JCI The Journal of Clinical Investigation

I’ve got algorithm: predicting tumor and autoimmune peptide targets for CD8⁺ T cells

Devin Dersh, Jonathan W. Yewdell

J Clin Invest. 2016;126(12):4399-4401. https://doi.org/10.1172/JCI91302.

Commentary

$CD8^+$ T cells play a central role in eradicating intracellular pathogens, but also are important for noninfectious diseases, including cancer and autoimmunity. The ability to clinically manipulate $CD8^+$ T cells to target cancer and autoimmune disease is limited by our ignorance of relevant self-peptide target antigens. In this issue of the *JCI*, Pearson et al. describe 25,270 MHC class I–associated peptides presented by a wide range of HLA A and B allomorphs expressed by 18 different B cell lines. Via extensive bioinformatic analysis, the authors make surprising conclusions regarding the selective nature of peptide generation at the level of individual gene products and create a predictive algorithm for disease-relevant self-peptides that will be of immediate use for clinical and basic immunological research.

I’ve got algorithm: predicting tumor and autoimmune peptide targets for CD8$^+$ T cells

Devin Dersh and Jonathan W. Yewdell

Cellular Biology Section, Laboratory of Viral Diseases, National Institute of Allergy and Infectious Diseases (NIAID), NIH, Bethesda, Maryland, USA.

CD8$^+$ T cells play a central role in eradicating intracellular pathogens, but also are important for noninfectious diseases, including cancer and autoimmunity. The ability to clinically manipulate CD8$^+$ T cells to target cancer and autoimmune disease is limited by our ignorance of relevant self-peptide target antigens. In this issue of the *JCI*, Pearson et al. describe 25,270 MHC class I–associated peptides presented by a wide range of HLA A and B allomorphs expressed by 18 different B cell lines. Via extensive bioinformatic analysis, the authors make surprising conclusions regarding the selective nature of peptide generation at the level of individual gene products and create a predictive algorithm for disease-relevant self-peptides that will be of immediate use for clinical and basic immunological research.

MHC class I immunosurveillance

Jawed vertebrates evolved a remarkable system of immunosurveillance based on the ability of CD8$^+$ T cells to monitor gene expression at the level of individual cells. T cell activation is triggered in response to the clonally restricted T cell receptor (TCR) engaging with MHC class IA molecules (HLA A and B molecules in humans) that are bound to oligopeptides. Virtually all cells in the body constitutively present peptides on surface-expressed class I complexes. Through the process of thymic selection, T cell activation by self-peptides is minimized, focusing CD8$^+$ T cells on foreign peptides. Tolerance is imperfect, however, and self-peptide–reactive CD8$^+$ T cells are important for autoimmunity and cancer immunosurveillance; therefore, the ability to accurately predict peptide target antigens would represent a major clinical breakthrough.

MHC class I–associated peptides (MAPs) are typically generated through the action of proteasomes, which degrade polypeptides into shorter peptide fragments, thereby generating the COOH-terminal anchor residues that are critical for binding class I molecules. After peptide transport into the endoplasmic reticulum (ER) by a peptide transporter (TAP), the amino terminus is trimmed by ER aminopeptidases (ERAP1 or ERAP2) to create a high-affinity binding peptide. Most proteasome-generated peptides are recycled into amino acids, and only a tiny faction of these peptides are carried to the cell surface by class I molecules. The identity of the class I molecule is a critical factor in the determination of which peptides will reach the surface. Humans have thousands of HLA A and B alleles at appreciable population frequencies. Each individual can express up to four different HLA A and B alleles, with each allele binding to a distinct repertoire of peptides based on differences in the peptide-binding site. The odds of binding any given peptide of suitable length with physiological affinity ($K_d < 1$ μM) are approximately 1 in 100.

Currently, reasonably accurate algorithms for predicting whether a given peptide will bind a given class I allomorph (Immune Epitope Database and Analysis Resource, http://www.iedb.org) are available, but predicting the self-immunopeptidome — the repertoire of self-peptides presented by class I molecules — requires knowledge of which peptides are available for presentation. Peptides for immunosurveillance derive from two general sources: (a) so-called “retirees,” proteins degraded at the end of their natural life span, and (b) defective ribosomal products (DRiPs), translation products that are degraded during (1) or shortly after their synthesis due to a failure to attain a stable conformation as the result of errors in synthesis or folding or an inability to locate a stabilizing binding partner in a reasonable time frame (2). As even identical peptides can be presented with widely different efficiencies at a given rate of proteasomal degradation from similar or even ostensibly identical proteins (3), predicting the immunopeptidome from global ‘omes (transcriptome, exome, translatome, proteome, and even the degradome) is unlikely to be a simple matter.

Simple or not, detailed ‘omics characterization is essential to developing accurate predictive algorithms for the immunopeptidome. Fortunately, there have been quantum leaps in the power of all ‘omic technologies in the past few years. Most importantly, advances in mass spectrometry (4) have enabled the characterization of tens of thousands of class I peptide ligands (5–7), and coupled with kinetic analysis (8) or the use of isotopically labeled amino acids (9), these peptides can be assigned to derive either from DRiPs or retirees.

A predictive algorithm

The stage was therefore set for Pearson and colleagues to devise the first algorithm for predicting the potential of a target gene to generate peptides that are likely to bind a wide variety of HLA A or B molecules (10). Specifically, B cell lines were generated from 18 individuals that collectively express 27 HLA-A and HLA-B allomorphs common to individuals of European ancestry. Peptides were isolated from cell sur-

▶ Related Article: p. 4690

Conflict of interest: The authors have declared that no conflict of interest exists.
Reference information: *J Clin Invest.* 2016;126(12):4399–4401. doi:10.1172/JCI91302.

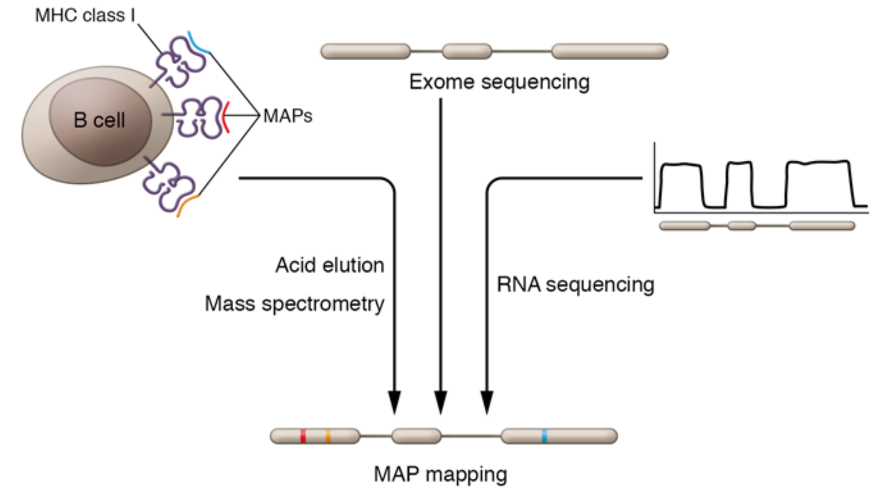

Figure 1. Summary of experimental approach and salient conclusions. As reported in this issue, Pearson et al. isolated B cells from 18 individuals, expressing a total of 27 MHC class I allomorphs, and identified surface-presented peptides using mild acid elution coupled with mass spectrometry. Simultaneously, personalized genetic databases were created using transcriptome and exome sequencing from each donor. Together, the peptide identification and sequencing information allowed for the mapping of MAPs at their source locations within the genome.

face class I molecules on live cells by acid elution, and mass spectrometry was used in conjunction with transcriptome and exome databases created for each cell line to identify over 25,000 peptides with high predictive binding scores for the HLA-A and HLA-B allomorphs expressed by each individual (Figure 1). This is the largest set of MAPs identified to date and the first to compare peptides from such a large collection of HLA allomorphs.

Pearson et al. discovered that peptides are not randomly generated across the exome (10). Remarkably, 41% of the 10,000+ genes expressed in the cell lines provide no detectable peptides. In contrast, approximately 40% of genes generated more than one peptide, with some genes sourcing as many as 64 MAPs. Altogether, the 25,000+ peptides derived from just 10% of the total exome divided into 25 amino acid blocks (if peptides were randomly distributed, we calculate that they would be distributed in approximately 20% of such windows, given the size of the exome). This bias is not related to the presence of predicted high-affinity peptides, which were randomly distributed in the exome, as expected (11). Remarkably, fully 20% of MAPs were determined to be part of a set of overlapping peptides, which typically bind different class I allomorphs (10). Because overlapping peptides will bind to different class I allomorphs using different anchor residues, the presence of MAP hot spots implies that there is preferential access of these regions to the class I processing pathway.

Pearson and colleagues analyzed multiple features of source compared with nonsource gene sets (10) to better understand the very curious bias in MAP localization in the translatome. RNA transcript levels and the corresponding source proteins were significantly higher on average for MAP source genes compared with nonsource genes. However, other factors clearly influenced MAP generation (summarized in Table 1). Source genes were biased toward longer transcripts with more exons, smaller 5′ UTRs, and few upstream open reading frames (uORFs). The greater number of exons supports previous findings that pioneer the concept that translation associated with nonsense-mediated decay (NMD) is a significant source of MAPs (12). The bias toward upstream mRNA simplicity is consistent with efficient translation boosting MAP generation.

At the polypeptide level, MAP source proteins are strongly biased toward proteins in macromolecular complexes, an observation that is consistent with DRiP generation from unassembled subunits (9), and nuclear targeted proteins, which is consistent with peptide generation by nuclear proteasomes (13) and against ER-targeted proteins, perhaps pointing to the exclusion of these proteins from the retiree pool by virtue of being extracellular or within endolysosomal compartments. MAP source proteins are on average larger, more disordered, and enriched in β sheets and known degradation signals.

Most importantly, Pearson et al. used their detailed analysis to create a model to predict potential target MAPs from exomic data (10). The model was effective not only on the training data set, but also in predicting MAPs previously reported by other mass spectroscopy studies, and should be immediately useful in identifying MAPs that are important for cancer immunology and autoimmunity.

Discussion

A major advance of the study by Pearson et al. is the sheer quantity of data generated from multiple HLA allomorphs and depth of informatics analysis, which enables statistically powerful conclusions regarding the origins of MAPs (10). The most intriguing finding is the presence of MAP hot spots in the genome, many of which promiscuously provide peptides to multiple HLA-A and HLA-B allomorphs with divergent peptide-binding motifs. The detailed analysis implies that both the chemical nature of the translation products and the

Table 1. Key features of MAP source and nonsource genes identified with bioinformatics

Feature	MAP source genes	Nonsource genes
Potential HLA-binding peptides	Equal	Equal
Transcript level	Higher	Lower
Protein level	Higher	Lower
Number of exons	More	Fewer
Length	Longer	Shorter
uORFs	Fewer	More
Polypeptide degradation motifs	More	Fewer
Predicted polypeptide disorder	Higher	Lower

act of translation itself contribute to the positional bias of MAPs in the exome.

As proteolytic generation of peptide termini is a critical feature in peptide antigenicity, inclusion of proteasome and ERAP predictive algorithms (14, 15) should enhance the accuracy of the predictive algorithm developed by Pearson et al. (10). Future studies should focus on the addition of ribosome profiling (16) to the analysis, which will provide a number of important parameters, including the density of actively translating ribosomes on mRNA, rates of initiation and elongation, and pausing. It will also be important to query the mass spectrometry data against all of the possible coding information in cells (17). This includes nonspliced RNA, as MAPs can also derive from introns (18, 19), +1 and +2 reading frames, and stop codon skipping as well as from proteasome-mediated splicing of degradation intermediates (20, 21). Remarkably, proteasome-spliced peptides appear to constitute perhaps 30% of immunopeptidome and may bind to class I molecules by different rules (22).

Importantly, the finding by Pearson et al. can be extended to cells altered by stress, infection, and oncogenic transformation and will likely provide insight into flexibility regarding the sources of MAPs and pathways used to generate MAPs under pathogenic conditions. Tumor cells are particularly important, given the exciting recent progress in immunotherapy and the paucity of defined target MAPs. Although the algorithm designed by Pearson and colleagues had predictive value for the five human cancer cell lines tested (10), the algorithm is likely to be improved by detailed analysis of cancer cells, including those present in biopsied tumors.

Parallel studies on MHC class II–associated peptides, which are also important targets for CD4$^+$ (and potentially even CD8$^+$; ref. 23) T cell immunosurveillance, will be interesting and important to perform. As most class II peptides are generated by endolysosomal processing of retirees, this should lead to a different and telling bias among the various bioinformatic parameters compared with class I peptides, and if not, what gives? Further, it will be interesting, and perhaps even useful, to specifically study HLA-E peptides, given recent findings that suggest a potentially broad role for HLA-E in T cell immunosurveillance (24).

The road to improving MAP-predictive algorithms also entails increasing the sensitivity of peptide detection of the mass spectrometric analysis to ultimately reach detection of a single peptide per cell (or less). Ironically, due to immune tolerance, in conjunction with the remarkable sensitivity of CD8$^+$ T cells (easily less than 10 complexes per cell for the most sensitive clones; ref. 25), cancer cell immunotherapy target peptides are likely to be biased toward low copy number peptides. It is therefore crucial to extend analysis to low abundance peptides. Regardless of these limitations, Pearson et al.'s algorithm represents a milestone in predicting the immunopeptidome and will have an immediate impact on a range of human diseases and, crucially, is sure to spur experimental and bioinformatic research to create ever better algorithms.

Acknowledgments

The authors are supported by the Division of Intramural Research, NIAID.

Address correspondence to: Jonathan W. Yewdell, Room 2E13.1C, Bldg. 33, 33 North Drive, NIH, Bethesda, Maryland 20892, USA. Phone: 301.402.4602; E-mail: jyewdell@NIH.gov.

1. Wang F, Durfee LA, Huibregtse JM. A cotranslational ubiquitination pathway for quality control of misfolded proteins. *Mol Cell*. 2013;50(3):368–378.
2. Antón LC, Yewdell JW. Translating DRiPs: MHC class I immunosurveillance of pathogens and tumors. *J Leukoc Biol*. 2014;95(4):551–562.
3. Dolan BP, Sharma AA, Gibbs JS, Cunningham TJ, Bennink JR, Yewdell JW. MHC class I antigen processing distinguishes endogenous antigens based on their translation from cellular vs. viral mRNA. *Proc Natl Acad Sci U S A*. 2012;109(18):7025–7030.
4. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of Major Histocompatibility Complex (MHC) immunopeptidomes using mass spectrometry. *Mol Cell Proteomics*. 2015;14(12):3105–3117.
5. Hassan C, et al. The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol Cell Proteomics*. 2013;12(7):1829–1843.
6. Mommen GP, et al. Expanding the detectable HLA peptide repertoire using electron-transfer/higher-energy collision dissociation (EThcD). *Proc Natl Acad Sci U S A*. 2014;111(12):4507–4512.
7. 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.
8. Croft NP, et al. Kinetics of antigen expression and epitope presentation during virus infection. *PLoS Pathog*. 2013;9(1):e1003129.
9. Bourdetsky D, Schmelzer CE, Admon A. The nature and extent of contributions by defective ribosome products to the HLA peptidome. *Proc Natl Acad Sci U S A*. 2014;111(16):E1591–E1599.
10. Pearson H, et al. MHC class I–associated peptides derive from selective regions of the human genome. *J Clin Invest*. 2016;126(12):4690–4701.
11. Istrail S, et al. Comparative immunopeptidomics of humans and their pathogens. *Proc Natl Acad Sci U S A*. 2004;101(36):13268–13272.
12. Apcher S, et al. Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *Proc Natl Acad Sci U S A*. 2011;108(28):11572–11577.
13. Antón LC, et al. Intracellular localization of proteasomal degradation of a viral antigen. *J Cell Biol*. 1999;146(1):113–124.
14. Singh SP, Mishra BN. Major histocompatibility complex linked databases and prediction tools for designing vaccines. *Hum Immunol*. 2016;77(3):295–306.
15. Guasp P, et al. The peptidome of Behcet's disease-associated HLA-B*51:01 includes two subpeptidomes differentially shaped by endoplasmic reticulum aminopeptidase 1. *Arthritis Rheumatol*. 2016;68(2):505–515.
16. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet*. 2014;15(3):205–213.
17. Laumont CM, et al. Global proteogenomic analysis of human MHC class I-associated peptides derived from non-canonical reading frames. *Nat Commun*. 2016;7:10238.
18. Robbins PF, El-Gamil M, Li YF, Fitzgerald EB, Kawakami Y, Rosenberg SA. The intronic region of an incompletely spliced gp100 gene transcript encodes an epitope recognized by melanoma-reactive tumor-infiltrating lymphocytes. *J Immunol*. 1997;159(1):303–308.
19. Apcher S, Millot G, Daskalogianni C, Scherl A, Manoury B, Fåhraeus R. Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. *Proc Natl Acad Sci U S A*. 2013;110(44):17951–17956.
20. Hanada K, Yewdell JW, Yang JC. Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature*. 2004;427(6971):252–256.
21. Vigneron N, et al. An antigenic peptide produced by peptide splicing in the proteasome. *Science*. 2004;304(5670):587–590.
22. Liepe J, et al. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science*. 2016;354(6310):354–358.
23. Hansen SG, et al. Cytomegalovirus vectors violate CD8$^+$ T cell epitope recognition paradigms. *Science*. 2013;340(6135):1237874.
24. Hansen SG, et al. Broadly targeted CD8$^+$ T cell responses restricted by major histocompatibility complex E. *Science*. 2016;351(6274):714–720.
25. Sykulev Y, Joo M, Vturina I, Tsomides TJ, Eisen HN. Evidence that a single peptide-MHC complex on a target cell can elicit a cytolytic T cell response. *Immunity*. 1996;4(6):565–571.