Supplementary Information for

Beyond MHC binding: immunogenicity prediction tools to refine neoantigen selection in cancer patients

Ibel Carri[1,2], Erika Schwab[3], Enrique Podaza[4], Heli M. Garcia Alvarez[1,2], José Mordoh[3,5,6], Morten Nielsen[1,2,7], María Marcela Barrio[3*]

[1]Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín (UNSAM)—Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires B1650HMP, Argentina

[2]Escuela de Bio y Nanotecnologías (EByN), Universidad Nacional de San Martín, Buenos Aires B1650HMP, Argentina

[3]Centro de Investigaciones Oncológicas, Fundación Cáncer, Ciudad Autónoma de Buenos Aires C1426ANZ, Argentina

[4]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA

[5]Instituto Alexander Fleming, Ciudad Autónoma de Buenos Aires C1426ANZ, Argentina

[6]Laboratory of Cancerology, Fundación Instituto Leloir, Ciudad Autónoma de Buenos Aires C1405BWE, Argentina

[7]Section of Bioinformatics, Department of Health Technology, Technical University of Denmark, 2800 Lyngby, Denmark

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Supplementary methods

In-house neoepitope dataset

Three melanoma patients that received VACCIMEL [1] were selected for this analysis. Their samples were obtained following the protocol described in [2]. Tumor samples were processed to perform Whole-exome Sequencing (WES), RNA sequencing (RNAseq), and Human Leukocyte Antigens (HLA) typing as described in [3]. To identify somatic variants, MuTect2 [4] was used following Genome Analysis Toolkit (GATK) best practices. Neopeptides were obtained with mutant peptide extractor and informer (MuPeXi) [5] and selected considering a predicted rank score of binding affinity to Major Histocompatibility Complex (MHC) $\leq$ 2 by using NetMHCpan 4.0 EL [6] and corresponding wild-type > 2. From all the candidates obtained with this pipeline, a group of peptides was manually selected to synthesize and assess the immune response. For promiscuous neopeptides, the allele with better predicted binding affinity was selected for further analysis. Neopeptide source mutated proteins were obtained with SeqTailor [7] and manually curated. Quantification of transcript expression from RNAseq data was obtained with Kallisto [8].

Neoepitope immunogenicity assessment

The interferon gamma (IFNγ) enzyme-linked immunospot (ELISPOT) Assay for the predicted neopeptides was performed with peripheral blood mononuclear cells (PBMC) at three time points after vaccination with VACCIMEL (P1= 6 months, P2= 1 yr, P3= 2yr) as previously described [3]. The background baseline for each patient and time point was calculated as the average number of spots present in non-stimulated cells. Quantitative values of immune response were derived from the ratio between the number of spots and the corresponding background baseline. The maximum value from any time point was considered, and neopeptides were labeled as positive or immunogenic if the number of spots is 2,5 times higher than baseline.

Amino acid enrichment analysis

To analyze the amino acid composition of immunogenic peptide datasets from different sources, we downloaded neopeptides from Neoepitope Database (NEPdb) [9] and Cancer Epitope Database and Analysis Resource (CEDAR) [10] (December 2022) and viral peptides from Immune Epitope Database (IEDB) (August 2021) that were experimentally evaluated for T cell responses. For viral peptides, the query included linear peptides that bind to MHC class I, with human as host organism, and virus as source organism. Entries that have MHCs with low resolution or not included in NetMHCpan 4.1 were discarded, as well as entries of peptides with non-conventional amino acids. In the case of neopeptides from NEPdb, we selected only 8 to 10 mers. Further, we predicted the likelihood of binding to the corresponding MHC using NetMHCpan, excluded peptides with a rank score higher than 2 (predicted non-binders),

and selected the predicted Icore as the minimal peptidic sequence that can form a pMHC-TCR complex. From this sequence, the first and last 3 amino acids containing anchor positions were discarded. In this way, the central region of the peptide, which is associated with the interaction with the T cell receptor (TCR), was conserved for further analysis. The enrichment (*e*) of each amino acid (*aa*) was calculated using the following formula

$$e_{aa} = log_2\big((P_{aa} + 0.01) \div (N_{aa} + 0.01)\big)$$

Where *P* is the proportion of the amino acid (aa) among the selected region in immunogenic neopeptides and *N* in non-immunogenic neopeptides.

Methods used to predict immunogenicity in-house dataset

ProteaSMM [11] predictions were retrieved via IEDB API.

Only the predictions of transporter associated with antigen-processing (TAP) binding affinity and proteasome cleavage from NetCTLpan [12] were considered for this study. NetCTLpan was downloaded and executed locally following the author's recommendations.

NetMHCstabpan [13] predictions considered for this study do not include affinity predictions (-ia 0). NetMHCstabpan was downloaded and executed locally following the author's recommendations.

NetMHCpanexp [14] was downloaded and executed locally following the author's recommendations.

Expression data to perform HLAthena [15] predictions were obtained with NetMHCpanexp. HLAthena was executed from docker containers following the author's recommendations.

Improved Proteasome Cleavage Prediction Server (iPCPS) [16] models used in this review were selected by best specificity (immunoproteasome model 1 and proteasome model 2). iPCPS predictions were obtained from the web server.

Kernel similarity was calculated as described in [17] using a block of amino acid substitution matrix (BLOSUM) 62 matrix. This metric was also applied to peptides with removed anchor positions. Anchors were defined as the positions with the highest information content in the motifs of each MHC molecule preference.

Pairwise sequence similarity was calculated as described in [18].

DeepNetBim [19] only accepts peptides of 9 amino acids. The predicted binding core with NetMHCpan 4.0 [6] was used to analyze all peptides included in the dataset. In 8-

meres, the predicted position of insertion was replaced with X. DeepNetBim was downloaded and executed following the author's recommendations.

DeepImmuno [20] only accepts peptides of 9 and 10 amino acids. The predicted binding core with NetMHCpan 4.0 [6] was used to analyze all peptides included in the dataset. In 8-meres, the predicted position of insertion was replaced with alanine, and in 11meres, 1 amino acid was deleted at the deletion predicted position. DeepImmuno predictions were obtained from the web server.

The neoantigen immunogenicity prediction model of immunogenic epitope/neoepitope prediction (INeo-Epp) [21] only supports single amino acid mutations. To analyze peptides originating in frameshift variants, we modified the wild-type peptide input sequence to be equal to the mutated peptide, except in the amino acid nearest to the anchor position, which was conserved as in the wild-type. (i.e. Mutant peptide: EADLRVQSL, Wild-type peptide: EA**T**LR**T**QSL, Wild-type sequence used as input to the program: EA**T**LRVQSL)

IEDB immunogenicity [22], Predictor of Immunogenic Epitopes (PRIME) [23], NetCleave [24], NetMHCpan and MHCflurry [25] were downloaded and executed locally following the author's recommendations.

Antigen.garnish [26], and DeepHLApan [27] were executed from docker containers following the author's recommendations.

Tumor Antigen predictor (TA predictor) [28] and identification of Tumor T cell Antigens-Random Forest iTTCA-RF [29] predictions were obtained from the corresponding web servers.

Variant allele frequency was obtained with MuTect2 [4].

Comparative evaluation of predictive methods

Evaluation metrics were calculated using the scikit-learn package [30] in Python 3.8.10. Statistical and correlation analysis were performed with SciPy 1.6.3.

References

1. Mordoh A, Aris M, Carri I, Bravo AI, Podaza E, Pardo JCT, et al. An Update of Cutaneous Melanoma Patients Treated in Adjuvancy With the Allogeneic Melanoma Vaccine VACCIMEL and Presentation of a Selected Case Report With In-Transit Metastases. Frontiers in immunology. 2022;13:842555-842555.

2. Mordoh J, Pampena MB, Aris M, Blanco PA, Lombardo M, Von Euw E, et al. Phase II Study of Adjuvant Immunotherapy with the CSF-470 Vaccine Plus Bacillus

Calmette–Guerin Plus Recombinant Human Granulocyte Macrophage-Colony Stimulating Factor vs Medium-Dose Interferon Alpha 2B in Stages IIB, IIC, and III Cutaneous Melanoma Patients: A Single Institution, Randomized Study. Frontiers in Immunology. 2017;8:625.

3. Podaza E, Carri I, Aris M, Von Euw E, Bravo AI, Blanco P, et al. Evaluation of T-Cell responses against shared melanoma associated antigens and predicted neoantigens in cutaneous melanoma patients treated with the CSF-470 allogeneic cell vaccine plus BCG and GM-CSF. Frontiers in immunology. 2020;11:1147.

4. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling somatic SNVs and indels with Mutect2. BioRxiv. 2019;861054.

5. Bjerregaard AM, Nielsen M, Hadrup SR, Szallasi Z, Eklund AC. MuPeXI: prediction of neo-epitopes from tumor sequencing data. Cancer Immunology, Immunotherapy. 2017;66(9):1123-1130.

6. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. The Journal of Immunology. 2017;199(9):3360-3368.

7. Zhang P, Boisson B, Stenson PD, Cooper DN, Casanova JL, Abel L, et al. SeqTailor: a user-friendly webserver for the extraction of DNA or protein sequences from next-generation sequencing data. Nucleic acids research. 2019;47(W1):W623-W631.

8. Bray NL, Pimentel H, Melsted P, Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology. 2016;34(5):525-527.

9. Xia J, Bai P, Fan W, Li Q, Li Y, Wang D, et al. NEPdb: a database of T-cell experimentally-validated neoantigens and pan-cancer predicted neoepitopes for cancer immunotherapy. Frontiers in Immunology. 2021;12:644637.

10. Koşaloğlu-Yalçın ZK, Blazeska N, Vita R, Carter H, Nielsen M, Schoenberger S, et al. The Cancer Epitope Database and Analysis Resource (CEDAR). *Nucleic Acids Research*. 2022;1.

11. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, et al. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. Cellular and Molecular Life Sciences CMLS. 2005;62(9):1025-1037.

12. Stranzl T, Larsen MV, Lundegaard C, Nielsen M. NetCTLpan: pan-specific MHC class I pathway epitope predictions. Immunogenetics. 2010;62(6):357-368.

13. Rasmussen M, Fenoy E, Harndahl M, Kristensen AB, Nielsen IK, Nielsen M, et al.

Pan-specific prediction of peptide–MHC class I complex stability, a correlate of T cell immunogenicity. The Journal of Immunology. 2016;197(4):1517-1524.

14. Alvarez HMG, Koşaloğlu-Yalçın Z, Peters B, Nielsen M. The role of antigen expression in shaping the repertoire of HLA presented ligands. Iscience. 2022;104975.

15. Sarkizova S, Klaeger S, Le PM, Li LW, Oliveira G, Keshishian H, et al. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nature biotechnology. 2020;38(2):199-209.

16. Gomez-Perosanz M, Ras-Carmona A, Lafuente EM, Reche PA. Identification of CD8+ T cell epitopes through proteasome cleavage site predictions. BMC bioinformatics. 2020;21(17):1-11.

17. Shen WJ, Wong HS, Xiao QW, Guo X, Smale, S. Towards a mathematical foundation of immunology and amino acid chains. arXiv:1205.6031. [Preprint]. 2012

18. Zhou C, Wei Z, Zhang Z, Zhang B, Zhu C, Chen K, et al. pTuneos: prioritizing tumor neoantigens from next-generation sequencing data. Genome medicine. 2019;11(1):1-17.

19. Yang X, Zhao L, Wei F, Li J. DeepNetBim: deep learning model for predicting HLA-epitope interactions based on network analysis by harnessing binding and immunogenicity information. BMC bioinformatics. 2021;22(1):1-16.

20. Li G, Iyer B, Prasath VS, Ni Y, Salomonis N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. Briefings in Bioinformatics. 2021;22(6):1–10.

21. Wang G, Wan H, Jian X, Li Y, Ouyang J, Tan X, et al. INeo-Epp: a novel T-cell HLA class-I immunogenicity or neoantigenic epitope prediction method based on sequence-related amino acid features. BioMed research international. 2020.

22. Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A. Properties of MHC class I presented peptides that enhance immunogenicity. PLoS computational biology. 2013;9(10):e1003266.

23. Schmidt J, Smith AR, Magnin M, Racle J, Devlin JR, Bobisse S, et al. Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. Cell Reports Medicine. 2021;2(2):100194.

24. Amengual-Rigo P, Guallar V. NetCleave: an open-source algorithm for predicting C-terminal antigen processing for MHC-I and MHC-II. Scientific reports. 2021;11(1):1-8.

25. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. Cell systems. 2020;11(1):42-48.

26. Richman LP, Vonderheide RH, Rech AJ. Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. Cell Systems. 2019;9(4):375-382.

27. Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, et al. DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. Frontiers in immunology. 2019;10:2559.

28. Herrera-Bravo J, Belén LH, Farias JG, Beltrán JF. TAP 1.0: a robust immunoinformatic tool for the prediction of tumor T-cell antigens based on AAindex properties. Computational Biology and Chemistry. 2021;91:107452.

29. Jiao S, Zou Q, Guo H, Shi L. iTTCA-RF: a random forest predictor for tumor T cell antigens. Journal of translational medicine. 2021;19(1):1-11.

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;2825–2830.

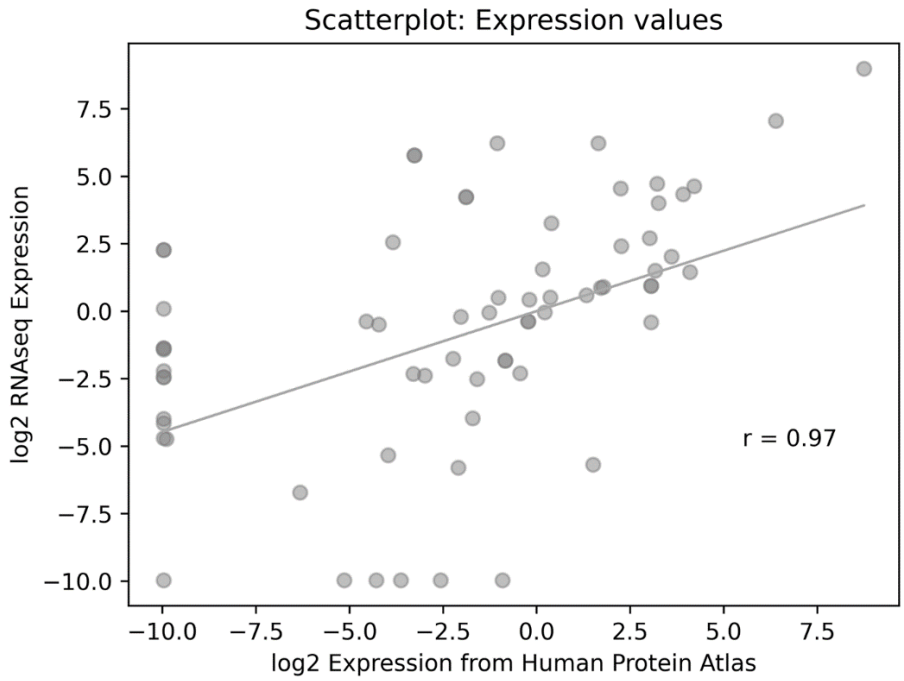

Figure S1. A positive correlation between Human Protein Atlas inferred expression values and RNAseq derived values for the proteins generating neopeptides in patients 005 and 006 (Pearson's correlation test, r = 0.97).

Table S2. Comparison of AUC of methods trained with neopeptides contained in the in-house neoepitope dataset, including and excluding such peptides

	AUC including over-lapped peptides	AUC excluding overlapped peptides
DeepHLApan immunogenicity	**0.53**	0.52
DeepImmuno	0.47	0.47
NetCleave	**0.41**	0.4
NetMHCpanExp	**0.54**	0.53

Table S3. Performance metrics of all the methods reviewed

Method	AUC ROC	AUC ROC 0.1	AUC ROC 0.2	Spearman correlation	Pearson correlation
MHCflurry AP	**0.609**	0.53	0.542	0.169	0.152
PRIME score	0.604	**0.536**	0.533	**0.188**	0.178
PRIME rank	0.601	0.494	0.466	-0.182	-0.12
Variant Allele Frequency	0.6	0.492	0.52	0.154	0.098
INeo-Epp neoantigen	0.584	0.49	0.489	0.139	0.032
ProteaSMM constitutive proteasome	0.58	0.514	0.513	0.139	0.148
HLAthena MSiCE	0.58	0.474	0.462	-0.149	-0.135
HLAthena MSiC	0.576	0.474	0.458	-0.141	-0.135
MHCflurry PS	0.571	0.53	0.547	0.154	0.186
NetCTLpan TAP	0.568	0.496	0.505	0.098	0.065
ProteaSMM immunoproteasome	0.561	0.49	0.485	0.097	0.081
MixMHCpred	0.556	0.474	0.456	-0.095	-0.128
TA predictor	0.552	0.531	0.529	0.093	0.062
HLAthena MSiE	0.549	0.474	0.47	-0.105	-0.103
IEDB immunogenicity	0.548	0.504	0.52	0.072	0.062
NetMHCpanExp rank	0.54	0.531	0.517	-0.09	-0.051
Antigen.garnish Dissimilarity	0.533	0.476	0.494	0.087	-0.105
DeepNetBim binding	0.53	0.493	0.495	0.081	0.124
DeepHLApan immunogenic score	0.529	0.528	0.517	0.046	0.031
INeo-Epp antigen	0.524	0.525	0.533	0.064	0.035
NetMHCstabpan Thalf(h)	0.521	0.508	0.494	0.053	0.021
iTTCA-RF	0.516	0.501	0.502	0.019	-0.006
DeepNetBim immunogenicity	0.513	0.502	0.504	0.045	0.054
Kernel Self Similarity without anchors	0.509	0.499	0.499	-0.041	-0.062
Antigen.garnish Foreignness score	0.505	0.511	0.519	-0.004	0.018
HPA expression	0.5	0.493	0.51	0.009	-0.049
NetMHCpan 4.0	0.499	0.53	**0.548**	-0.032	-0.099
Kernel Self Similarity	0.497	0.496	0.488	-0.001	0.055
DeepNetBim immunogenicity probability	0.495	0.499	0.498	0.017	0.044
MHCflurry BA	0.491	0.487	0.488	-0.021	-0.095
NetMHCstabpan rank	0.487	0.48	0.484	0	-0.075
DeepHLApan binding score	0.485	0.482	0.48	-0.008	0.024
iPCPS Proteasome C-terminal	0.48	0.479	0.477	-0.018	0.014
iTTCA-RF probability	0.48	0.474	0.457	-0.017	-0.019
DeepImmuno	0.466	0.491	0.491	-0.068	-0.098
iPCPS Immunoproteasome C-terminal	0.464	0.494	0.478	-0.04	-0.021
iPCPS Proteasome	0.46	0.495	0.49	-0.09	-0.11
NetCTLpan Cleavage	0.459	0.474	0.456	-0.075	-0.011
iPCPS Proteasome internal	0.444	0.477	0.481	0.115	0.072
Paired sequence similarity	0.434	0.507	0.524	0.071	0.07
iPCPS Immunoproteasome	0.423	0.487	0.472	-0.147	-0.145
iPCPS Immunoproteasome internal	0.415	0.505	0.499	0.143	**0.206**
NetCleave	0.411	0.488	0.468	-0.152	-0.148