistic regression model predicts whether MAP output for a gene will be high, low or nonexistent. (A) Model performance measured by a ROC plot; the AUCs are 0.87 for the high category, 0.65 for the low category and 0.82 for the none category. (B) The relative weight of the top 10 features contributing to prediction scores. Variables normalized by the length of the corresponding UTR, transcript or protein are denoted with †, GO terms denoted with *. (C) Prediction scores for each category grouped by experimentally defined source classification.

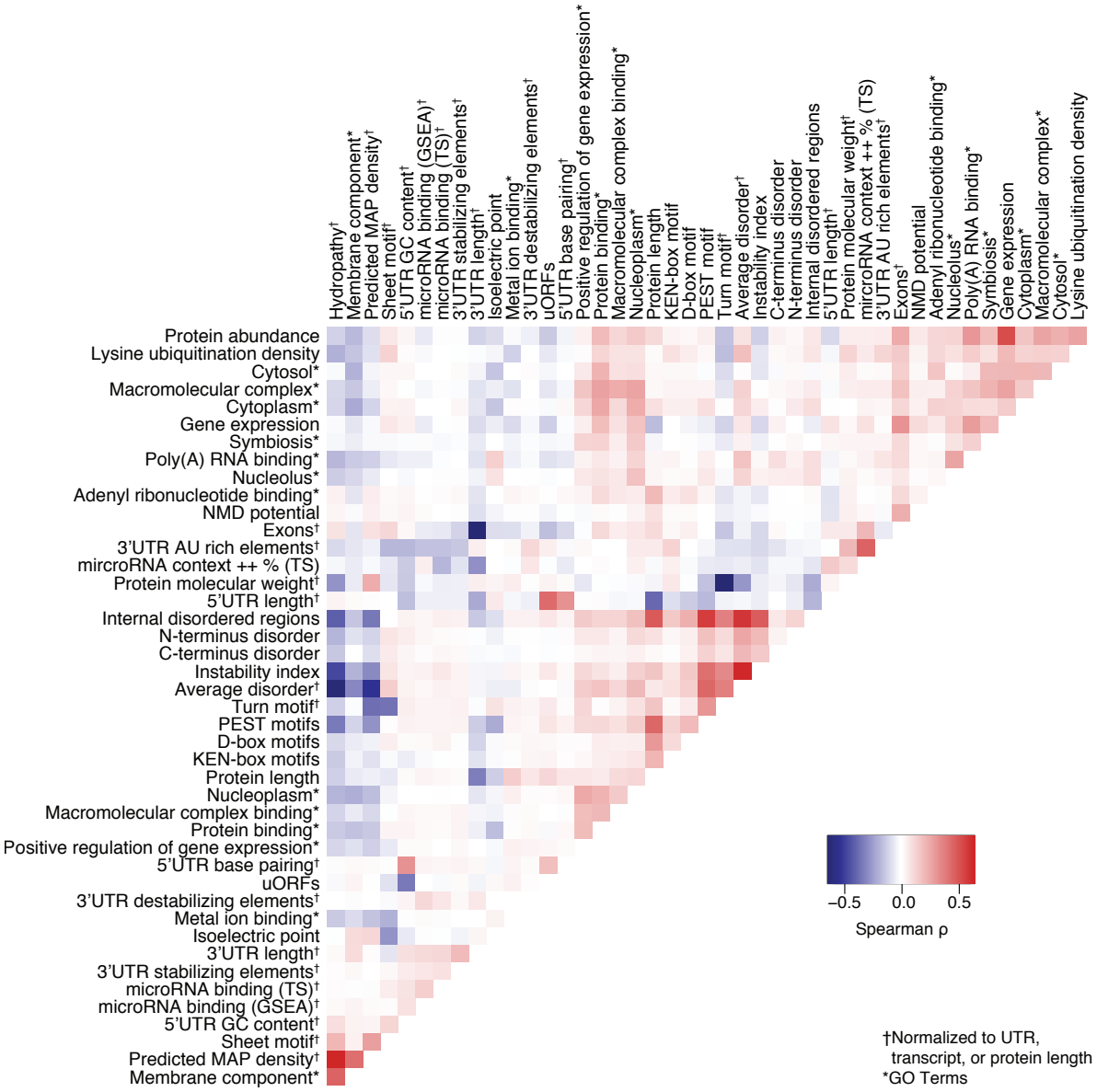

Figure 23. Correlation matrix of all model input variables using Spearman's ρ. Variables normalized by the length of the corresponding UTR, transcript or protein are denoted with †, GO terms denoted with *. TS: TargetScan, GSEA: Gene Set Enrichment Analysis database.

3.7 Acknowledgements

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 and the PT lab by the Genome Canada Innovation Network.

3.8 Additional Information

Funding

FUNDER	GRANT REFERENCE NUMBER	AUTHOR
Quebec Breast Cancer Foundation	Strategic Grants for Breast Cancer	Claude Perreault Sylvie Mader
Canadian Cancer Society	Impact Grant 701564	Claude Perreault Pierre Thibault

Ethics

This study was approved by the Comité d'Éthique de la Recherche de l'Hôpital Maisonneuve-Rosemont (Permit Number CÉR 14095).

3.9 Author contributions

HP, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting and revising the article. DPG, TD, CD, EB, MC, AR, JPL, CC, Acquisition of data, Analysis and interpretation of data, Revising the article. SL, PT, CP, Conception and design, Analysis and interpretation of data, Revising the article. SM, Analysis and interpretation of data, Revising the article. CP, Conception and design, Analysis and interpretation of data, Drafting and revising the article.

3.10 References

Agarwal,V., Bell,G.W., Nam,J.W., and Bartel,D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife. *4*.

Alexa, A. and Rahnenfuhrer, J. topGO: Enrichment analysis for Gene Ontology. R package[2.22.0]. 2010. Bioconductor.

Ref Type: Computer Program

Anton,L.C. and Yewdell,J.W. (2014). Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. J Leukoc. Biol *95*, 551-562.

Apcher,S., Millot,G., Daskalogianni,C., Scherl,A., Manoury,B., and Fahraeus,R. (2013). Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. Proc. Natl. Acad. Sci. U. S. A *110*, 17951-17956.

Bassani-Sternberg,M., Pletscher-Frankild,S., Jensen,L.J., and Mann,M. (2015). Mass spectrometry of HLA-I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol. Cell Proteomics. *14*, 1042-1052.

Benoist,C., Germain,R.N., and Mathis,D. (2006). A plaidoyer for 'systems immunology'. Immunol. Rev. *210*, 229-234.

Blankenstein,T., Leisegang,M., Uckert,W., and Schreiber,H. (2015). Targeting cancer-specific mutations by T cell receptor gene therapy. Curr. Opin. Immunol. *33*, 112-119.

Blum,J.S., Wearsch,P.A., and Cresswell,P. (2013). Pathways of antigen processing. Annu. Rev Immunol *31*, 443-473.

Brennecke,P., Reyes,A., Pinto,S., Rattay,K., Nguyen,M., Kuchler,R., Huber,W., Kyewski,B., and Steinmetz,L.M. (2015). Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. Nat Immunol *16*, 933-941.

Butler,T.C., Kardar,M., and Chakraborty,A.K. (2013). Quorum sensing allows T cells to discriminate between self and nonself. Proc. Natl. Acad. Sci U. S. A *110*, 11833-11838.

Calvo,S.E., Pagliarini,D.J., and Mootha,V.K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc. Natl. Acad. Sci. U. S. A *106*, 7507-7512.

Caron,E., Espona,L., Kowalewski,D.J., Schuster,H., Ternette,N., Alpizar,A., Schittenhelm,R.B., Ramarathinam,S.H., Lindestam Arlehamn,C.S., Chiek,K.C., Gillet,L.C., Rabsteyn,A., Navarro,P., Kim,S., Lam,H., Sturm,T., Marcilla,M., Sette,A., Campbell,D.S., Deutsch,E.W., Moritz,R.L., Purcell,A.W., Rammensee,H.G., Stevanovic,S., and Aebersold,R. (2015a). An open-source computational and data resource to analyze digital maps of immunopeptidomes. Elife. *4*.

Caron,E., Kowalewski,D.J., Koh,C.C., Sturm,T., Schuster,H., and Aebersold,R. (2015b). Analysis of MHC immunopeptidomes using mass spectrometry. Mol. Cell Proteomics. *14*, 3105-3117.

Caron,E., Vincent,K., Fortier,M.H., Laverdure,J.P., Bramoullé,A., Hardy,M.P., Voisin,G., Roux,P., Lemieux,S., Thibault,P., and Perreault,C. (2011). The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. Mol. Syst. Biol. *7*, 533.

Chakraborty,A.K. and Weiss,A. (2014). Insights into the initiation of TCR signaling. Nat Immunol *15*, 798-807.

Chen,X., Qiu,J.D., Shi,S.P., Suo,S.B., Huang,S.Y., and Liang,R.P. (2013). Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites. Bioinformatics. *29*, 1614-1622.

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

de Verteuil,D., Granados,D.P., Thibault,P., and Perreault,C. (2012). Origin and plasticity of MHC I-associated self peptides. Autoimmun. Rev. *11*, 627-635.

de Verteuil,D., Muratore-Schroeder,T.L., Granados,D.P., Fortier,M.H., Hardy,M.P., Bramoullé,A., Caron,E., Vincent,K., Mader,S., Lemieux,S., Thibault,P., and Perreault,C. (2010). 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.

Dosztanyi,Z., Csizmok,V., Tompa,P., and Simon,I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. *21*, 3433-3434.

Dosztanyi,Z., Meszaros,B., and Simon,I. (2010). Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief. Bioinform. *11*, 225-243.

Eisenlohr,L.C., Huang,L., and Golovina,T.N. (2007). Rethinking peptide supply to MHC class I molecules. Nat. Rev. Immunol. *7*, 403-410.

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

Floor,S.N. and Doudna,J.A. (2016). Tunable protein synthesis by transcript isoforms in human cells. Elife. *5*.

Gasteiger,E., Hoogland,C., Gattiker,A., Duvaud,S., Wilkins,M.R., Appel,R.D., and Bairoch,A. (2005). Protein identification and analysis tools on the ExPASy server. In The Proteomics Protocols Handbook, J.M.Walker, ed. Humana Press Inc.), pp. 571-607.

Goodenough,E., Robinson,T.M., Zook,M.B., Flanigan,K.M., Atkins,J.F., Howard,M.T., and Eisenlohr,L.C. (2014). 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.

Govern,C.C., Paczosa,M.K., Chakraborty,A.K., and Huseby,E.S. (2010). Fast on-rates allow short dwell time ligands to activate T cells. Proc. Natl. Acad. Sci. U. S. A *107*, 8724-8729.

Granados,D.P., Durette,C., Pearson,H., Bonneil,E., Laverdure,J.P., Côté,C., Carli,C., Delisle,J.S., Lemieux,S., Thibault,P., and Perreault,C. (2016). Proteogenomic-based discovery of human minor histocompatibility antigens with suitable features for immunotherapy of hematologic cancers. Leukemia *PMID: 26857467*, epub Feb 9, 2016.

Granados,D.P., Laumont,C.M., Thibault,P., and Perreault,C. (2015). The nature of self for T cells - a systems-level perspective. Curr. Opin. Immunol. *34*, 1-8.

Granados,D.P., Sriranganadane,D., Daouda,T., Zieger,A., Laumont,C.M., Caron-Lizotte,O., Boucher,G., Hardy,M.P., Gendron,P., Côté,C., Lemieux,S., Thibault,P., and Perreault,C. (2014). Impact of genomic polymorphism on the repertoire of human MHC class I-associated peptides. Nat Commun *5*, 3600.

Granados,D.P., Yahyaoui,W., Laumont,C.M., Daouda,T., Muratore-Schroeder,T.L., Cote,C., Laverdure,J.P., Lemieux,S., Thibault,P., and Perreault,C. (2012). MHC I-associated peptides preferentially derive from transcripts bearing miRNA response elements. Blood *119*, e181-e191.

Hammer,G.E., Kanaseki,T., and Shastri,N. (2007). The final touches make perfect the peptide-MHC class I repertoire. Immunity. *26*, 397-406.

Hassan,C., Kester,M.G., de Ru,A.H., Hombrink,P., Drijfhout,J.W., Nijveen,H., Leunissen,J.A., Heemskerk,M.H., Falkenburg,J.H., and van Veelen,P.A. (2013). The human leukocyte antigen-presented ligandome of B lymphocytes. Mol. Cell Proteomics. *12*, 1829-1843.

Hickman,H.D., Luis,A.D., Buchli,R., Few,S.R., Sathiamurthy,M., VanGundy,R.S., Giberson,C.F., and Hildebrand,W.H. (2004). Toward a definition of self: proteomic evaluation of the class I peptide repertoire. J. Immunol. *172*, 2944-2952.

Hoof,I., van Baarle,D., Hildebrand,W.H., and Kesmir,C. (2012). Proteome sampling by the HLA class I antigen processing pathway. PLoS. Comput. Biol. *8*, e1002517.

Jones,D.T. and Cozzetto,D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. *31*, 857-863.

Jovanovic,M., Rooney,M.S., Mertins,P., Przybylski,D., Chevrier,N., Satija,R., Rodriguez,E.H., Fields,A.P., Schwartz,S., Raychowdhury,R., Mumbach,M.R., Eisenhaure,T., Rabani,M., Gennert,D., Lu,D., Delorey,T., Weissman,J.S., Carr,S.A., Hacohen,N., and Regev,A. (2015). Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. Science *347*, 1259038.

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

Kim,M.S., Pinto,S.M., Getnet,D., Nirujogi,R.S., Manda,S.S., Chaerkady,R., Madugundu,A.K., Kelkar,D.S., Isserlin,R., Jain,S., Thomas,J.K., Muthusamy,B., Leal-Rojas,P., Kumar,P., Sahasrabuddhe,N.A., Balakrishnan,L., Advani,J., George,B., Renuse,S., Selvan,L.D., Patil,A.H., Nanjappa,V., Radhakrishnan,A., Prasad,S., Subbannayya,T., Raju,R., Kumar,M., Sreenivasamurthy,S.K., Marimuthu,A., Sathe,G.J., Chavan,S., Datta,K.K., Subbannayya,Y., Sahu,A., Yelamanchi,S.D., Jayaram,S., Rajagopalan,P., Sharma,J., Murthy,K.R., Syed,N., Goel,R., Khan,A.A., Ahmad,S., Dey,G., Mudgal,K., Chatterjee,A., Huang,T.C., Zhong,J., Wu,X., Shaw,P.G., Freed,D., Zahari,M.S., Mukherjee,K.K., Shankar,S., Mahadevan,A., Lam,H., Mitchell,C.J., Shankar,S.K., Satishchandra,P., Schroeder,J.T., Sirdeshmukh,R., Maitra,A., Leach,S.D., Drake,C.G., Halushka,M.K., Prasad,T.S.,

Hruban,R.H., Kerr,C.L., Bader,G.D., Iacobuzio-Donahue,C.A., Gowda,H., and Pandey,A. (2014a). A draft map of the human proteome. Nature *509*, 575-581.

Kim,Y., Sidney,J., Buus,S., Sette,A., Nielsen,M., and Peters,B. (2014b). Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. BMC. Bioinformatics. *15*, 241.

Kuhn, M. caret: Classification and Regression Training. [6.0-64]. 2016.
Ref Type: Computer Program

Laumont,C.M., Daouda,T., Laverdure,J.P., Bonneil,E., Caron-Lizotte,O., Hardy,M.P., Granados,D.P., Durette,C., Lemieux,S., Thibault,P., and Perreault,C. (2016). Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. Nat. Commun. *7*, 10238.

Liu,Y. and Aebersold,R. (2016). The interdependence of transcript and protein abundance: new data-new complexities. Mol. Syst. Biol. *12*, 856.

Liu,Z., Yuan,F., Ren,J., Cao,J., Zhou,Y., Yang,Q., and Xue,Y. (2012). GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. PLoS. ONE. *7*, e34370.

Lorenz,R., Bernhart,S.H., Honer Zu,S.C., Tafer,H., Flamm,C., Stadler,P.F., and Hofacker,I.L. (2011). ViennaRNA Package 2.0. Algorithms. Mol. Biol. *6*, 26.

Lundegaard,C., Lamberth,K., Harndahl,M., Buus,S., Lund,O., and Nielsen,M. (2008). 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. *36*, W509-W512.

Mester,G., Hoffmann,V., and Stevanovic,S. (2011). Insights into MHC class I antigen processing gained from large-scale analysis of class I ligands. Cell Mol. Life Sci. *68*, 1521-1532.

Paul,S., Weiskopf,D., Angelo,M.A., Sidney,J., Peters,B., and Sette,A. (2013). HLA class I alleles are associated with peptide-binding repertoires of different size, affinity and immunogenicity. J. Immunol. *191*, 5831-5839.

Prakash,S., Tian,L., Ratliff,K.S., Lehotzky,R.E., and Matouschek,A. (2004). An unstructured initiation site is required for efficient proteasome-mediated degradation. Nat. Struct. Mol. Biol. *11*, 830-837.

Princiotta,M.F., Finzi,D., Qian,S.B., Gibbs,J., Schuchmann,S., Buttgereit,F., Bennink,J.R., and Yewdell,J.W. (2003). Quantitating protein synthesis, degradation, and endogenous antigen processing. Immunity *18*, 343-354.

Rammensee,H., Bachmann,J., Emmerich,N.P., Bachor,O.A., and Stevanovic,S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics *50*, 213-219.

Rechsteiner,M. and Rogers,S.W. (1996). PEST sequences and regulation by proteolysis. Trends Biochem. Sci. *21*, 267-271.

Rice,P., Longden,I., and Bleasby,A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276-277.

Robbins,P.F., Lu,Y.C., El-Gamil,M., Li,Y.F., Gross,C., Gartner,J., Lin,J.C., Teer,J.K., Cliften,P., Tycksen,E., Samuels,Y., and Rosenberg,S.A. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. Nat Med *19*, 747-752.

Rock,K.L., Farfan-Arribas,D.J., Colbert,J.D., and Goldberg,A.L. (2014). Re-examining class-I presentation and the DRiP hypothesis. Trends Immunol. *35*, 144-152.

Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J., and Dunker,A.K. (2001). Sequence complexity of disordered protein. Proteins *42*, 38-48.

Sansom,S.N., Shikama,N., Zhanybekova,S., Nusspaumer,G., Macaulay,I.C., Deadman,M.E., Heger,A., Ponting,C.P., and Hollander,G.A. (2014). Population and single cell genomics reveal the Aire-dependency, relief from Polycomb silencing and distribution of self-antigen expression in thymic epithelia. Genome Res. *24*, 1918-1931.

Schott,J. and Stoecklin,G. (2010). Networks controlling mRNA decay in the immune system. Wiley. Interdiscip. Rev RNA. *1*, 432-456.

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

Sidney,J., Peters,B., Frahm,N., Brander,C., and Sette,A. (2008). HLA class I supertypes: a revised and updated classification. BMC. Immunol. *9*, 1.

St-Pierre,C., Trofimov,A., Brochu,S., Lemieux,S., and Perreault,C. (2015). Differential features of AIRE-induced and AIRE-independent promiscuous gene expression in thymic epithelial cells. J Immunol *195*, 498-506.

Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., and Mesirov,J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A *102*, 15545-15550.

Szostak,E. and Gebauer,F. (2013). Translational control by 3'-UTR-binding proteins. Brief. Funct. Genomics *12*, 58-65.

van der Lee,R., Lang,B., Kruse,K., Gsponer,J., Sanchez de,G.N., Huynen,M.A., Matouschek,A., Fuxreiter,M., and Babu,M.M. (2014). Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell Rep. *8*, 1832-1844.

Venables,W.N. and Ripley,B.D. (2002). Modern applied statistics with S. Springer).

Vigneron,N. and Van den Eynde,B.J. (2012). Proteasome subtypes and the processing of tumor antigens: increasing antigenic diversity. Curr. Opin. Immunol. *24*, 84-91.

Vrisekoop,N., Monteiro,J.P., Mandl,J.N., and Germain,R.N. (2014). Revisiting thymic positive selection and the mature T cell repertoire for antigen. Immunity *41*, 181-190.

Weinzierl,A.O., Lemmel,C., Schoor,O., Muller,M., Kruger,T., Wernet,D., Hennenlotter,J., Stenzl,A., Klingel,K., Rammensee,H.G., and Stevanovic,S. (2007). Distorted relation between mRNA copy number and corresponding major histocompatibility complex ligand density on the cell surface. Mol. Cell. Proteomics *6*, 102-113.

Yadav,M., Jhunjhunwala,S., Phung,Q.T., Lupardus,P., Tanguay,J., Bumbaca,S., Franci,C., Cheung,T.K., Fritsche,J., Weinschenk,T., Modrusan,Z., Mellman,I., Lill,J.R., and Delamarre,L. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. Nature *515*, 572-576.

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

Zerbino,D.R., Wilder,S.P., Johnson,N., Juettemann,T., and Flicek,P.R. (2015). The ensembl regulatory build. Genome Biol. *16*, 56.

Zhao,W., Pollack,J.L., Blagev,D.P., Zaitlen,N., McManus,M.T., and Erle,D.J. (2014). Massively parallel functional annotation of 3' untranslated regions. Nat Biotechnol. *32*, 387-391.

Chapter 4 - Discussion & perspectives

4.1 Elucidating the dynamics of MHCI expression

We show integral differences in absolute expression and peptide binding properties of different HLA allotypes. The completion of this story is of fundamental importance and will inform future studies of antigen presentation. For example, we found that mild acid elution efficiency is likely allotype dependent; future immunopeptidome studies should take this into account. Quantitative analyses also allow one to estimate adjustments to immunopeptidome discovery protocols for cells lineages with fewer MHCI molecules expressed at the cell surface.

What is the global scope of variation in MHCI expression across allotypes? Is expression shaped by superfamily, tapasin and PLC interaction, inherent molecular stability or the availability of peptides with suitable motifs? We suspect each of these factors will contribute to expression dynamics. A model that considers each step in expression offers the most comprehensive approach to comparing HLA allotype expression.

To complete the story of differential HLA expression would required a broader comparison of allotypes with different properties and therefore additional HLA-A,B antibodies. Thus far, antibodies specific to HLA-A*01:01, HLA-A*23:01, HLA*24:01 and HLA-B*27:05 have been identified and titrated (details presented in appendix 1). A panel of 8 allotypes (these and the 4 studied in Chapter 2) would offer more generally applicable results. A limiting factor is quite simply the availability of appropriate high affinity antibodies with specificity for different HLA allotypes - one solution to would be custom recombination of available antibodies with appropriate isotypes (mouse IgG) for quantitation. Evidently, by conducting further replicates of each experiment one could draw more concrete conclusions about patterns of absolute MHCI expression and relative dynamics.

To obtain a comprehensive view of the entire life cycle of MHCI expression in B-LCLs, additional quantification of intracellular MHCI and analyses of surface stability should be carried out. A simple fixation protocol combined with the quantitative assay would measure intracellular allotype expression using flow cytometry. Preliminary results indicate

approximately 50% of MHCI are retained intracellularly. Retention may be linked to efficiency of peptide loading, quality filtering and/or inherent stability of peptide-allotype complexes. Complete intracellular and extracellular quantitation profiles would also describe the global abundance of each allotype.

Furthermore, the duration of extracellular expression could be assessed *in vitro* by inhibiting the transport through secretory pathway. Reagents such as monensin or Brefeldin A inhibit the secretory pathway and should block expression of newly formed pMHCI complexes.[104] By measuring the decline in pMHCI surface expression due to loss of peptides and instability without replacement by new pMHCI complexes, one could determine surface lifespan of various allotypes. We hypothesize surface expression time will be closely related to stability. Another approach could use a recently described surface density profiling assay preformed on MHCII to gauge the stability of each allotype.[98]

From these data it would be possible to build a complete model of the inter-allele dynamics from gene expression, absolute intracellular expression, MHCI recovery or expression rate, absolute surface expression, and surface stability or internalization rate. Mathematical modeling could highlight global and inter-allotype differences at each step of the MHCI life cycle. Differences in expression patterns would help elucidate if HLA allotypes have fundamentally different purposes, for example by drawing parallels to thymic selection,[88] or the generalist-specialist paradigm.[29] These findings will also be relevant to understanding the mechanisms linking HLA alleles to autoimmune disease and HIV control.

4.2 Developing immunopeptidome predictions

Our study of the genetic origins of the immunopeptidome revealed only 58% of genes produce MAPs for 27 allotypes in 18 individuals studied. MAP source genes and gene products showed distinct features contributing to MAP generation, for example features reflecting greater translational efficiency and preferential proteasomal degradation. Using these features, a logistic regression model was built and predicted with good accuracy whether or not

a give gene would generate MAPs. We confirmed the defining features of MAP source genes included gene expression, protein length and protein abundance with significant contributions from other GO annotation, transcript, and protein related features. Our model therefore represents a comprehensive look at antigen processing independent of HLA alleles and antigen binding. The next steps will be to test and apply the model in different settings.

The utility of any tool depends on its' availability. The impact of our findings will certainly be more important if others can easily make use of these results. To achieve this, there are three essential steps to complete. First, the model must be tested on other contexts using gene expression data. For example, studies of renal, breast, and lymphoid cell lines could serve to validate and adjust model parameters.[52,54] Second, we must optimize the model to use a few features as possible to speed up the processing time; this could be easily done since a model limited to 10 variables had comparable performance (average ROC AUC = 0.81 compared to 0.82 for the complete model, data not shown). Third, a user-friendly script to gather the appropriate features must be made available alongside the model.

The logistic regression model approach offers the advantages of simplicity, speed and allows one to asses the weight of each variable. The model had good performance (ROC AUC = 0.82) and accuracy (0.75). There are however limitations using a best-fit logistic function to stratify all samples, not the least of which is the extent of misclassification. While false positives are of less concern since our MAP identification is not exhaustive, false negatives are of interest for improving model performance. The most striking difference between the genes classified as false negatives and all positive source genes was lower gene expression. Secondary differences included fewer exons, shorter proteins, longer 3'UTRs and lower predicted lysine ubiquitination in the false negative population. Remaining features were comparable with the positive population. We conclude a limitation of our model is inflexibility with respect to dominant variables.

Alternatively, more complex statistical models such as artificial neural networks or support vector machines may show more flexibility with respect to single variables. Other options include rule-based algorithms such as random forest or decision tree models. Notably,

rule-based algorithms will be more comprehensible than statistical models. However, since we have generally understood the role of each feature in our study, application of a statistical model - which preform well with continuous variables - could augment preformance. Otherwise a combination of results from different models could be more powerful.[105,106]

We were unable to identify specific features that stratified highly from lowly MAP-producing source genes. However the diversity of unique MAP sequences produced is but one dimension of MAP production. Quantitative information in terms of the copy number of each MAP per cell is becoming available for large scale MS studies using DIA. This techniques draws from multiple reaction monitoring and DDA techniques using comprehensive peptide libraries for comparison with immunopeptidome samples. Incorporating diversity and copy number in future analyses will offer a clearer stratification between highly and lowly MAP producing genes.

While we are able to predict MAP source genes with good accuracy, the specific location of true MAPs along source proteins remains a challenging question. We have shown that some regions or 'hotspots' preferentially generate multiple MAPs. Further studies might identify characteristics of hotspot regions to predict start sites. Do MAPs derive preferentially from regions with particular motifs, disorder, or degradation sites? A two step prediction model defining i) potential source genes and ii) the most probable locations of MAPs would allow for more sensitive predictions. Each predicted 'hotspot' could then be tailored to the HLA alleles using binding affinity predictions.

4.3 Applications of immunopeptidome predictions

As a frontier treatment for cancer, implementation of immunotherapy faces many challenges not the least of which is the identification of suitable targets. Immunotherapy co-opts host or engineered T cells to identify and eliminate cancer cells. MAPs containing neoplastic mutations (neo-MAPs) are a particularly fruitful subset of therapeutic targets because they may already be targeted by host tumour infiltrating lymphocytes (TILs) and are

specific to tumour cells.[107] Perhaps the most striking finding of our work is that MAPs derive from a small portion of the genome (< 17%). In the context of immunotherapy, this would explain at least in part why many predicted neo-MAPs (< 90%) are not presented at the cell surface.[61,108,109]

We propose three approaches to applying immunopeptidome predictions to discover neo-MAPs. The first is to apply prediction algorithms within current pipelines to optimize peptide selection. Presently neo-MAP discovery relies on exome sequencing, identification of mutations, and selection of mutations located in peptide regions predicted to have a good MHC binding affinity. An addition step predicting the MAP generating potential of genes harbouring tumour mutations would prioritize potential neo-MAPs.

A second approach would predict MAP source genes by incorporating cancer-specific gene expression in the logistic regression model. RNA and exome sequencing of tumour samples from databases such as The Cancer Genome Atlas are excellent resources to study antigen presentation in cancer. HLA alleles can now be accurately predicted from sequencing data using OptiType.[110] By comparing predicted MAP source genes with mutations identified in each sample, a list of candidate neo-MAPs may be generated.[111–113] We also note that based on the short range effect, examining experimentally defined MAPs in similar tissues may facilitate identification of regions that preferentially generate MAPs within source genes.

Thirdly, one could study the MAP generating potential of frequently mutated oncogenes and tumour suppressor genes to identify theoretically common neo-MAPs.[114,115] Notably, mutated and non-functional proteins may undergo rapid proteasomal degradation and preferentially generate MAPs.[13]

The application of MAP source gene predictions in tumour sequencing data may allow one to bypass sample-hungry MS while being more selective than approaches relying solely on binding motifs. The feasibility of this approach remains to be tested but we believe that progress in the field neo-MAP discovery would be greatly facilitated by large scale analyses of cancer cell immunopeptidomes.

4.4 How diverse is the MAP repertoire ?

Our proteogenomic approach identified 25,172 MAPs binding 27 HLA-A,B allotypes deriving from 6,231 genes in B-LCLs. Although the use of multiple subjects bearing different alleles in this study allowed us to look at the global repertoire of MAP producing genes, we are acutely aware that out collection of MAPs is incomplete. MAPs may be lost at any step of the proteogenomic workflow: peptides remaining bound to MHCI during elution, lowly abundant peptides not detected in MS,[57] peptides with noncanonical binding interactions removed by prediction algorithms, or peptides presented uniquely under specific conditions. A central question remains: how diverse is the MAP repertoire?

Considering a single cell type in one individual, the diversity of MAPs is limited first by the HLA genotype and corresponding binding motifs. Results presented here and elsewhere show different HLA allotypes will present different diversity of peptides which may be inversely correlated with allele specific expression.[29,88,89] To begin to answer the question of MAP diversity, one might derive a diversity index for each allele:

Allotype diversity index = number of unique MAPs / allotype specific expression

Although the binding properties of each allotype are consistent across contexts, the diversity index may be shaped by competition from other allotypes. In principle, the net diversity of MAPs can be described quite simply:

Number of unique MAPs x copy number per cell = Total MHCI expression

One could estimate each of these parameters relatively easily using available techniques. Copy number can be loosely estimated in MS using known amounts of a few nonamers as Schellens et al. have done,[93] or more accurately with emerging DIA MS techniques.[57] Thus far, MAP copy numbers are known to range from 1 to 10^4 per cell.[93] Notably, the B-LCL model cell line exhibits particularly high MHCI expression, other groups have quantified HLA-A,B,C expression at ~100,000 molecules per cell on blood monocytes, roughly 10% of B-LCL expression.[89] The number of unique MAPs may be derived from studies like ours.

Together, these analyses would estimate MAP diversity and may shed light on how much of the immunopeptidome has been discovered so far.

How much of the exome do the entire set of MAPs cover? Our results show only a subset of genes can produce MAPs and certain regions of source proteins will preferentially produce MAPs in a largely allele-independent manner. We estimate the immunopeptidome covers <17% of the whole exome in B-LCLs. Importantly, from the perspective of a CD8+ T cell controlling an active infection, the immunopeptidome would still effectively capture highly expressed pathogen-derived proteins. Indeed, some mechanisms of viral immune evasion simply diminish gene expression.[75,116]

Globally, across cell types and HLA allotypes in a given individual the immunopeptidome will present a comprehensive overview of the self. To the extent that MAP processing depends on the cellular transcriptome, we expect differences in the immunopeptidome reflecting the dynamic intracellular and extracellular environments of each tissue. At the population level, considerable inter-individual variability in gene expression is well established and is known to shape immune response as well as disease sceptibility.[117,118,90] In reality, the tissue-specific immunopeptidomes of each allotype is likely similar between individuals but reflects inter-individual heterogeneity and is shaped by interacting dynamic systems.

Conclusion

MHCI is the centerpiece of adaptive immune surveillance and shapes the progression of numerous diseases. In this work we use cutting-edge proteomics, genomics, and informatics to answer two central questions in antigen presentation: what is the impact of allelic diversity on expression and peptide binding? What are the genetic origins of MAPs? Our results highlight consequences of HLA allotype variability beyond peptide binding motifs and underscore the importance of antigen processing in selection of the immunopeptidome.

We conclude there are fundamental differences in absolute HLA allotype expression and peptide binding properties brought about by allelic variation. Our results are consistent with the emerging hypothesis that allotype expression is inversely correlated with peptide repertoire diversity. The next step will be complete characterization of the allotype specific expression cycle to reveal different functional properties.

Our study of the genetic origins of MAPs stands out because we studied an exceptional number of HLA allotypes which was essential to understanding common sources of MAPs. We contribute the largest dataset of HLA-A,B MAPs identified to date (25,172 MAPs). Our findings show MAPs derive from a subset of genes and gene products with distinct features. These features can be used to predict with good accuracy whether or not a given gene can produce MAPs. Perhaps our most controversial finding is that CD8+ T cells monitor a fraction of the protein coding genome (<17%) because only 58% of genes generate MAPs and MAPs occur in clusters. We suggest applying more flexible machine learning techniques to the same data and incorporating quantitative information from MS may improve predictive power. We also expect our predictive model will facilitate the identification of neo-MAPs for immunotherapy based treatments of cancer.

Our results raise fundamental questions to direct future research:

- What is the impact of differential HLA expression in associated disease phenotypes?
- What local sequence features contribute to the short range effect where MAPs co-localize along source proteins?

- What distinguishes genes that produce many versus few MAPs?
- To what extent can we predict neo-MAPs using modeling and sequencing data?
- What are the functional consequences of the finding that CD8+ T cells monitor only a fraction of the genome?

Bibliography

1. 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). doi:10.1038/leu.2016.22.

2. Litman GW, Rast JP, Fugmann SD. The origins of vertebrate adaptive immunity. *Nat Rev Immunol*. 2010;10(8):543-553. doi:10.1038/nri2807.

3. Murphy K. *Janeway's Immunobiology*.; 2014. doi:10.1007/s13398-014-0173-7.2.

4. Pancer Z, Cooper M. the Evolution of Adaptive Immunity. *Annu Rev Immunol*. 2006;24(1):497-518. doi:10.1146/annurev.immunol.24.021605.090542.

5. den Haan JMM, Arens R, van Zelm MC. The activation of the adaptive immune system: Cross-talk between antigen-presenting cells, T cells and B cells. *Immunol Lett*. 2014;162(2):103-112. doi:10.1016/j.imlet.2014.10.011.

6. Gardiner CM, Mills KHG. The cells that mediate innate immune memory and their functional significance in inflammatory and infectious diseases. *Semin Immunol*. 2016:1-8. doi:10.1016/j.smim.2016.03.001.

7. Smith-Garvin JE, Koretzky GA, Jordan MS. T Cell Activation. *Annu Rev Immunol*. 2009;27(1):591-619. doi:10.1146/annurev.immunol.021908.132706.

8. Parker DC. T Cell-Dependent B Cell Activation. *Annu Rev Immunol*. 1993;11(1):331-360. doi:10.1146/annurev.iy.11.040193.001555.

9. Harwood NE, Batista FD. Early Events in B Cell Activation. *Annu Rev Immunol*. 2010;28(1):185-210. doi:10.1146/annurev-immunol-030409-101216.

10. McHeyzer-Williams LJ, McHeyzer-Williams MG. ANTIGEN-SPECIFIC MEMORY B CELL DEVELOPMENT. *Annu Rev Immunol*. 2004;23(1):487-513. doi:10.1146/annurev.immunol.23.021704.115732.

11. Kurosaki T, Shinohara H, Baba Y. B Cell Signaling and Fate Decision. *Annu Rev*

Immunol. 2010;28(1):21-55. doi:10.1146/annurev.immunol.021908.132541.

12. Neefjes J, Jongsma ML, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol*. 2011;11(12):823-836. doi:10.1038/nri3084.

13. Blum JS, Wearsch P a., Cresswell P. *Pathways of Antigen Processing*.; 2013. doi:10.1146/annurev-immunol-032712-095910.

14. Verteuil D de, Granados D, Thibault P, Perreault C. Origin and plasticity of MHC I-associated self peptides. *Autoimmun Rev*. 2012;11(9):627-635. doi:10.1016/j.autrev.2011.11.003.

15. 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.

16. Alexandropoulos K, Danzl NM. Thymic epithelial cells: Antigen presenting cells that regulate T cell repertoire and tolerance development. *Immunol Res*. 2012;54(1-3):177-190. doi:10.1007/s12026-012-8301-y.

17. Lo W-L, Allen PM. Thymic Development and Selection of T Lymphocytes. In: Boehm T, Takahama Y, eds. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014:49-67. doi:10.1007/82_2013_319.

18. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* . 43 (D1):D423-D431. http://nar.oxfordjournals.org/content/43/D1/D423.abstract.

19. Blum JS, Wearsch PA, Cresswell P. Pathways of Antigen Processing. *Annu Rev Immunol*. 2013;31(1):443-473. doi:10.1146/annurev-immunol-032712-095910.

20. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. *BMC Immunol*. 2008;9:1. doi:10.1186/1471-2172-9-1.

21. Harjanto S, Ng LFP, Tong JC. Clustering HLA class I superfamilies using structural

interaction patterns. *PLoS One*. 2014;9(1):e86655. doi:10.1371/journal.pone.0086655.

22. Gherardi E. The antibody, T cell receptor and MHC loci. *Univ Pavia*. http://nfs.unipv.it/nfs/minf/dispense/immunology/lectures/files/loci_abs_tcr_mhc.html . Accessed March 20, 2016.

23. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet*. 2013;14:301-323. doi:10.1146/annurev-genom-091212-153455.

24. Prugnolle F, Manica A, Charpentier M, Guégan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol*. 2005;15(11):1022-1027. doi:10.1016/j.cub.2005.04.050.

25. 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*. 2011;68(9):1521-1532. doi:10.1007/s00018-011-0659-9.

26. Brodsky FM, Parham P, Barnstable CJ, Crumpton MJ, BOdmer WF. Monoclonal Antibodies for Analysis of the HLA System. *Immunol Rev*. 1979;47(1):3-61. doi:10.1111/j.1600-065X.1979.tb00288.x.

27. Falk K, Rötzschke O, Stevanović S, Jung G, Rammensee HG. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature*. 1991;351(6324):290-296. doi:10.1038/351290a0.

28. Stevanovic S, Schild H. Quantitative aspects of T cell activation—peptide generation and editing by {MHC} class I molecules. *Semin Immunol*. 1999;11(6):375-384. doi:http://dx.doi.org/10.1006/smim.1999.0195.

29. Chappell P, Meziane EK, Harrison M, et al. Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding. *Elife*. 2015;(April).

30. 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.

31. Vigneron N, Van den Eynde B. Proteasome Subtypes and Regulators in the Processing of Antigenic Peptides Presented by Class I Molecules of the Major Histocompatibility Complex. *Biomolecules*. 2014;4(4):994-1025. doi:10.3390/biom4040994.

32. de Verteuil D, Muratore-Schroeder TL, Granados DP, 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. doi:10.1074/mcp.M900566-MCP200.

33. Van Hateren A, James E, Bailey A, Phillips A, Dalchau N, Elliott T. The cell biology of major histocompatibility complex class i assembly: Towards a molecular understanding. *Tissue Antigens*. 2010;76(4):259-275. doi:10.1111/j.1399-0039.2010.01550.x.

34. Paulsson KM, Wang P. Quality control of MHC class I maturation. *FASEB J*. 2004;18(1):31-38. doi:10.1096/fj.03-0846rev.

35. Geironson L, Thuring C, Harndahl M, et al. Tapasin facilitation of natural HLA-A and -B allomorphs is strongly influenced by peptide length, depends on stability, and separates closely related alloconmorphs. *J Immunol*. 2013;191(7):3939-3947. doi:10.4049/jimmunol.1201741.

36. Yanaka S, Ueno T, Shi Y, et al. Peptide-dependent conformational fluctuation determines the stability of the human leukocyte antigen class I complex. *J Biol Chem*. 2014;289(35):24680-24690. doi:10.1074/jbc.M114.566174.

37. Joffre OP, Segura E, Savina A, Amigorena S. Cross-presentation by dendritic cells. *Nat Rev Immunol*. 2012;12(8):557-569. doi:10.1038/nri3254.

38. Vyas JM, Van Der Veen AG, Ploegh HL. The known unknowns of antigen processing and presentation. *Nat Rev Immunol*. 2008;8(8):607-618. doi:10.1038/nri2368.

39. Apcher S, Daskalogianni C, Fåhraeus R. Pioneer translation products as an alternative

source for MHC-I antigenic peptides. *Mol Immunol*. 2015:2008-2011. doi:10.1016/j.molimm.2015.04.019.

40. Laumont CM, Daouda T, Laverdure J-P, 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.

41. Yewdell JW, Antón LC, Bennink JR. Defective Ribosomal Products (DRiPs): A Major Source of Antigenic Peptides for MHC Class I Molecules? *J Immunol*. 1996;157(5):1823-1826. http://www.scopus.com/inward/record.url?eid=2-s2.0-0030239693&partnerID=tZOtx3y1.

42. 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-e191. doi:10.1182/blood-2012-02-412593.

43. Apcher S, Daskalogianni C, Lejeune F, et al. Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *Proc Natl Acad Sci U S A*. 2011;108(28):11572-11577. doi:10.1073/pnas.1104104108.

44. 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.

45. 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:http://dx.doi.org/10.1016/j.it.2014.01.002.

46. Yewdell JW. DRiPs solidify: progress in understanding endogenous MHC class I antigen processing. *Trends Immunol*. 2011;32(11):548-558. doi:10.1016/j.it.2011.08.001.

47. Huppa JB, Davis MM. Chapter One - The Interdisciplinary Science of T-cell Recognition. In: Immunology FWABT-A in, ed.Vol Volume 119. Academic Press; 2013:1-50. doi:http://dx.doi.org/10.1016/B978-0-12-407707-2.00001-1.

48. Wooldridge L, Ekeruche-Makinde J, van den Berg HA, et al. A Single Autoimmune T Cell Receptor Recognizes More Than a Million Different Peptides. *J Biol Chem* . 2012;287 (2):1168-1177. http://www.jbc.org/content/287/2/1168.abstract.

49. Girdlestone J. Regulation of HLA Class I Loci by Interferons. *Immunobiology*. 1995;193(2–4):229-237. doi:http://dx.doi.org/10.1016/S0171-2985(11)80548-6.

50. Paul S, Weiskopf D, Angelo M a., 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. doi:10.4049/jimmunol.1302101.

51. Vita R, Overton JA, Greenbaum JA, et al. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* . 43 (D1):D405-D412. http://nar.oxfordjournals.org/content/43/D1/D405.abstract.

52. Caron E, Espona L, Kowalewski DJ, et al. An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife*. 2015;4(JULY 2015):1-17. doi:10.7554/eLife.07661.

53. Hassan C, Kester MGD, de Ru AH, et al. The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol Cell Proteomics*. 2013;12(7):1829-1843. doi:10.1074/mcp.M112.024810.

54. 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. doi:10.1074/mcp.M114.042812.

55. 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. doi:10.1371/journal.pcbi.1002517.

56. 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. doi:10.1038/ncomms4600.

57. Caron E, Kowalewski DJ, Koh CC, Sturm T. Analysis of MHC immunopeptidomes using mass spectrometry. 2015:3105-3117. doi:10.1074/mcp.O115.052431.

58. Nielsen M, Lundegaard C, Worning P, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*. 2003;12(5):1007-1017. doi:10.1110/ps.0239403.

59. Sieker F, Zacharias AM and M. Predicting Affinity and Specificity of Antigenic Peptide Binding to Major Histocompatibility Class I Molecules. *Curr Protein Pept Sci*. 2009;10(3):286-296. doi:http://dx.doi.org/10.2174/138920309788452191.

60. Overwijk WW, Wang E, Marincola FM, Rammensee H-G, Restifo NP. Mining the mutanome: developing highly personalized Immunotherapies based on mutational analysis of tumors. *J Immunother cancer*. 2013;1(1):11. doi:10.1186/2051-1426-1-11.

61. Robbins PF, Lu Y-C, El-Gamil M, 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. http://dx.doi.org/10.1038/nm.3161.

62. van Rooij N, van Buuren MM, Philips D, et al. Tumor Exome Analysis Reveals Neoantigen-Specific T-Cell Reactivity in an Ipilimumab-Responsive Melanoma. *J Clin Oncol*. 2013;31(32):e439-e442. doi:10.1200/JCO.2012.47.7521.

63. 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(suppl 2):W509-W512.

64. Karosiene E, Lundegaard C, Lund O, Nielsen M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*. 2011;64(3):177-186. doi:10.1007/s00251-011-0579-8.

65. R Core Team. R: A Language and Environment for Statistical Computing. 2015. http://www.r-project.org/.

66. Jianhong, Ou; Zhu LJ. motifStack: Plot stacked logos for single or multiple DNA,

RNA and amino acid sequence. 2014.

67. Trolle T, McMurtrey CP, Sidney J, 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. doi:10.4049/jimmunol.1501721.

68. Tynan FE, Borg NA, Miles JJ, et al. High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I: Implications for T-cell receptor engagement and T-cell immunodominance. *J Biol Chem*. 2005;280(25):23900-23909. doi:10.1074/jbc.M503060200.

69. Burrows SR, Rossjohn J, McCluskey J. Have we cut ourselves too short in mapping CTL epitopes? *Trends Immunol*. 2006;27(1):11-16. doi:10.1016/j.it.2005.11.001.

70. Samino Y, López D, Guil S, Saveanu L, Van Endert PM, Del Val M. A long N-terminal-extended nested set of abundant and antigenic major histocompatibility complex class I natural ligands from HIV envelope proteinSamino, Yolanda, Daniel López, Sara Guil, Loredana Saveanu, Peter M. Van Endert, and Margarita Del Val. 2006. *J Biol Chem*. 2006;281(10):6358-6365. doi:10.1074/jbc.M512263200.

71. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York; 2009. http://had.co.nz/ggplot2/book.

72. Blazar BR, Murphy WJ, Abedi M. Advances in graft-versus-host disease biology and therapy. *Nat Rev Immunol*. 2012;12(6):443-458. doi:10.1038/nri3212.

73. Berkers CR, de Jong A, Schuurman KG, et al. Definition of Proteasomal Peptide Splicing Rules for High-Efficiency Spliced Peptide Presentation by MHC Class I Molecules. *J Immunol*. 2015;195(9):4085-4095. doi:10.4049/jimmunol.1402455.

74. Dunn GP, Old LJ, Schreiber RD. The three Es of cancer immunoediting. *Annu Rev Immunol*. 2004;22(4):329-360. doi:10.1146/annurev.immunol.22.012703.104803.

75. Zhou F. Molecular mechanisms of viral immune evasion proteins to inhibit MHC class I antigen processing and presentation. *Int Rev Immunol*. 2009;28(5):376-393.

doi:10.1080/08830180903013034.

76. McLaren PJ, Carrington M. The impact of host genetic variation on infection with HIV-1. *Nat Immunol*. 2015;16(6):577-583. doi:10.1038/ni.3147.

77. Seldin MF. The genetics of human autoimmune disease: A perspective on progress in the field and future directions. *J Autoimmun*. 2015;64:1-12. doi:10.1016/j.jaut.2015.08.015.

78. Brown MA, Kenna T, Wordsworth BP. Genetics of ankylosing spondylitis—insights into pathogenesis. *Nat Rev Rheumatol*. 2015;12(2):1-11. doi:10.1038/nrrheum.2015.133.

79. Schittenhelm RB, Sian TCCLK, Wilmann PG, Dudek NL, Purcell AW. Revisiting the arthritogenic peptide theory: Quantitative not qualitative changes in the peptide repertoire of HLA-B27 allotypes. *Arthritis Rheumatol*. 2015;67(3):702-713. doi:10.1002/art.38963.

80. Schittenhelm RB, Sivaneswaran S, Lim Kam Sian TCC, Croft NP, Purcell AW. HLA-B27 allotype-specific binding and candidate arthritogenic peptides revealed through heuristic clustering of DIA-MS data. *Mol Cell Proteomics*. 2016. http://www.mcponline.org/content/early/2016/02/29/mcp.M115.056358.abstract.

81. Schlom J, Hodge JW, Palena C, et al. Chapter Two - Therapeutic Cancer Vaccines. In: Tew KD, Fisher PB, eds.Vol 121. Advances in Cancer Research. Academic Press; 2014:67-124. doi:http://dx.doi.org/10.1016/B978-0-12-800249-0.00002-0.

82. Desrichard A, Snyder A, Chan TA. Cancer Neoantigens and Applications for Immunotherapy. *Clin Cancer Res* . 2016;22 (4):807-812. http://clincancerres.aacrjournals.org/content/22/4/807.abstract.

83. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science (80-)*. 2015;348(6230):62-68. doi:10.1126/science.aaa4967.

84. Guo C, Manjili MH, Subjeck JR, Sarkar D, Fisher PB, Wang X-Y. Therapeutic

Cancer Vaccines. 2013;125(9):421-475. doi:10.1016/B978-0-12-407190-2.00007-1.

85. Hundal J, Carreno BM, Petti AA, et al. pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med*. 2016;8(1):11. doi:10.1186/s13073-016-0264-5.

86. Hui-Yuen J, McAllister S, Koganti S, Hill E, Bhaduri-McIntosh S. Establishment of Epstein-Barr Virus Growth-transformed Lymphoblastoid Cell Lines. *J Vis Exp*. 2011;(57):1-6. doi:10.3791/3321.

87. Demanet C, Mulder A, Deneys V, et al. Down-regulation of HLA-A and HLA-Bw6, but not HLA-Bw4, allospecificities in leukemic cells: An escape mechanism from CTL and NK attack? *Blood*. 2004;103(8):3122-3130. doi:10.1182/blood-2003-07-2500.

88. Košmrlj A, Read EL, Qi Y, et al. Effects of thymic selection of the T cell repertoire on HLA-class I associated control of HIV infection. 2011;465(7296):350-354. doi:10.1038/nature08997.Effects.

89. Berlin C, Kowalewski DJ, Schuster H, et al. Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. *Leukemia*. 2014;29(April):1-13. doi:10.1038/leu.2014.233.

90. Houldcroft CJ, Petrova V, Liu JZ, et al. Host genetic variants and gene expression patterns associated with epstein-barr virus copy number in lymphoblastoid cell lines. *PLoS One*. 2014;9(10). doi:10.1371/journal.pone.0108384.

91. Garstka M a., Fritzsche S, Lenart I, et al. Tapasin dependence of major histocompatibility complex class I molecules correlates with their conformational flexibility. *FASEB J*. 2011;25(11):3989-3998. doi:10.1096/fj.11-190249.

92. Belicha-Villanueva A, McEvoy S, Cycon K, Ferrone S, Gollnick SO, Bangia N. Differential contribution of TAP and tapasin to HLA class I antigen expression. *Immunology*. 2008;124(1):112-120. doi:10.1111/j.1365-2567.2007.02746.x.

93. Schellens IMM, Hoof I, Meiring HD, et al. Comprehensive Analysis of the Naturally

Processed Peptide Repertoire: Differences between HLA-A and B in the Immunopeptidome. *PLoS One*. 2015;10(9):e0136417. doi:10.1371/journal.pone.0136417.

94. Theodossis A. On the trail of empty MHC class-I. *Mol Immunol*. 2013;55(2):131-134. doi:10.1016/j.molimm.2012.10.012.

95. Galati G, Arcelloni C, Paroni R, et al. Quantitative cytometry of MHC class I digestion from living cells. *Cytometry*. 1997;27(1):77-83. doi:10.1002/(SICI)1097-0320(19970101)27:1<77::AID-CYTO10>3.0.CO;2-P.

96. Antwi K, Hanavan PD, Myers CE, Ruiz YW, Thompson EJ, Lake DF. Proteomic identification of an MHC-binding peptidome from pancreas and breast cancer cell lines. *Mol Immunol*. 2009;46(15):2931-2937. doi:10.1016/j.molimm.2009.06.021.

97. Badrinath S, Huyton T, Blasczyk R, Bade Doeding C. HLA Class I Polymorphism and Tapasin Dependency. In: *HLA and Associated Important Diseases*. InTech; 2014. doi:10.5772/57495.

98. Miyadera H, Ohashi J, Lernmark Å, Kitamura T, Tokunaga K. Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA. *J Clin Invest*. 2015;125(1):275-291. doi:10.1172/JCI74961.

99. Reddy P, Maeda Y, Liu C, Krijanovski OI, Korngold R, Ferrara JLM. A crucial role for antigen-presenting cells and alloantigen expression in graft-versus-leukemia responses. *Nat Med*. 2005;11(11):1244-1249.

100. Rammensee HG, Bachmann J, Stevanovic S. MHC ligands and peptide motifs Landes Bioscience. *Georg TX*. 1997.

101. Daouda T, Perreault C, Lemieux S. pyGeno: A Python package for precision medicine and proteogenomics. *F1000Research*. 2016;5(701564):381. doi:10.12688/f1000research.8251.1.

102. González-Galarza FF, Takeshita LYC, Santos EJM, et al. Allele frequency net 2015

update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res*. 2014:gku1166.

103. Gasteiger E, Hoogland C, Gattiker A, et al. Protein Identification and Analysis Tools on the ExPASy Server. *Proteomics Protoc Handb*. 2005:571-607. doi:10.1385/1-59259-890-0:571.

104. Schuerwegh AJ, Stevens WJ, Bridts CH, De Clerck LS. Evaluation of monensin and brefeldin A for flow cytometric determination of interleukin-1 beta, interleukin-6, and tumor necrosis factor-alpha in monocytes. *Cytometry*. 2001;46(3):172-176.

105. Tan a C, Gilbert D. An empirical comparison of supervised machine learning techniques in bioinformatics. 2009;19(Apbc). http://hdl.handle.net/2438/3020.

106. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: A methodology review. *J Biomed Inform*. 2002;35(5-6):352-359. doi:10.1016/S1532-0464(03)00034-0.

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

108. Yadav M, Jhunjhunwala S, Phung QT, et al. Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*. 2014;515(7528):572-576. http://dx.doi.org/10.1038/nature14001.

109. Blankenstein T, Leisegang M, Uckert W, et al. Targeting cancer-specific mutations by T cell receptor gene therapy. *Curr Opin Immunol*. 2015;33(8):112-119. doi:10.1016/j.coi.2015.02.005.

110. Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014;30(23):3310-3316. doi:10.1093/bioinformatics/btu548.

111. Korthauer KD, Kendziorski C. MADGiC: A model-based approach for identifying driver genes in cancer. *Bioinformatics*. 2014;31(10):1526-1535. doi:10.1093/bioinformatics/btu858.

112. Radenbaugh AJ, Ma S, Ewing A, et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One*. 2014;9(11). doi:10.1371/journal.pone.0111516.

113. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Poznań, Poland)*. 2015;19(1A):A68-A77. doi:10.5114/wo.2014.47136.

114. Pon JR, Marra MA. Driver and Passenger Mutations in Cancer. *Annu Rev Pathol Mech Dis*. 2015;10:25-50. doi:10.1146/annurev-pathol-012414-040312.

115. Zhang J, Liu J, Sun J, Chen C, Foltz G, Lin B. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. *Brief Bioinform*. 2014;15(2):244-255. doi:10.1093/bib/bbt042.

116. 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. doi:10.1002/wrna.1262.

117. Pacis A, Nedelec Y, Barreiro LB. When genetics meets epigenetics: Deciphering the mechanisms controlling inter-individual variation in immune responses to infection. *Curr Opin Immunol*. 2014;29(1):119-126. doi:10.1016/j.coi.2014.06.002.

118. Fairfax BP, Knight JC. Genetics of gene expression in immunity to infection. *Curr Opin Immunol*. 2014;30(1):63-71. doi:10.1016/j.coi.2014.07.001.

Appendix 1 - Protocol for QIFIKIT quantitation of MHCI expression on B-LCLs

Introduction

This protocol is for quantitation of MHCI surface molecules on B-LCLs using the QIFIKIT and indirect immunofluorescence protocol for flow cytometry.

Materials

- FACS tubes (BD Falcon)
- Cold PBS/BSA (0.1%)
- Unlabeled Mouse IgG primary antibody (ex. Mouse IgG2b anti-Human HLA-A2 clone BB7.2, BD cat#551230)
- Unlabeled Mouse IgG primary isotype control (ex. Mouse IgG2b Isotype Control clone MPC-11, BD cat#557351)
- Fluorochrome labeled secondary antibody (BV421 Goat anti-Mouse Ig polyclonal, BD cat#563846)
- QIFIKIT® vials 1 and 2 (Dako, cat#K0078)

Table III. Primary antibody dilutions and product information for indirect immunofluorescence and quantitation of various HLA-A,B allotypes.

Antibody target	Clone	Isotype	Supplier	Product #	Dilution
HLA-ABC	W6/32	IgG2a	Abcam	ab7855	1:4
HLA-A1	4i93	IgG2a	Abcam	ab33641	1:8
HLA-A2	BB7.2	IgG2b	BD	551230	1:16
HLA-A3	GAP.A3	IgG2a	eBioscience	custom	1:8
HLA-A11	4i93	IgG2a	Abcam	4i93	1:16
HLA-A23/A24	4i94	IgG2b	Abcam	4i94	1:3
HLA-B7	BB7.1	IgG1	Santa Cruz	sc-53304	1:3
HLA-B27	HLA.ABC.M3	IgG2a	Merek	MAB1285	1:64

Notes

Caution: Use Biosafety level 2 in cell culture room and lab.

Note 1: Reserve FACS Canto II in advance. Estimate 5-15 minutes for setup + 1-1.5 minutes per tube + 15 minutes for fluidics shutdown.

Note 2: Aseptic technique is not necessary unless part of the cells will be kept in culture.

Note 3: Keep the cells on ice (4°C) all the times. All centrifugation steps must be carried out at 4°C, if possible. Use *cold* PBS/BSA 0.1% for all steps.

Note 4: Protect the cells from light after labeling with secondary antibody.

Note 5: Carry out 3-5 replicates of each cell line / condition.

Note 6: Tests with the Human BD Fc block (Cat#564219) showed no significant difference in extracellular staining.

Preparation

1. If fresh cell are required, thaw vials 3 days in advance.
2. Resuspend cells to 0.4 x 10^6 cells/ml 1 day in advance of FACS start time:
    a. Preform a cell count for each cell line with the Countess. Ensure samples are mixed well before count. Mix 10 µl 0.4% Trypan blue + 10µl cell suspension and pipette into counting slide. Record all measurements (including average viable cell size) on a USB drive for reference;
3. Prepare the following in advance of the experiment:
    a. PBS solution (at least 10ml x total number of tubes in experiment);
    b. Labeled FACS tubes;
    c. Experimental setup recording folder on FACS Canto II.

Procedure

Cell culture room

4. Preform a cell counts noting the average concentration and viability of cells for each line.
5. Calculate the volume of cell culture required for the number of cells for each FACS tube (50,000 cells).

    # cells to be labeled / [ ] x 10^6 (cells/ml) = volume to transfer

    ex. 50,000 / 0.3 x 10^6 cells per ml = 0.166 or 16.6µl
6. Transfer slightly more that the total volume required for all FACS tubes to an Eppendorf or 15ml Falcon tube for each cell line.

    Ex. (50 samples + 5 (1 per 10 samples)) x 0.166 = 9.130 ml total

Benchtop

7. Pipette calculated volume for 50,000 cells into each FACS tube.
8. **Wash to remove medium**:
    a. Add 1-2ml of cold PBS/BSA 0.1% to each tube;
    b. Centrifuge at 1000 rpm x 5 min at 4°C, primary antibody solutions can be prepared during centrifugation (step 9);
    c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.
9. **Prepare the antibody and isotype solutions** for primary labeling:
    a. Primary HLA-A2 at a 1:16 dilution;

Ex. For 100 tubes:	Total Volume	[100 + (100*0.15)] * 10 = 1150 µl
	Antibody volume	1150*(1/16) = 71.88 µl
	PBS/BSA (0.1%) volume	1150*(15/16) = 1078.13 µl

10. **Label cells with 1° antibody:**
    a. Pipette 10µl of diluted primary purified antibody to each test FACS tube and 10µl of diluted primary isotype antibody into isotype control tubes;
    b. *Vortex gently*;
    c. Incubate at 4°C for 30 min, preparation of QIFIKIT® beads can be done during this step.
11. **Prepare QIFIKIT® Beads:**
    a. Vortex and pipette 50µl of from Vial 1 and Vial 2 into 2 separate FACS tubes;
    b. Treat tubes with beads as samples for the rest of the protocol.
12. **Wash to remove excess 1° antibody**:
    a. Add 1-2ml of cold PBS/BSA 0.1% to each tube, vortex gently;
    b. Centrifuge at 1000 rpm x 5 min at 4°C, secondary antibody solutions can be prepared during centrifugation (step 12);
    c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.
13. **Prepare the antibody solution** for secondary labeling:
    a. Secondary BV421 antibody at a 1:2 dilution;

Ex. For 100 tubes:	Total Volume	[100 + (100*0.25)] * 10 =1250 µl
	Antibody volume	1250*(1/2) = 625.00 µl
	PBS/BSA (0.1%) volume	1250*(1/2) = 625.00 µl

14. **Label cells with 2° antibody:**

a. Pipette 10µl of diluted secondary purified antibody to each test FACS tube, primary isotype control tubes, secondary isotype control tubes, and QIFIKIT® beads tubes;

b. *Vortex gently*;

c. Incubate at 4˚C for 30 min.

15. **Wash to remove excess 2˚ antibody**:

a. Add 1-2ml of cold PBS/BSA 0.1% to each tube;

b. Centrifuge at 1000 rpm x 5 min at 4˚C,

c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.

16. Resuspend the cells in 50 ul of cold PBS/BSA 0.1%.

17. Store at 4˚C until ready for FACS analysis.

18. Analyze the cells in the FACS CANTO II, keep samples on ice until analysis.

a. To maximize statistical power, increase # events recorded to 25,000.

b. Refer to QIFIKIT® protocol (7th edition) for Data Acquisition guidelines.

19. **Setting Window of Analysis**

a. Using the set up beads, adjust laser voltage so positive and negative populations fall within 15% of the reference values (below).

Table IV. Reference values for consistent quantitation using the QIFIKIT on a BD FACSCANTO II. For best results, approximate voltages should be adjusted to place set up bead populations within near to approximate MFI.

			Approximate values
Voltages (starting point)		FSC (beads)	400
		SSC (beads)	400
		FSC (cells)	280
		SSC (cells)	520
		BV421	245
		7-AAD	640
MFI	Set Up Beads	Negative	~100 a.u.

		Positive	~20,000 a.u.

20. **Flow Cytometry**

a. Pass each FACS tube into appropriately labeled experiments, gather ~10,000 events per tube, ensuring good viability and sufficient populations in the live cell gate.

b. Save experiment and run clean cycle (5 minutes bleach, 2 x 5 minute rinse).

21. Data Analysis (See QIFIKIT Manual for reference)

a. Gate on live cells, obtain BV421 MFI for each population of calibration beads and each experiment tube.

b. Obtain the calibration curve by a log10 transformation of the MFI and Antibody Binding Capacity (ABC, lot specific values included with each kit) and plotting log(ABC) = a x log(MFI) + b.

c. Use the linear regression equation of the calibration curve to transform each sample MFI value to ABC.

d. Subtract the ABC of the appropriate isotope control (background antibody equivalent) from each sample to obtain the Specific Antibody-Binding Capacity (SABC).

Appendix 2 - Protocol for mild acid elution of surface MHCI peptides on B-LCLs

Introduction

This protocol describes mild acid elution of B-LCLs using a citrate phosphate buffer to release MHCI bound peptides from the binding groove and induce MHCI internalization. The protocol is adapted from various sources (see References). The protocol may be used prior to FACS analysis and may have applications elsewhere (ex. isolation of MHCI peptides for immunopeptidome analysis, see references).

Precautions

Caution: Use Biosafety level 2 in cell culture room and lab.

Note 1: As the cells will be put in culture during the procedure, aseptic technique is required.

Note 2: Keep the cells on ice (4°C) unless otherwise noted. All centrifugation steps must be carried out at 4°C, if possible. Use *cold* RPMI for washing steps.

Note 3: Carry out 3-5 replicates of each cell line / condition.

Materials

- Citric Acid (Sigma, C2402, CAS# 77-92-9)
- Cold PBS/BSA (0.1%)
- Sodium Phosphate Dibasic (Sigma Aldrich, S9390, CAS#7782-85-6)
- Sodium Chloride (Sigma Aldrich, S7653, CAS#7647-14-5)
- RPMI Complete (RPMI 1640 + HEPES + L-Glutamine, 10% FBS, 1% L-Glutamine, 1% Pen-Strep, ThermoFisher #74200047)

Reagent Preparation

Table V. Recipe for preparation of citrate phosphate buffer for mild acid elution.

Reagent	**Molecular Weight**	**Concentration**	**Quantity per 100 ml**
Citric Acid ($C_6H_8O_7$)	192.1 g/mol	131 mM	2.5160 g
Sodium Phosphate Dibasic	268.1 g/mol	66 mM	0.9369 g

(Na_2HPO_4 $7H_2O$)			
Sodium Chloride (NaCl)	58.44 g/mol	150 mM	0.8766 g
Water	-	-	100 mL

Preparation: Buffer can be prepared in advance, autoclaved, and stored at room temperature; verify no contamination has occurred prior to each experiment.

Procedure

1. Adjust the citrate phosphate buffer pH to 3.3 using a pH meter and concentrated NaOH or HCl accordingly.
2. Preform a cell count for each cell line. Ensure samples are mixed well before count.
    a. Mix 10 µl 0.4% Trypan blue + 10µl cell suspension and pipette into counting slide.
    b. Calculate the volume of cell culture required for 600,000 cells.
3. Transfer volume for 600,000 cells to a FACS test tube.
4. Wash to remove medium:
    a. Add 1-2ml of cold RPMI to each tube;
    b. Centrifuge at 1000 rpm x 5 min at 4˚C, primary antibody solutions can be prepared during centrifugation (step 9);
    c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.
5. Mild acid elution:
    a. Attach a tube rack to a plate vortex machine.
    b. Add 200µL of citrate phosphate buffer to each sample and vortex gently in rack for desired time (15 seconds - 1.5 minutes for partial elution; 5 minutes for complete elution; exceeding 8 minutes will lead to eventual cell death if cells are returned to culture)
    c. Neutralize the medium by adding 1-2 mL RPMI 1640 complete.
6. Wash to remove citrate phosphate buffer and RPMI:
    a. Add 1-2ml of cold RPMI to each tube;
    b. Centrifuge at 1000 rpm x 5 min at 4˚C, primary antibody solutions can be prepared during centrifugation (step 9);
    c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.
7. Resuspend each pellet in 1 mL PBS/BSA.

References

Granados, Diana Paola, Dev Sriranganadane, Tariq Daouda, Antoine Zieger, Céline M Laumont, Olivier Caron-Lizotte, Geneviève Boucher, et al. 2014. "Impact of Genomic Polymorphisms on the Repertoire of Human MHC Class I-Associated Peptides." *Nature Communications* 5 (January): 3600. doi:10.1038/ncomms4600.

Storkus, Walter J, Herbert J Zen III, Russell D Salter, and Michael T Lotze. 1993. "Identification of T-Cell Epitopes: Rapid Isolation of Class I-Presented Peptides from Viable Cells by Mild Acid Elution." *Journal of Immunotherapy* 14 (2). LWW: 94–103.

Appendix 3 - Protocol for papain digestion of surface MHCI on B-LCLs

Introduction

This protocol describes treatment of B-LCLs with the papain protease to cleave surface MHC Class I molecules without inducing cell death. The protocol is adapted from various sources (see References). The protocol may be used prior to FACS analysis and may have applications elsewhere (ex. isolation of peptide MHC complexes for immunopeptidome analysis, see references).

Precautions

Caution: Use Biosafety level 2 in cell culture room and lab.

Note 1: As the cells will be put in culture during the procedure, aseptic technique is required.

Note 2: Keep the cells on ice (4°C) unless otherwise noted. All centrifugation steps must be carried out at 4°C, if possible. Use *cold* RPMI for washing steps.

Note 3: Carry out 3-5 replicates of each cell line / condition.

Materials

- Papain from papaya latex (Sigma, P1325, CAS# 9001-73-4)
- L-Cysteine (Sigma Aldrich, C7352, CAS#52-90-4)
- EDTA (Sigma Aldrich, 60-00-4)
- RPMI Complete (RPMI 1640 + HEPES + L-Glutamine, 10% FBS, 1% L-Glutamine, 1% Pen-Strep, ThermoFisher #74200047)

Reagent Preparation

Table VI. Recipe for preparation of papain buffer.

Reagent	Initial parameters	Final Concentration	Quantity per ml
Papain	868 U/ml*	75 units/ml	86.4 ul
L-Cysteine	121.6 g/mol	20 mM	0.002423 mg

EDTA	372.24 g/mol	1 mM	0.00037224 mg
RPMI Complete	-	-	913.6 ul

Preparation: dilute L-Cysteine and EDTA in RPMI complete < 24 hours before experiment start. Add Papain just prior to incubation to avoid loss of Proteolytic activity. * Depends on lot.

Procedure

8. Preform a cell count for each cell line. Ensure samples are mixed well before count.
    a. Mix 10 µl 0.4% Trypan blue + 10µl cell suspension and pipette into counting slide.
    b. Calculate the volume of cell culture required for 600,000 cells.
9. Transfer volume for 600,000 cells to a FACS test tube.
10. Wash to remove medium:
    a. Add 1-2ml of cold RPMI to each tube;
    b. Centrifuge at 1000 rpm x 5 min at 4°C, primary antibody solutions can be prepared during centrifugation (step 9);
    c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.
11. Incubation
    a. Resuspend cells in 913.6 ul Papain Buffer without Papain in a 24 well plate;
    b. Add 86.4 ul Papain to each well;
    c. Incubate at 37°C, 5% CO_2 for 45 minutes.
12. Wash to remove Papain 3X:
    a. Add 1-2ml of cold RPMI to each tube;
    b. Centrifuge at 1000 rpm x 5 min at 4°C, primary antibody solutions can be prepared during centrifugation (step 9);
    c. Discard the supernatant and tap gently the bottom of the tube to destroy the pellet.
13. Optional Re-expression Analysis: Incubate 600,000 cells in 1ml RPMI Complete at 37°C, 5% CO_2 for desired time to allow cells to re-express pMHC (1- 12 hours).

References

Antwi, Kwasi, Paul D Hanavan, Cheryl E Myers, Yvette W Ruiz, Eric J Thompson, and Douglas F Lake. 2009. "Proteomic Identification of an MHC-Binding Peptidome from

Pancreas and Breast Cancer Cell Lines." *Molecular Immunology* 46 (15): 2931–37. doi:10.1016/j.molimm.2009.06.021.

Galati, Giacomo, Cinzia Arcelloni, Rita Paroni, Silvia Heltai, Patrizia Rovere, Claudio Rugarli, and Angelo A. Manfredi. 1997. "Quantitative Cytometry of MHC Class I Digestion from Living Cells." *Cytometry* 27 (1): 77–83. doi:10.1002/(SICI)1097-0320(19970101)27:1<77::AID-CYTO10>3.0.CO;2-P.

Gebreselassie, Daniel, Hans Spiegel, and Stanislav Vukmanovic. 2006. "Sampling of Major Histocompatibility Complex Class I-Associated Peptidome Suggests Relatively Looser Global Association of HLA-B*5101 With Peptides." *Human Immunology* 67 (11): 894–906.

Ladasky, John J, Sarah Boyle, Malini Seth, Hewang Li, Tsvetelina Pentcheva, Fumiyoshi Abe, Steven J Steinberg, and Michael Edidin. 2006. "Bap31 Enhances the Endoplasmic Reticulum Export and Quality Control of Human Class I MHC Molecules." *Journal of Immunology (Baltimore, Md. : 1950)* 177 (9): 6172–81.

Pickl, W F, W Holter, J Stöckl, O Majdic, and W Knapp. 1996. "Expression of Beta 2-Microglobulin-Free HLA Class I Alpha-Chains on Activated T Cells Requires Internalization of HLA Class I Heterodimers." *Immunology* 88 (1): 104–9.

Appendix 4 - MiHA Annotation

Detected Peptide	Length	nsSNV Pos.	Affinity (nM)	Allele	Alternate Peptide	Affinity (nM)	RS ID	Chromosome	nsSNV Pos.	Gene ID	Gene Name	Transcript ID	nsSNV Pos.	Protein ID	nsSNV Pos.	Literature Status
AEELGGPVHAL	11	5	917.0 B*44:03	AEELAGPVHAL	408.0	rs1044808	1	149871904	ENSG00000178096	BOLA1	ENST00000369153	957	ENSP00000358149	98	Novel
AEIEQKIKEY	10	1	14.0 B*44:03	SEIEQKIKEY	20	rs2073498	3	50369545	ENSG00000068028	RASSF1	ENST00000359365	504	ENSP00000352323	133	MiHA
AEILRGVRL	9	3	193.0 B*44:03	AEVLRGVRL	339	rs2273137	20	2635211	ENSG00000101361	NOP56	ENST00000329276	877	ENSP00000370589	121	Novel
AEIQEKKEI	9	2	91.0 B*44:03	AAIQEKKEI	30788	rs17244028	11	27362358	ENSG00000109881	CCDC34	ENST00000328697	1465	ENSP00000330240	264	Novel
AELQGFHRSF	10	4	30.0 B*44:03	AELKGFHRSF	37	rs3746101	19	2050822	ENSG00000099875	MKNK2	ENST00000250896	273	ENSP00000250896	10	MiHC Binder
AELQSRLAA	9	5	472.0 B*44:03	AELQARLAA	351.0	rs6952	19	7747292	ENSG00000181029	TRAPPC5	ENST00000559148	341	ENSP00000470262	52	Novel
AENDFVHRI	9	7	10.0 B*44:03	AENDFVRRI	10.0	rs11546682	8	23146094	ENSG00000104679	R3HCC1	ENST00000411463	179	ENSP00000397555	60	MiHA
AENIDAQLKRM	11	5	99.0 B*44:03	AENIEAQLKRM	85.0	rs892028	19	50411741	ENSG00000213024	NUP62	ENST00000407331	1323	ENSP00000407331	441	MiHA
AEVEHVYNA	9	9	525.0 B*44:03	AEVEHVYNT	1232	rs61746217	17	79993918	ENSG00000169738	DCXR	ENST00000330869	702	ENSP00000330356	218	MiHA
AEVRAENL	8	3	419.0 B*44:03	AEIRAENL	234.0	rs61744267	1	186321241	ENSG00000047410	TPR	ENST00000367478	2632	ENSP00000356448	779	Novel
ALSGHLETV	9	9	19.0 A*02:01	ALSGHLETL	53	rs17845226	15	60653204	ENSG00000182718	ANXA2	ENST00000396024	452	ENSP00000379342	98	MHC Binder
DEIVCTQHW	9	6	14.0 B*44:03	DEIVCIQHW	11	rs1372085	13	25008587	ENSG00000102699	PARP4	ENST00000381989	4797	ENSP00000371419	1564	MiHA
DEIVCTQHW	9	6	14.0 B*44:04	DEIVCRQHW	32	rs1372086	14	25008588	ENSG00000102700	PARP5	ENST00000381990	4798	ENSP00000371420	1565	MiHA
EEALGLYHW	9	8	42.0 B*44:03	EEALGLYQW	52	rs41283558	20	47269232	ENSG00000124126	PREX1	ENST00000371941	2381	ENSP00000361009	786	Novel
EEEQSQSRW	9	2	86.0 B*44:03	EKEQSQSRW	145604.0	rs997173	12	55355032	ENSG00000135426	TESPA1	ENST00000449076	1619	ENSP00000400892	496	Novel
EEINLQRNI	9	7	96.0 B*44:03	EEINLQINI	104	rs2083914	8	6302153	ENSG00000147316	MCPH1	ENST00000344683	987	ENSP00000342924	304	MiHA
EEIPVSSHY	9	5	15.0 B*44:03	EEIPSSHY	13	rs7167216	15	91354520	ENSG00000197299	BLM	ENST00000355112	4079	ENSP00000347232	1321	MiHA
EEIPVSSHYF	10	5	18.0 B*44:03	EEIPSSHYF	17	rs7167216	15	91354520	ENSG00000197299	BLM	ENST00000355112	4079	ENSP00000347232	1321	MiHA
EELLAVGKF	9	7	80.0 B*44:03	EELLAVSKF	39.0	rs2152143	10	129906979	ENSG00000148773	MKI67	ENST00000368654	3500	ENSP00000357643	1042	Novel
EELLAVSKF	9	7	39.0 B*44:03	EELLAVGKF	80.0	rs2152143	10	129906979	ENSG00000148773	MKI67	ENST00000368654	3500	ENSP00000357643	1042	Novel
EENGRKEIDIKKY	13	10	45.0 B*44:03	EENGRKEIDVKKY	44.0	rs11548200	1	156290655	ENSG00000163468	CCT3	ENST00000413555	668	ENSP00000413308	219	Novel
EERGENTSY	9	2	220.0 B*44:03	EGRGENTSY	39259	rs10511	2	234749257	ENSG00000123485	HJURP	ENST00000441486	2234	ENSP00000414109	723	Novel
EESAVPERSW	10	7	45.0 B*44:03	EESAVPKRSW	33.0	rs2295283	1	12082925	ENSG00000116691	MIIP	ENST00000235332	668	ENSP00000235333	167	Novel
EESAVPKRSW	10	7	33.0 B*44:04	EESAVPERSW	45.0	rs2295284	2	12082926	ENSG00000116692	MIIP	ENST00000235333	669	ENSP00000235333	168	MiHA
EETAIYKGDY	10	1	57.0 B*44:03	QETAIYKGDY	64	rs1804080	4	89625426	ENSG00000138641	HERC3	ENST00000402738	3075	ENSP00000385684	946	Novel
EEYEELLHY	9	8	22.0 B*44:03	EEYEELLRY	16.0	rs2307111	5	75003677	ENSG00000152359	POC5	ENST00000428202	297	ENSP00000410216	36	Novel
EEYEELLRY	9	8	16.0 B*44:03	EEYEELLHY	22.0	rs2307111	5	75003677	ENSG00000152359	POC5	ENST00000428202	297	ENSP00000410216	36	Novel
EQQLLYRSW	9	2	925.0 B*44:03	ERQLLYRSW	9032.0	rs2006771	22	31998611	ENSG00000198089	SFI1	ENST00000432498	1946	ENSP00000402679	518	Novel
FLQAKQIAL	9	8	14.0 A*02:01	FLQAKQITL	12	rs2275660	13	25029217	ENSG00000102699	PARP4	ENST00000381989	2801	ENSP00000371419	899	Novel
FLSSANEHL	9	3	19.0 A*02:01	FLPSANEHL	34	rs2274217	10	131959149	ENSG00000108010	GLRX3	ENST00000481034	389	ENSP00000435445	123	MHC Binder
GEDKGIKAL	9	3	3242.0 B*44:03	GEGKGIKAL	1753.0	rs10082391	10	129901824	ENSG00000148773	MKI67	ENST00000368654	8279	ENSP00000357643	2760	Novel
GEYFAIKAL	9	2	285.0 B*44:03	GDYFAIKAL	18202.0	rs2230494	3	53220214	ENSG00000163932	PRKCD	ENST00000394729	1447	ENSP00000378217	373	Novel
GLSPLLQKI	9	2	156.0 A*02:01	GSSPLLQKI	19296	rs767987	4	54011525	ENSG00000184178	SCFD2	ENST00000401642	1669	ENSP00000384182	512	Novel
GQYTDLLRL	9	2	763.0 A*02:01	GHYTDLLRL	22095	rs1128416	2	29001690	ENSG00000213639	PPP1CB	ENST00000358506	361	ENSP00000351298	67	MHC Binder
HLEEQIAKV	9	7	118.0 A*02:01	HLEEQIAKV	63	rs9051	18	2616456	ENSG00000080986	NDC80	ENST00000261597	1995	ENSP00000261597	605	MHC Binder
HLEEQIPKV	9	7	63.0 A*02:01	HLEEQIAKV	118	rs9051	18	2616456	ENSG00000080986	NDC80	ENST00000261597	1995	ENSP00000261597	605	Novel
IEATGFDRL	9	5	4649.0 B*44:03	IEATEFDRL	3159.0	rs2304748	8	33369943	ENSG00000129696	TTI2	ENST00000431156	807	ENSP00000411169	63	Novel
IEDRQYKDY	9	2	1101.0 B*44:03	IKDRQYKDY	34506	rs187325799	19	50188064	ENSG00000126457	PRMT1	ENST00000532489	978	ENSP00000433556	158	Novel
ILAPCKLETV	10	1	19.0 A*02:01	SLAPCKLETV	21	rs1129495	5	115428380	ENSG00000145781	COMMD10	ENST00000274458	445	ENSP00000274458	128	Novel
ILLEDGTFKV	10	1	6.0 A*02:01	TLLEDGTFKV	10	rs11557236	16	4519438	ENSG00000153406	NMRAL1	ENST00000574733	798	ENSP00000458762	23	Novel
ILSEVERNL	9	9	335.0 A*02:01	ILSEVERNF	13191	rs2230194	1	156280970	ENSG00000163468	CCT3	ENST00000295688	1452	ENSP00000295688	391	Novel
KEFEDDINW	10	6	47.0 B*44:03	KEFEDGINW	39.0	rs3826007	15	80263216	ENSG00000140379	BCL2A1	ENST00000267953	572	ENSP00000267954	82	MiHA
KEFEDGIINW	10	6	39.0 B*44:04	KEFEDDIINW	47.0	rs3826008	16	80263217	ENSG00000140380	BCL2A2	ENST00000267954	573	ENSP00000267954	83	MiHA
KEIAKTVLI	9	4	878.0 B*44:03	KEITKTVLI	1126	rs1678674	6	38759378	ENSG00000124721	DNAH8	ENST00000327475	2794	ENSP00000333363	932	Novel
KENEKSNIL	10	8	1019.0 B*44:03	KEINEKSSIL	861	rs33999879	3	160131346	ENSG00000113810	SMC4	ENST00000337388	1518	ENSP00000349961	356	Novel
KENQEAEKL	9	3	455.0 B*44:03	KENQEAEKL	455	rs7935564	11	5718516	ENSG00000132274	TRIM22	ENST00000379965	740	ENSP00000369299	155	Novel
KILEKEIRV	9	8	36.0 A*02:01	KILEKEICV	34	rs299284	5	162896649	ENSG00000072571	HMMR	ENST00000393915	420	ENSP00000377492	93	Novel
KIRGVINQL	9	2	4880.0 A*02:01	KLRGVINQL	489	rs11556157	2	44028012	ENSG00000138036	DYNC2LI1	ENST00000260605	788	ENSP00000260605	230	Novel
KLAENIDAQL	10	7	49.0 A*02:01	KLAENIEAQL	55	rs8892028	19	50411741	ENSG00000213024	NUP62	ENST00000426090	1323	ENSP00000407331	441	MHC Binder
KLRGVINQL	9	2	489.0 A*02:01	KIRGVINQL	4880	rs11556157	2	44028012	ENSG00000138036	DYNC2LI1	ENST00000260605	788	ENSP00000260605	230	Novel
KTDKTLVLL	9	8	970.0 A*02:01	KTDKTLVML	1419	rs13204	17	4849283	ENSG00000108518	PFN1	ENST00000225655	954	ENSP00000225655	112	MHC Binder
KTDKTLVLL	9	8	970.0 A*02:02	KTDKTLVVL	2866	rs13205	18	4849284	ENSG00000108519	PFN2	ENST00000225656	955	ENSP00000225656	113	MHC Binder
LEADLPRSW	9	1	26.0 B*44:03	MEADLPRSW	7.0	rs11556913	8	125463543	ENSG00000183665	TRMT12	ENST00000328599	497	ENSP00000329858	126	Novel
LEADLPRSW	9	1	26.0 B*44:03	VEADLPRSW	16.0	rs11556913	8	125463543	ENSG00000183665	TRMT12	ENST00000328599	497	ENSP00000329858	126	Novel
LIDTSRHYL	9	2	1364.0 A*02:01	LVDTSRHYL	2915	rs10805890	5	73992880	ENSG00000049860	HEXB	ENST00000261416	736	ENSP00000261416	207	Novel
LLAAPAQA	8	5	41.0 A*02:01	LLWAGPVTA	124	rs1048719	5	150632831	ENSG00000196743	GM2A	ENST00000357164	380	ENSP00000349687	19	MHC Binder
LLWAGPVIA	9	5	71.0 A*02:01	LLLATPAQA	25	rs361437	7	142250987	ENSG00000211713	TRBV6-4	ENST00000390360	80	ENSP00000374883	20	Novel
MESINPHKY	9	4	17.0 B*44:03	MESMNPHKY	15	rs2230794	9	111659438	ENSG00000070061	IKBKAP	ENST00000374647	2798	ENSP00000363779	830	Novel

Detected Peptide	Length	nsSNV Pos.	Affinity (nM)	Allele	Alternate Peptide	Affinity (nM)	RS ID	Chromosome	nsSNV Pos.	Gene ID	Gene Name	Transcript ID	nsSNV Pos.	Protein ID	nsSNV Pos.	Literature Status
MLRSELLL	8	5	1110.0	A*02:01	MLRSQLLL	870.0	rs27044	5	96118851	ENSG00000164307	ERAP1	ENST00000296754	2446	ENSP00000296754	730	Novel
NESNTQKTY	9	1	80.0	B*44:03	absence	N/A	-	Y	15363036	ENSG00000183878	UTY	ENST00000331397	4884	ENSP00000328939	1293	Novel
NEVLIHSSQY	10	1	47.0	B*44:03	DEVLIHSSQY	109.0	rs61732383	5	145508635	ENSG00000133706	LARS	ENST00000394434	2841	ENSP00000377954	892	MHC Binder
QEAPESATVIF	11	1	283.0	B*44:03	EEAPESATVIF	160.0	rs2294689	6	24653272	ENSG00000111802	TDP2	ENST00000378198	916	ENSP00000367440	249	Novel
QEEQTRVAL	9	6	4725.0	B*44:03	QEEQTKVAL	3226	rs75139274	10	95278682	ENSG00000138180	CEP55	ENST00000371485	1347	ENSP00000360540	348	MHC Binder
QEFIDNPKW	9	5	27.0	B*44:03	QEFINNPKW	16	rs58447305	2	219526648	ENSG00000074582	BCS1L	ENST00000359273	765	ENSP00000352219	210	Novel
QELDGVFQKL	10	5	210.0	B*44:03	QELDRVFQKL	339	rs2241666	10	58118629	ENSG00000122952	ZWINT	ENST00000373944	598	ENSP00000363055	187	MiHA
QELDRVFQKL	10	5	339.0	B*44:03	QELDGVFQKL	210.0	rs2241666	10	58118629	ENSG00000122952	ZWINT	ENST00000373944	598	ENSP00000363055	187	MiHA
QELEEKLNIL	10	9	819.0	B*44:03	QELEEKLNML	780	rs60910145	22	36662033	ENSG00000100342	APOL1	ENST00000397278	1381	ENSP00000380448	384	Novel
QELETSIKKI	10	7	422.0	B*44:03	QELETSNKKI	409	rs12572012	10	91487648	ENSG00000138182	KIF20B	ENST00000416354	2339	ENSP00000411545	756	MiHA
QELIGKKEY	9	2	145.0	B*44:03	QDLIGKKEY	10291	rs34101857	14	45696036	ENSG00000129534	MIS18BP1	ENST00000310806	2208	ENSP00000309790	583	MiHa
QENIQNLQL	9	4	518.0	B*44:03	QEDIHNLQL	3813.0	rs3748693	1	214818222	ENSG00000117724	CENPF	ENST00000366955	5478	ENSP00000355922	1770	Novel
QENIQNLQL	9	4	518.0	B*44:03	QEDIQNLQL	291.0	rs3748693	1	214818222	ENSG00000117724	CENPF	ENST00000366955	5478	ENSP00000355922	1770	Novel
QENIQNLQL	9	4	518.0	B*44:03	QENIHNLQL	2465.0	rs3748693	1	214818222	ENSG00000117724	CENPF	ENST00000366955	5478	ENSP00000355922	1770	Novel
QENQDPRRW	9	7	21.0	B*44:03	QENQDPGRW	69	rs8285	12	49717532	ENSG00000135451	TROAP	ENST00000257909	218	ENSP00000257909	48	Novel
QEPFVFHEF	9	8	637.0	B*44:03	QEPFVFREF	656.0	rs201944488	19	58421127	ENSG00000173480	ZNF417	ENST00000312026	682	ENSP00000311319	173	Novel
QEYQVKLQA	9	1	1630.0	B*44:03	REYQVKLQA	2304.0	rs10838525	11	5701000	ENSG00000132256	TRIM5	ENST00000412903	833	ENSP00000388031	136	Novel
REVLELDSI	9	3	3731.0	B*44:03	REALELDSI	7223.0	rs664226	11	64789193	ENSG00000213465	ARL2	ENST00000246747	517	ENSP00000246747	141	Novel
REYQVKLQA	9	1	2304.0	B*44:03	QEYQVKLQA	1630.0	rs10838525	11	5701000	ENSG00000132256	TRIM5	ENST00000412903	833	ENSP00000388031	136	MiHA
RLAPTLSQL	9	1	60.0	A*02:01	QLAPTLSQL	191	rs4823054	22	30198024	ENSG00000100325	ASCC2	ENST00000307790	1640	ENSP00000305502	509	Novel
RLLQEQHQL	9	2	120.0	A*02:01	RVLQEQHQL	3429	rs61745599	17	78355493	ENSG00000173821	RNF213	ENST00000582970	14088	ENSP00000464087	4649	MHC Binder
RLQEELEKL	9	1	88.0	A*02:01	LLQEELEKL	58	rs1132274	20	17596154	ENSG00000125844	RRBP1	ENST00000360807	2961	ENSP00000354045	891	MHC Binder
RQPDLVLRL	9	2	668.0	A*02:01	REPDLVLRL	17633	rs2740348	17	649934	ENSG00000179409	GEMIN4	ENST00000319004	1467	ENSP00000321706	450	Novel
RSVDVTNTTFL	11	8	4489.0	A*02:01	RSVDVTNITFL	3519.0	rs10330	19	19017861	ENSG00000105669	COPE	ENST00000262812	399	ENSP00000262812	117	Novel
SEESAVPERSW	11	8	39.0	B*44:03	SEESAVPKRSW	30.0	rs2295283	1	12082925	ENSG00000116691	MIIP	ENST00000235332	668	ENSP00000235332	167	MiHA
SEESAVPKRSW	11	8	30.0	B*44:03	SEESAVPERSW	39	rs2295283	1	12082925	ENSG00000116691	MIIP	ENST00000235332	668	ENSP00000235332	167	MiHA
SEIEQKIKEY	10	1	20.0	B*44:03	AEIEQKIKEY	14	rs2073498	3	50369545	ENSG00000068028	RASSF1	ENST00000359365	504	ENSP00000352323	133	Novel
SESEDRLVA	9	9	4772.0	B*44:03	SESEDRLVG	7356	rs117236526	15	55664109	ENSG00000260916	CCPG1	ENST00000310958	886	ENSP00000311656	196	MiHA
SESKIRVLL	9	6	**918.0**	B*44:03	SESKICVLL	528	rs299284	5	162896649	ENSG00000072571	HMMR	ENST00000393915	420	ENSP00000377492	93	Novel
SLAPCKLETV	10	1	21.0	A*02:01	ILAPCKLETV	19	rs1129495	5	115428380	ENSG00000145781	COMMD10	ENST00000274458	445	ENSP00000274458	128	Novel
SLDDHVVAV	9	2	4.0	A*02:01	SPDDHVVAV	8249	rs4236176	6	170070722	ENSG00000184465	WDR27	ENST00000333572	918	ENSP00000330265	133	Novel
SLLESSRSQEL	11	11	74.0	A*02:01	SLLESSRSQEP	9627	rs2287174	12	110474069	ENSG00000076513	ANKRD13A	ENST00000261739	1680	ENSP00000261739	505	MHC Binder
SLQEKVAKA	9	6	73.0	A*02:01	SLQEKAAKA	295	rs299295	5	162909715	ENSG00000072571	HMMR	ENST00000393915	1597	ENSP00000377492	485	Novel
SQAEIEQKI	9	3	2800.0	A*02:01	SQSEIEQKI	5001.0	rs2073498	3	50369545	ENSG00000068028	RASSF1	ENST00000359365	504	ENSP00000352323	133	MiHa
TEGDALDALGLKRY	14	4	72.0	B*44:03	TEGEALDALGLKRY	109.0	rs4895	11	840476	ENSG00000177700	POLR2L	ENST00000534030	120	ENSP00000432807	33	Novel
TEMEISRAA	9	6	235.0	B*44:03	TEMEIPRAA	420.0	rs12702	15	44093926	ENSG00000242028	HYPK	ENST00000406925	4424	ENSP00000384474	105	Novel
TEMEISRAA	9	6	235.0	B*44:03	TEMEIPRAA	420.0	rs12702	15	44093926	ENSG00000140264	SERF2	ENST00000594896	566	ENSP00000471672	151	Novel
TETQEKNTL	9	5	4556.0	B*44:03	TETQDKNTL	5052	rs7148	2	85133215	ENSG00000034510	TMSB10	ENST00000233143	184	ENSP00000233143	25	Novel
TEVDEAGSQL	10	4	937.0	B*44:03	TEVGEAGSQL	1260	rs1135216	6	32814974	ENSG00000168394	TAP1	ENST00000354258	2252	ENSP00000346206	697	MHC Binder
TLLEDGTFKV	10	1	10.0	A*02:01	ILLEDGTFKV	6	rs11557236	16	4519438	ENSG00000153406	NMRAL1	ENST00000574733	798	ENSP00000458762	23	MHC Binder
TLSPEIITV	9	6	12.0	A*02:01	TLSPEMITV	16	rs12828016	12	998364	ENSG00000060237	WNK1	ENST00000535572	5656	ENSP00000441972	1561	MiHA
VEEADGNKQW	10	7	149.0	B*44:03	VEEADGHKQW	170	rs1131857	3	98307695	ENSG00000080819	CPOX	ENST00000264193	1033	ENSP00000264193	272	Novel
VEVPEAHQL	9	1	691.0	B*44:03	absence	N/A	-	Y	21871519	ENSG00000012817	KDM5D	ENST00000541639	3078	ENSP00000444293	931	Novel
VIAEILRGV	9	5	68.0	A*02:01	VIAEVLRGV	72	rs2273137	20	2635211	ENSG00000101361	NOP56	ENST00000329276	877	ENSP00000370589	121	MHC Binder
VLQNVAFSV	9	4	10.0	A*02:01	VLQKVAFSV	13	rs1138358	15	80263344	ENSG00000140379	BCL2A1	ENST00000267953	444	ENSP00000267953	39	Novel
YTWEEVFRV	9	7	3.0	A*02:01	YTWEEVCRV	7	rs1131293	2	242046784	ENSG00000115687	PASK	ENST00000405260	4496	ENSP00000384016	1266	MiHA