Next-generation epitope prediction using mass spectrometry and integrative genomics

POSTER LB-179

Michael S. Rooney[1,8,9], Jennifer G. Abelin[1] , Derin B. Keskin[1,3,4,6] , Siranush Sarkizova[1,2], Christina Hartigan[1], Wandi Zhang[3], John Sidney[7], Jonathan Stevens[5], William Lane[5], Guang Lan Zhang[3,6,10], Karl R. Clauser[1], Nir Hacohen[1,3,11], Steven A. Carr[1], Catherine J. Wu[1,3,4,6]

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA; [3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA; [4]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA; [5]Tissue Typing Laboratory, Brigham and Women's Hospital, Boston, MA, USA; [6]Harvard Medical School, Boston, MA, USA; [7]La Jolla Institute for Allergy and Immunology, La Jolla, CA; [8]Harvard/MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA; [9]Neon Therapeutics, Cambridge, MA, USA; [10]Department of Computer Science, Metropolitan College, Boston University, Boston, MA, USA; [11]Center for Cancer Immunology, Massachusetts General Hospital

Abstract

Neoantigens are somatically mutated protein sequences presented on MHC Class I or II molecules. Mounting evidence suggests neoantigens potentiate anti-tumor immune responses during checkpoint blockade, and there is great interest in directly targeted neoantigens therapeutically with vaccine. Such vaccines must rely on personalized *in silico* epitope prediction given the multitude of MHC alleles in the human population (each with its own peptide binding repertoire) and the very small degree of somatic mutation overlap amongst cancer patients. The current prediction paradigm is based almost entirely on machine learning algorithms (*e.g.* NetMHC) trained on *in vitro* p:MHC binding affinity assays.

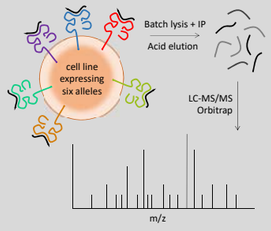
Mass spectrometry (MS), which can be used to directly identify endogenously processed and presented peptides, is an orthogonal approach to define the rules of the MHC ligandome. To date, MS studies have identified large numbers of MHC-bound peptides and produced general observations, such as bias toward highly expressed proteins, but have not yielded new predictors. The main bottleneck to prediction has been the use of cell lines expressing the natural complement of **six distinct MHC Class I molecules** (2 HLA-A alleles, 2 HLA-B alleles, and 2 HLA-C alleles). In this context, peptides can only be assigned to one of the six alleles using pre-existing predictors; many peptides cannot be confidently assigned to any allele. This convolution drastically reduces the amount of forward learning that can be achieved.

Here we modify the MS approach to focus on an **MHC-null cell line** transfected with a **single allele** of our choice. We iteratively applied rapid, high-resolution liquid chromatography mass spectrometry (LC-MS/MS) to 16 HLA alleles identifying >24,000 MHC-bound peptides. Using these data, we were able to learn peptide-binding and proteasomal cleavage motifs *de novo*. In addition, we uncovered new motifs absent in the public database IEDB (which would have been impossible using the multi-allelic MS approach) and validated these with traditional *in vitro* binding assays. Furthermore, we systematically assess variables hypothesized to predict presentation. Notably, we find that transcript expression is highly predictive of presentation and that source protein localization makes minimal incremental contribution.

Finally, we trained neural network models on our endogenous peptide dataset to produce integrated predictors that leverage both peptide sequence and auxiliary variables such as gene expression and the likelihood of efficient proteasomal processing. We demonstrate effective prediction in MS data sets collected by independent groups, achieving 2-3x more correct identifications than NetMHC or NetMHCPan. We also show compelling performance in predicting immunogenic HIV epitopes. We conclude that measuring mono-allelic MHC peptides using LC-MS/MS improves the utility of predictive algorithms and provides a rapid and scalable method to generate rules for neoantigen prediction.

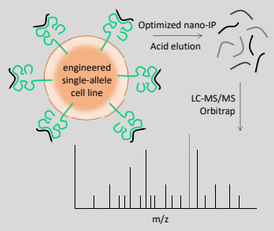
Traditional MS approach

- Cell lines express six different class I alleles
- Must rely on known motifs to assign peptides to alleles



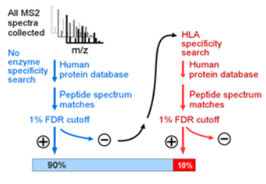
Our approach

- Class I-null B721.221 cells are transfected with single allele
- No ambiguity in allele assignment; higher depth per allele



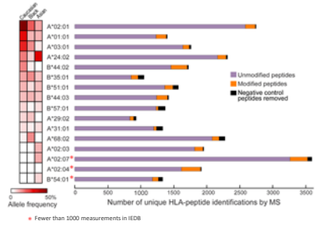
Iterative spectral search

Boosts identifications by 5-40%



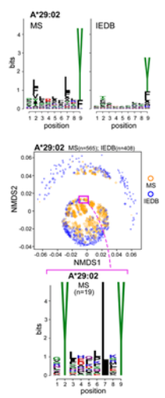
16 alleles, 24,000 peptides

Compare to total size of IEDB – 97,000 peptides

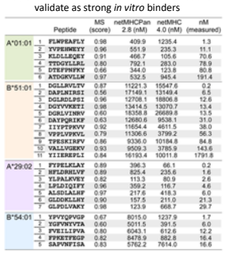


Single-allele system identifies novel motifs that are not discoverable using traditional multi-allelic approach

Isolation of a motif absent in IEDB

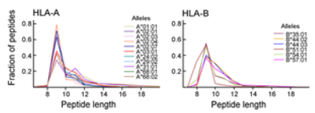


MS peptides with poor NetMHCPan predictions validate as strong *in vitro* binders



	Peptide	MS (score)	netMHCPan 2.8 (nM)	netMHC 4.0 (nM)	nM (measured)
A*01:01	FLWPEAFLY	0.98	409.9	1235.4	1.3
	YVPEMMEYV	0.96	551.9	235.3	11.1
	KLDLLEQEY	0.91	466.7	105.6	70.6
	TTDGTLLEL	0.80	792.1	283.0	78.9
	DTEFPNFKY	0.66	344.0	123.8	80.8
	ATDGRVVLW	0.97	532.5	945.4	191.4
B*51:01	DGLLAVLTV	0.87	11221.3	15547.6	0.2
	DAPLHIRSI	0.56	17149.1	13149.4	6.5
	DGLRDLPSI	0.96	12708.1	18806.8	12.6
	DGYYVKETI	0.98	13414.5	13070.7	13.4
	DGRLYINRV	0.60	18358.8	26889.8	13.5
	DAYPGRIKF	0.63	12680.6	9538.1	31.0
	IIYPTPKYV	0.92	11654.4	4611.5	38.0
	VVPSVMKYL	0.79	11306.6	3799.2	56.3
	TPESKIRFV	0.86	9336.0	10184.8	84.8
	YALLVGEKV	0.93	9509.3	3785.9	143.6
	YIIERERPLI	0.84	16193.4	10011.8	1791.8
A*29:02	FYPELKLAY	0.89	396.3	66.1	0.2
	NFLDRNLVP	0.89	825.4	235.6	1.6
	YLPALAVEY	0.82	113.3	80.9	2.6
	LPLDIQIFY	0.96	359.2	116.7	4.6
	ALQLALAHF	0.97	217.6	418.3	6.0
	GLDDKLLWY	0.90	157.5	211.0	21.3
	GLPSLVAKY	0.98	123.9	668.7	29.7
B*54:01	YPVYQPVGP	0.67	8015.0	1237.9	1.7
	YGFVNTVTA	0.60	5011.5	391.5	6.0
	IYRILIPVA	0.80	6043.1	612.6	12.2
	FPKETFEGP	0.82	8478.9	882.8	16.4
	SAPVWFIAA	0.83	5762.2	7614.0	16.6

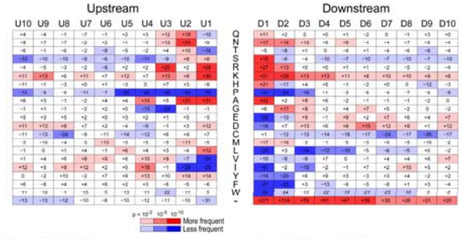
Direct observation of allele-specific length distributions

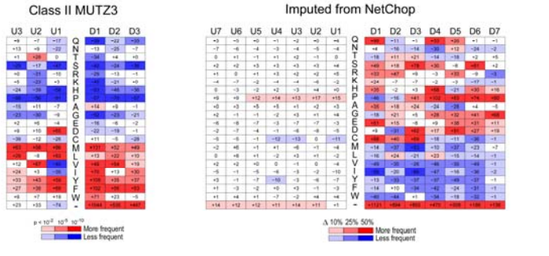


Quantitative analysis of *in vivo*-processed peptide shows tryptic signature, preference for alanine, and strong proline avoidance

Below:

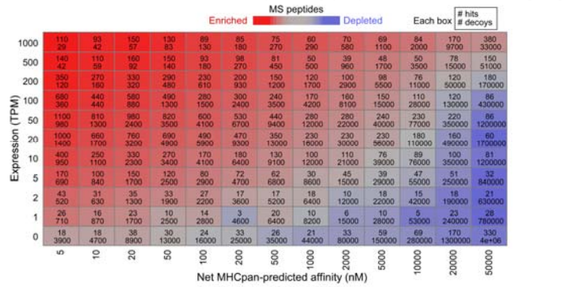
- The upstream and downstream sequence of each MS binder was compared to decoy sequences
- Heatmap shows percent change in frequency of given amino acid at given position when comparing binders to decoys
- Signature is preserved in breast, colorectal, and fibroblast cell lines as well as HeLa and PBMC.





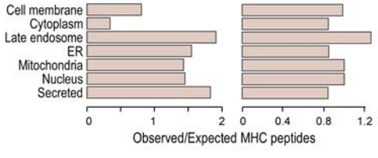
MHC peptides reveal multiplicative relationship between gene expression and binding affinity

- MS peptides and random genomic 9mer decoys were binned according expression (RNA-Seq) and predicted affinity
- Binder:decoy ratio per bin shows that a 10x increase in expression roughly compensates for a 90% reduction in binding strength

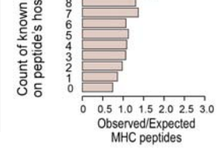


Most cellular localization bias can be explained by expression

After accounting for expression level, localization of source protein does not predict peptide presentation



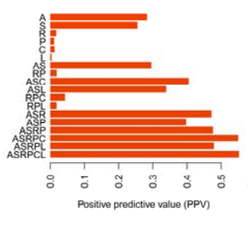
The count of known ubiquitination sites in the source protein is associated with enhanced peptide presentation



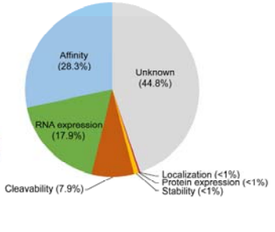
Single-allele data enables development of data-driven integrated predictors; expression and cleavability greatly increase accuracy

- Logistic regression models were developed to discriminate *n* MS-observed peptides from 999*n* genomic 9mer decoys
- Each model included a different subset of predictors, *e.g.* IsBinder ~ β_0 + β_1NetMHCScore + β_2Expression
- Model performance was assessed according to the fraction of true MS peptides among the most highly ranked peptides (top 0.1%); this score is called the positive predictive value (PPV)

Positive predictive value (PPV) of model using given subset of predictor variables

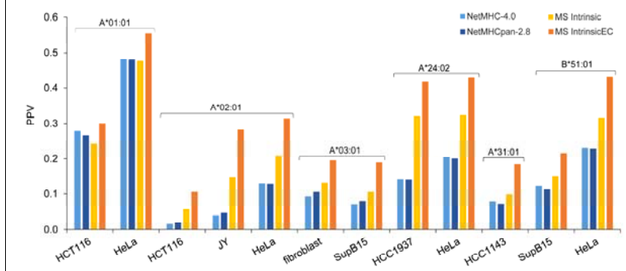


Stepwise addition of variables reveals the following explanatory contributions (averaged across all alleles)

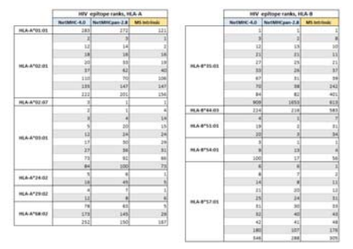


Novel predictors based on our single-allele MS data outperform NetMHC on external test data

- Bassani-Sternberg *et al* (2015) identified MHC-bound peptides for six human cell lines using mass spectrometry
- We evaluated our algorithms by seeing whether we could distinguish these peptides from decoys (999 decoys per true peptide)
- Our MS-based predictor bests NetMHC for almost every allele we could evaluate
- Adding expression and cleavage to the model ("+ EC") improves prediction further – **2-3x** the performance of NetMHC



- The Los Alamos 2013 HIV-1 CTL Epitope Update lists 52 HIV-1 epitopes associated with alleles included in our study
- We used our predictor to rank these relative to the remaining 9mers in the HIV-1 genome
- The MS predictor was almost always able to nominate true epitopes to higher rank than NetMHC-4.0 or NetMHCPan-2.8



Conclusions and Next Steps

- Single-allele system generates high coverage and enables deep integrative analysis by removing the problem of allele ambiguity
- MS data can be used to learn peptide motifs and cleavage motifs *de novo* and to systematically probe how various factors (protein localization, expression, *etc.*) impact epitope presentation
- These findings will yield new predictors that will be critical to the success of personalized neoantigen approaches
- Method can be rapidly deployed to expand allele coverage for Class I and can be adapted for Class II MHC predictions

Acknowledgments

We thank Wilfredo F. Garcia-Beltran of the Ragon Institute of MGH, MIT and Harvard for providing the cell lines. The work was supported by the Blavatnik Family Foundation (N.H. and C.W.), the NCI R01 CA155010 (C.W.), NCI CPTAC U24 CA160034 (S.A.C, and K.C.), NHGRI T32 HG002295 (S.S.), and Dana-Farber/Harvard Cancer Center Kidney Cancer SPORE 2P50CA101942-11A1 (D.B.K.). C.W. is a Scholar of the Leukemia and Lymphoma Society.