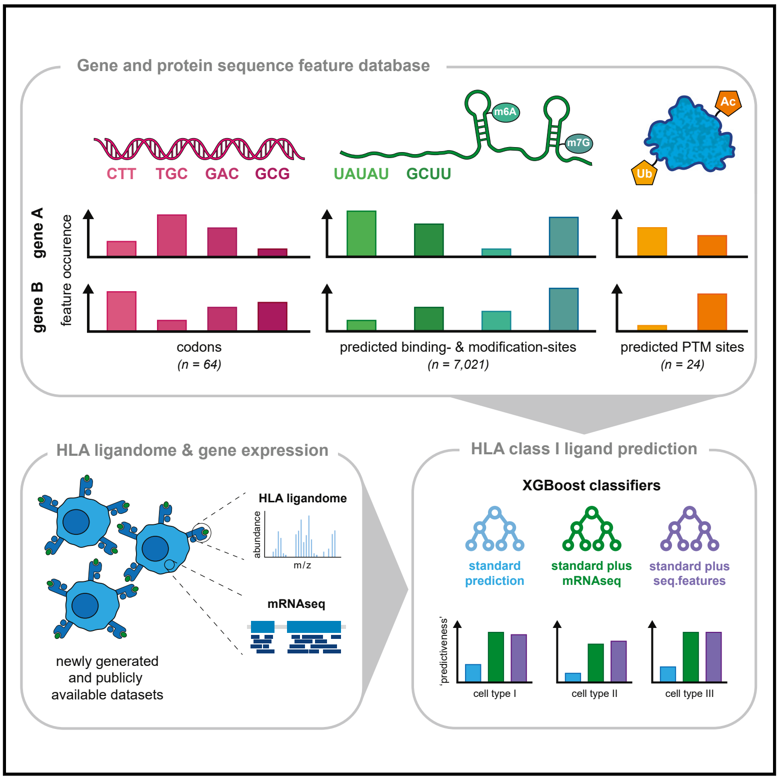
Cell Reports

Article

Gene and protein sequence features augment HLA class I ligand predictions

Graphical abstract



Highlights

- Gene and protein sequence features inform on HLA class I sampling
- Predicted RNA and protein modifications are most informative
- Models that integrate sequence features improve HLA class I ligand predictions

Authors

Kaspar Bresser, Benoit P. Nicolet,
Anita Jeko, ..., Albert J.R. Heck,
Monika C. Wolkers, Ton N. Schumacher

Correspondence

t.schumacher@nki.nl

In brief

Understanding the rules that control the composition of the HLA class I ligandome is highly important in the design of cancer immunotherapies. Bresser et al. show that the prediction of HLA class I ligands can be significantly improved by incorporation of gene and protein sequence features of source proteins.

Bresser et al., 2024, Cell Reports *43*, 114325
June 25, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.celrep.2024.114325

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Gene and protein sequence features augment HLA class I ligand predictions

Kaspar Bresser,[1,2,9,10,11] Benoit P. Nicolet,[3,4,5] Anita Jeko,[6] Wei Wu,[6] Fabricio Loayza-Puch,[7] Reuven Agami,[8] Albert J.R. Heck,[6] Monika C. Wolkers,[3,4,5] and Ton N. Schumacher[1,2,12,*]

[1]Department of Molecular Oncology and Immunology, Netherlands Cancer Institute, Oncode Institute, Amsterdam, the Netherlands
[2]Department of Hematology, Leiden University Medical Center, Leiden, the Netherlands
[3]Sanquin Blood Supply Foundation, Department of Research, T cell differentiation lab, Amsterdam, The Netherlands
[4]Amsterdam UMC, University of Amsterdam, Landsteiner Laboratory, Amsterdam, The Netherlands
[5]Oncode Institute, Utrecht, The Netherlands
[6]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Utrecht, the Netherlands
[7]Translational Control and Metabolism, German Cancer Research Center (DKFZ), Heidelberg, Germany
[8]Division of Oncogenomics, Oncode Institute, The Netherlands Cancer Institute, Amsterdam, the Netherlands
[9]Present address: Sanquin Blood Supply Foundation, Department of Research, T cell differentiation lab, Amsterdam, The Netherlands
[10]Present address: Amsterdam UMC, University of Amsterdam, Landsteiner Laboratory, Amsterdam, The Netherlands
[11]Present address: Oncode Institute, Utrecht, The Netherlands
[12]Lead contact
*Correspondence: t.schumacher@nki.nl

https://doi.org/10.1016/j.celrep.2024.114325

SUMMARY

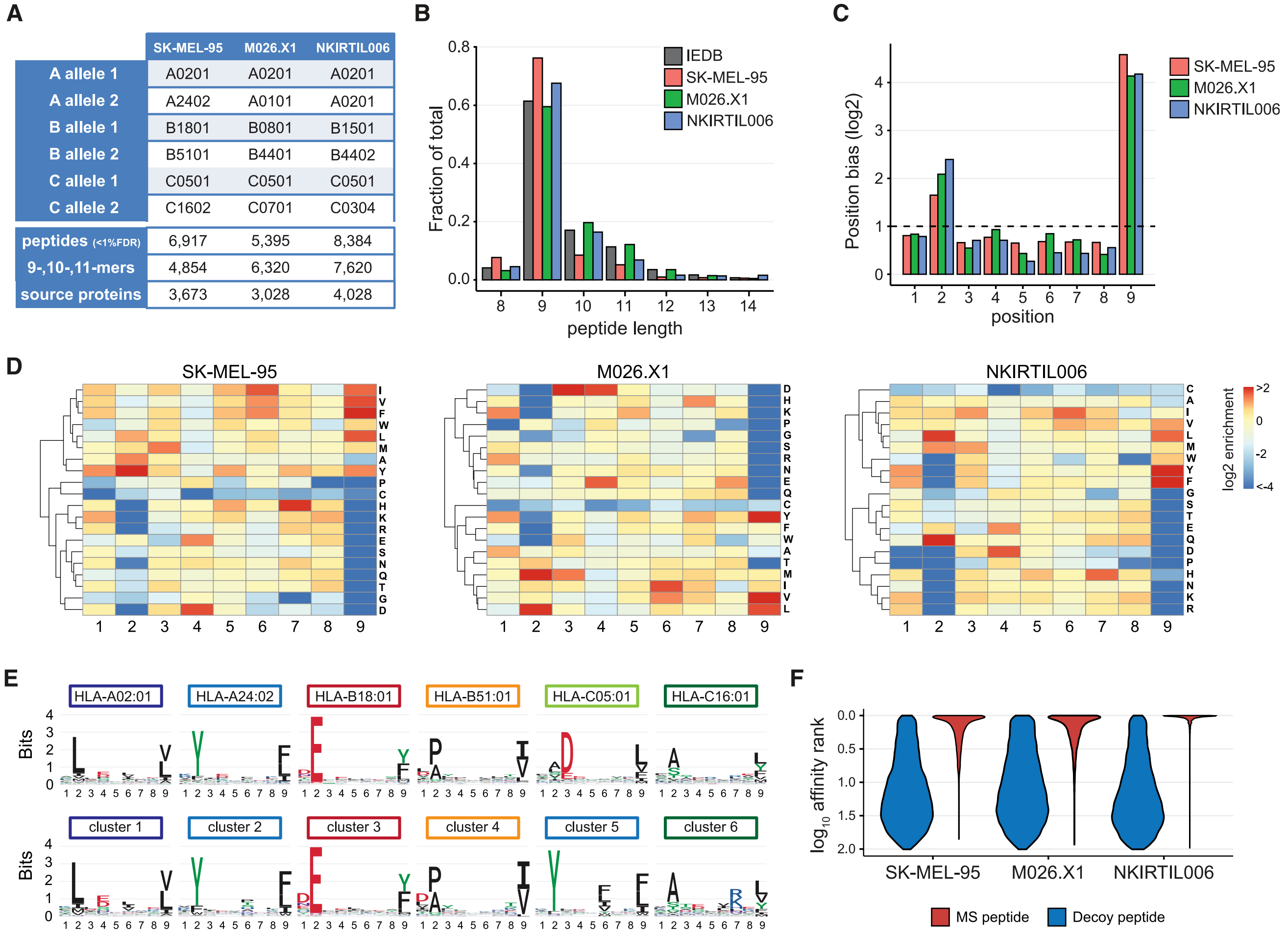
The sensitivity of malignant tissues to T cell-based immunotherapies depends on the presence of targetable human leukocyte antigen (HLA) class I ligands. Peptide-intrinsic factors, such as HLA class I affinity and proteasomal processing, have been established as determinants of HLA ligand presentation. However, the role of gene and protein sequence features as determinants of epitope presentation has not been systematically evaluated. We perform HLA ligandome mass spectrometry to evaluate the contribution of 7,135 gene and protein sequence features to HLA sampling. This analysis reveals that a number of predicted modifiers of mRNA and protein abundance and turnover, including predicted mRNA methylation and protein ubiquitination sites, inform on the presence of HLA ligands. Importantly, integration of such "hard-coded" sequence features into a machine learning approach augments HLA ligand predictions to a comparable degree as experimental measures of gene expression. Our study highlights the value of gene and protein features for HLA ligand predictions.

INTRODUCTION

Spontaneous or immunotherapy-induced recognition and destruction of malignant tissues by the CD8$^+$ T cell-based immune system is dependent on presentation of human leukocyte antigen (HLA) class I-bound peptides to antigen-specific CD8$^+$ T cells.[1–3] Consequently, the composition of the pool of peptide-HLA class I complexes at the cell surface—or the HLA class I ligandome—strongly determines the "visibility" of tumor cells to CD8$^+$ cytotoxic T cells. Understanding the various factors that determine the composition of this HLA ligandome is thus of major value for cancer immunotherapy.

The HLA class I ligandome is primarily generated through the intracellular degradation of proteins by the proteasome and subsequent translocation of peptide fragments into the endoplasmic reticulum (ER) lumen by the transporter associated with antigen processing. These peptides can undergo further trimming by ER-resident aminopeptidases, bind to the peptide-binding groove of HLA class I molecules, and finally traffic to the cell surface to be presented to the CD8$^+$ T cell pool.[4,5] The number of peptides that can theoretically be generated from the human proteome is vast, adding up to approximately 10^7 distinct peptides for 9-meric species alone.[6] This large space poses a substantial challenge in the prediction of the HLA ligandome of a cell population of interest. Over the past decades, significant advances have been made to reduce this complexity, primarily by focusing on characteristics of the peptide itself or its surrounding sequence. Specifically, HLA class I ligands bind to the peptide-binding groove of HLA class I through HLA class I allele-specific "anchor" residues, a feature that has been leveraged in the development of allele-specific predictive algorithms.[7–9] In addition, the cleavage preference of the proteasome[10] has been used to improve epitope prediction accuracy.[11,12]

Beyond local sequence characteristics, a number of other protein-level features may be expected to play an important role in the generation of HLA binding peptides; for instance, by influencing protein abundance and turnover.[13–15] In prior work, transcriptome measurements have been used as a proxy for protein expression to aid HLA ligand predictions. However, the variation in protein levels is only partly (on average 36%) explained by



	SK-MEL-95	M026.X1	NKIRTIL006
A allele 1	A0201	A0201	A0201
A allele 2	A2402	A0101	A0201
B allele 1	B1801	B0801	B1501
B allele 2	B5101	B4401	B4402
C allele 1	C0501	C0501	C0501
C allele 2	C1602	C0701	C0304
peptides (<1%FDR)	6,917	5,395	8,384
9-,10-,11-mers	4,854	6,320	7,620
source proteins	3,673	3,028	4,028

Figure 1. Identification of HLA ligandomes

(A) HLA class I haplotype of the melanoma lines used and number of peptides and source proteins identified.

(B) Peptide length distribution of each LC-MS dataset compared to the peptide length distribution of known melanoma-derived HLA class I ligands deposited into IEDB.

(C and D) Enrichment of the indicated amino acids relative to amino acid occurrence in the proteome, at each position of all 9-meric species in the datasets. A summary depicting the median of the absolute enrichment values of all amino acids for each position (C) and heatmaps visualizing hierarchical clustering of amino acid enrichment (D) are shown.

(E) Sequence logos of all 9-mer ligands deposited into IEDB for the HLA class I alleles expressed by SK-MEL-95 (top) and the sequence logos of 6 peptide clusters obtained using the GibbsCluster algorithm (bottom). The number of clusters was constrained to the number of expressed HLA class I alleles.

(F) Affinity percentile rank scores of LC-MS-detected peptides compared to randomly drawn peptides from the human proteome (decoy peptides).

See also Figure S1.

mRNA sequencing data in different mammalian cell types[16–18] because of factors such as post-transcriptional regulation. Such post-transcriptional regulation includes the activity of RNA-binding proteins and non-coding RNA species but also sequence-intrinsic features (e.g., GC content and codon usage) that can affect the translational output of mRNAs.[19,20] Furthermore, post-translational modifications, including ubiquitination and glycosylation, are known to modulate protein abundance, localization, and turnover rates[21,22] and may thereby influence epitope sampling.

In this study, we aimed to examine the potential value of gene and protein sequence features in the prediction of the HLA class I ligands. Implementing a machine learning approach, we show that the performance of such predictions can be improved through the addition of sequence features. Importantly, predictive models that make use of such features achieve the same level of predictive power as models that incorporate experimental measurements of gene expression levels, and the predictive value of these features was generalizable to external data. Our data exemplify how the "hard-coded" information of gene and protein sequence features can be exploited to infer a cell's proteomic content and its derivatives.

RESULTS

Identification of human melanoma HLA ligandomes

To investigate putative determinants of the HLA ligandome, we performed liquid chromatography-mass spectrometry (LC-MS) on pan-HLA immunoprecipitates of three melanoma lines

(Figure 1A), resulting in the identification of 20,696 peptides derived from 8,554 proteins at a false discovery rate of <1%. The length distribution of the LC-MS-detected peptides closely matched that of known melanoma-derived HLA ligands (Immune Epitope Database [IEDB][23]; Figure 1B), with the vast majority of peptides consisting of 9-meric species. Examination of positional frequencies of each amino acid revealed strong usage biases at positions 2 and 9 (Figures 1C and 1D). To assess whether this observed amino acid enrichment was explained by the known ligand preference of the HLA class I haplotypes expressed by these tumor lines, 9-meric peptide sequences from each melanoma line were clustered using the GibbsCluster algorithm.[24] This analysis revealed dominant motifs present in each of the HLA ligandomes that closely matched the corresponding HLA haplotype consensus binding motifs for 11 of 11 HLA A and B alleles and 5 of 6 HLA C alleles (Figures 1E, S1A, and S1B).

As a final quality check, LC-MS-detected peptides generally had a high predicted affinity for the expressed HLA class I alleles (Figures 1F and S1C), and 97.6%–99.4% of all MS-detected peptides could be reliably assigned to one of the expressed HLA class I alleles using the MHCMotifDecon algorithm[25] (Figures S1D and S1E). Notably, "unassigned" peptides exhibited an unusual length distribution (Figure S1F), and therefore the subsequent analyses were restricted to 9- to 11-meric species.

Gene and protein features inform on HLA sampling

Gene and protein sequence features, such as post-transcriptional or post-translational modification sites, have been shown to influence mRNA or protein abundance.[20,26–28] In line with this, sequence features can be integrated into machine learning models to predict mRNA and protein expression levels.[29] To determine whether such features can be employed to predict the presence of HLA ligands within the proteome, we made use of a library of 7,135 "hard-coded" sequence features. This feature library includes codon and amino acid usage, RNA-binding motifs from 142 RNA-binding proteins, predicted microRNA binding site scores, and RNA modification sites that were separately analyzed for the 5′ untranslated region (UTR), 3′ UTR, and coding sequence (CDS).[30] Predicted post-translational modification (PTM) sites, such as ubiquitination, acetylation, and malonylation sites, were additionally included. Of note, this sequence feature library comprises mRNA and protein sites than can potentially be modified or recognized, irrespective of whether these sites are, in fact, utilized, and these features are hence considered hard coded.

To assess whether individual sequence features can inform on HLA sampling, we first performed an exploratory analysis on a subset of features that could be assigned to five major feature classes (5′ UTR, CDS, 3′ UTR, miRNA binding, and PTM) and that displayed a substantial degree of variance across the proteome (Figure 2A; 5,782 of 7,135 features in the library). A set of 2,000 HLA ligands was drawn from each tumor line and supplemented with a 4-fold excess of decoy peptides that were randomly sampled from the human proteome. This dataset was then used to train individual random forest classifiers for each tumor line and each sequence feature class, which were subsequently used to determine the importance of these sequence features to each of the obtained classifiers (Figures 2B and S2A, showing normalized importance plots and random forest metrics). The importance of sequence features was highly consistent between the different melanoma ligandome datasets, indicating that a shared set of features reliably informed on the presence of HLA ligands (Figure 3C). Furthermore, direct comparison of the occurrence of high-importance sequence features within source proteins of HLA ligands and decoy peptides revealed significant differences for a set of sequence features (Figure S2). For example, HLA ligands were preferentially sampled from proteins that contained a higher number of predicted sites for ubiquitination and acetylation, two PTMs that can regulate targeted proteasomal degradation and protein stability[31–33] (Figure 2D). Predicted N1-methyladenosine (m1A) sites within the 5′ UTR were also enriched in the mRNA of source proteins of HLA ligands, an effect that appears to be consistent with the prior observation of enhanced translation efficiency of m1A-modified mRNA molecules.[27] In contrast, 5′ UTR length and occurrence of G-rich motifs in the CDS, features that have been suggested previously to negatively impact mRNA levels and translation, respectively,[34,35] were negatively associated with the presence of HLA ligands (Figure 2D).

To understand the ability of individual sequence features to contribute to HLA ligand prediction in a more quantitative manner, a custom enrichment score was calculated for each of the selected features (STAR Methods). In brief, the set of HLA ligands and decoy peptides was either sorted by the occurrence of each feature or was arranged randomly. Subsequently, the quantity of HLA ligands present in the top 50% of ranked peptides was compared between these two cases, reflecting the benefit of each feature when used as a single determinant. In concordance with the prior analysis (Figure 2C), miRNA binding site quantities exhibited no detectable bias toward HLA ligands or decoy peptides. In contrast, sequence features from the other classes showed a consistent capacity to enrich or deplete the presence of HLA ligands (Figures 2E and S2C). The most prominent associations were observed in the CDS and PTM classes (Figure 2F), with many features increasing the number of ligands detected by more than 20%. To assess whether the association between specific PTM sites and HLA class I sampling is also observed for experimentally observed PTMs, we interrogated the dataset from Abelin et al. that comprises matched HLA ligandome, ubiquitinome, acetylome, and phosphorome measurements of human lung adenocarcinomas.[36] This analysis suggests that the presence of acetyl and ubiquitin groups is associated with the presence of HLA ligands in source proteins (Figures S2D and S2E). Computed m1A and N7-methylguanosine sites were also predictive of the presence of HLA ligands in the protein product and irrespective of their location within either coding sequence or untranslated regions (Figure 2E), an observation that aligns with their general translation-enhancing capacity.[27,37] Intriguingly, even though GC content was consistently informative on HLA sampling, its directionality was context dependent (i.e., positively correlated in the 5′ UTR and negatively correlated in the 3′ UTR and CDS), in line with prior reports suggesting that GC content may influence mRNA levels in a location-dependent manner.[6,18,38] Together, the above analyses show that gene and protein sequence features can individually inform on the presence of HLA class I ligands.

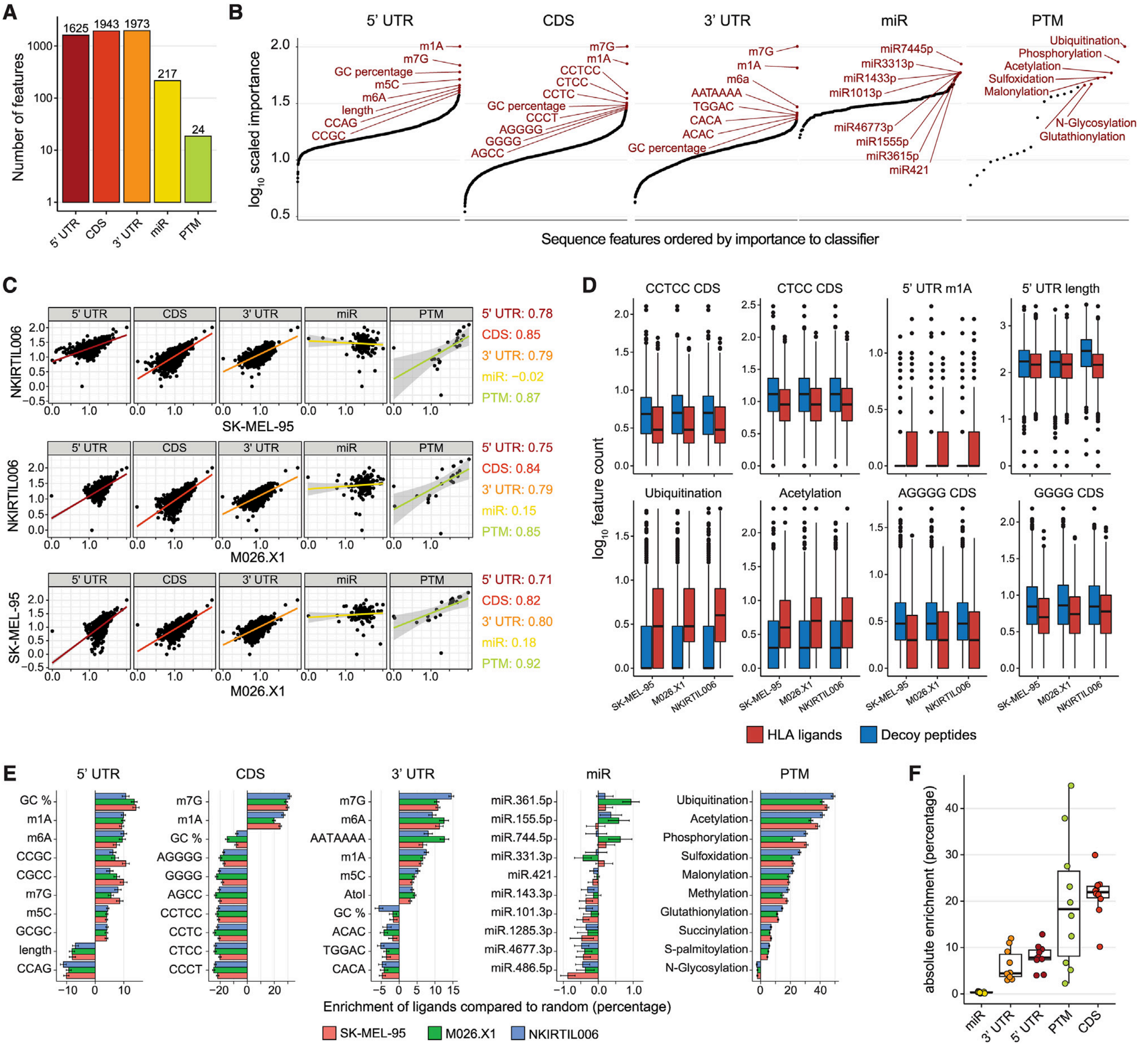


Figure 2. Sequence features inform on HLA sampling

Random forest models were trained using HLA ligandome data from each melanoma line and using individual classes of gene and protein sequence features to identify HLA ligands.

(A) Sequence feature classes used to fit random forest classifiers for each melanoma line. Values indicate the number of features per class.

(B) Mean importance of sequence features of each class to the random forest models. Feature importance represents the mean decrease in accuracy when that sequence feature was excluded from the model. Importance scores are re-scaled per feature class to a 0–100 scale. Dots indicate individual features.

(C) Comparison of the importance of all sequence features across the individual random forest models. Dots indicate individual features, and linear regressions are shown as colored lines and 95% confidence intervals as gray areas. Colored text denotes the respective Pearson correlation coefficients.

(D) Comparison of sequence feature occurrence between 500 LC-MS-detected HLA ligands and the same number of decoy peptides. Selected sequence features are shown. Boxplots indicate group median and 25th and 75th percentiles, whiskers indicate the interquartile range multiplied by 1.5, and dots signify individual peptides.

(E and F) HLA ligands and decoy peptides were either ranked at random or by the indicated sequence feature, and the number of HLA ligands in the top 50% ranked peptides was quantified. Data depict the relative increase in HLA ligands found comparing feature-ranked and random-ranked peptides; see STAR Methods for details. (E) Bars indicate the mean percentage increase of 50 bootstraps, and error bars depict 95% confidence intervals. (F) Comparison of averaged absolute enrichment values between feature classes. Boxplots indicate group median and 25th and 75th percentiles, whiskers indicate the interquartile range multiplied by 1.5, and dots signify individual features.

See also Figure S2.

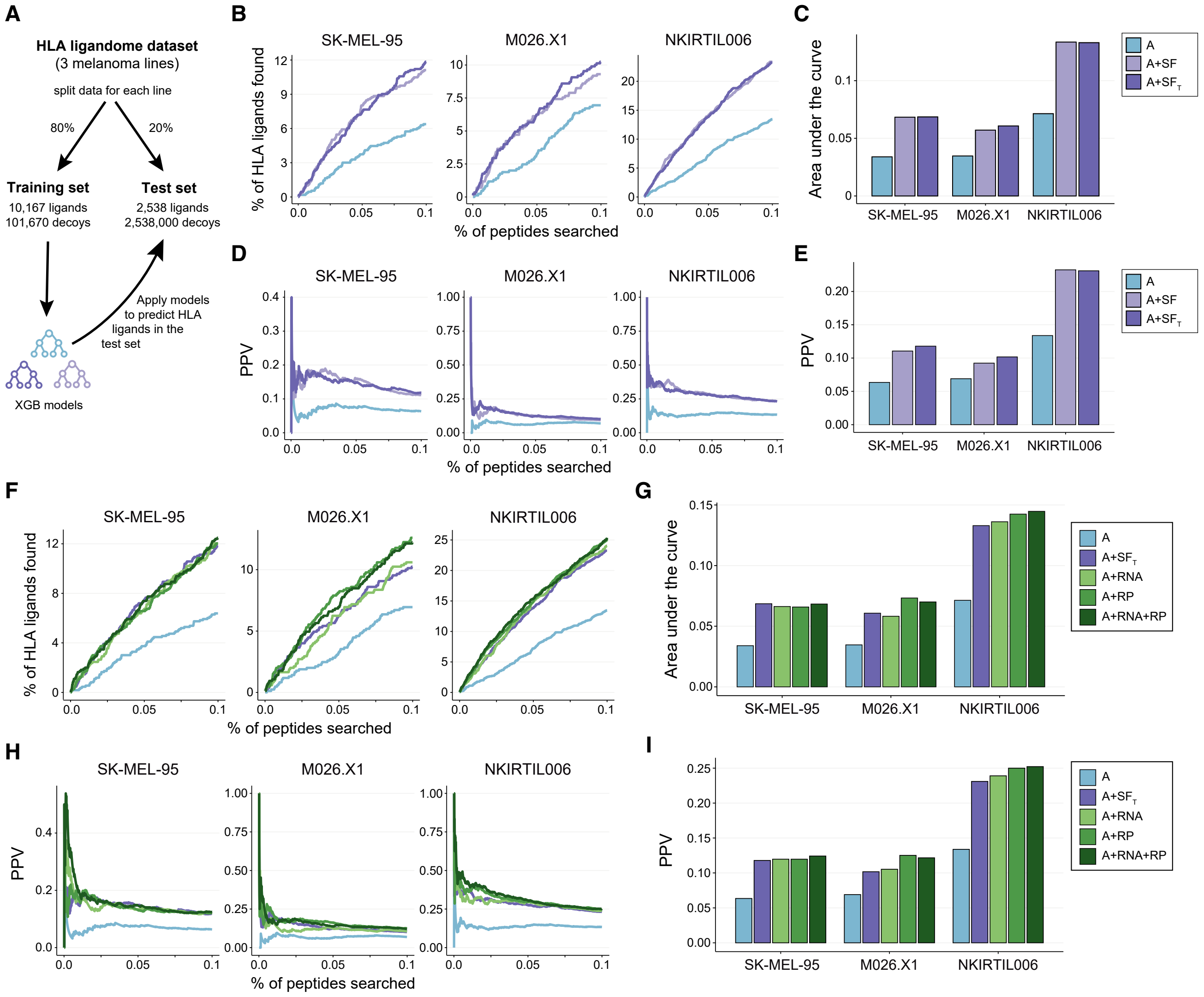


Figure 3. Value of sequence features in HLA ligand predictions

(A) The melanoma line dataset was split into a training set and test set at an 80/20 ratio. The training set was used to build XGB classifiers using different combinations of features. Classifiers were either trained using the full sequence feature library (n = 7,124) or a trimmed version (n = 680).

(B and C) Number of true HLA ligands observed in the top 0.1% of predicted peptides from the matched melanoma line test set by each of the indicated models. Line graphs depicting the cumulative sum (B) and bar charts depicting areas under the curve (AUCs) (C) are shown.

(D and E) Positive predictive value (PPV) at each peptide rank within the top 0.1% of predicted peptides from the melanoma line test set by each of the indicated models.

(F and G) Quantity of true HLA ligands observed in the top 0.1% of predicted peptides from the melanoma line test set by each of the indicated models. Line graphs depicting the cumulative sum (F) and bar charts depicting AUCs (G) are shown.

(H and I) PPV at each peptide rank within the top 0.1% of predicted peptides from the melanoma line test set by each of the indicated models. Features used to build classifiers were predicted HLA class I affinity (A), transcript abundance (RNA), ribosome occupancy (RP), sequence feature library (SF), and trimmed sequence feature library (SF_T).

Sequence features augment HLA ligand predictions

Having shown that individual sequence features can inform on HLA sampling, we next assessed whether these features can be leveraged to improve HLA ligand prediction models. To this end, the melanoma HLA class I ligand dataset was divided into a training set (80%) and test set (20%) that were supplemented with a 4-fold and 1,000-fold excess of decoy peptides, respectively. To evaluate the added value of sequence features to classical HLA ligand prediction methods, such as netMHCpan (HLA affinity), the training set was used to generate multiple XGBoost[39] classifier models (Figure 3A), each integrating a different set of explanatory variables. As reported previously,[9,12] computed HLA affinity was strongly predictive of HLA sampling (Figures S3A and S3B). Importantly, applying the obtained

XGBoost models to predict HLA ligands in the test set revealed that the classifier that included sequence feature information consistently and significantly outperformed models that lacked this information. Specifically, the model that included sequence features consistently ranked true HLA ligands at a higher position (Figures 3B and 3C) and increased positive predictive value by approximately 1.5-fold (Figures 3D and 3E). The selection of informative features is a frequently used strategy in machine learning to reduce model complexity and, thereby, computational cost and to improve model generalization. To evaluate this approach, an additional "trimmed" sequence feature XGBoost model (termed A+SF$_T$) was trained using a smaller set of sequence features (n = 680) that contributed most substantially to the accuracy of the original classifier. This more efficient XGBoost classifier demonstrated equal model performance compared to the original sequence feature classifier (Figures 3B–3E). Of note, the strongest contributing features to the A+SF$_T$ model belonged to the PTM and CDS classes (Figures S3B and S3C), in line with the analysis of the predictive value of individual sequence features (Figure 2).

Sequence features can match "wet lab" measures of gene transcription and translation

Prior studies have established that mRNA abundance measurements can increase the accuracy of HLA class I ligandome predictions,[40–42] and recent efforts have indicated that ribosome occupancy (as a measure of active protein translation[43,44]) can inform on HLA class I sampling.[45] To understand the value of hard-coded protein and gene sequence features relative to these experimental measurements of either gene expression or ribosome occupancy, we generated mRNA sequencing (transcript abundance) and ribosome profiling (protein translation activity) datasets for each of the melanoma lines. Consistent with previous reports, both measures of gene expression were strongly indicative of HLA sampling, and the combination of transcript-level data and ribosome occupancy data offered little additional benefit (Figures S4A−S4E). Next, to directly compare the predictive value of sequence features versus these "wet lab" measures of gene transcription and protein translation, additional XGBoost classifiers were trained that included these metrics. Application of this new set of models to the test dataset revealed that the XGBoost model that included sequence features was able to predict true HLA ligands at an equal potency as models that incorporated "wet lab" measurements of gene transcription or protein translation (Figures 3F–3I). Interestingly, addition of "wet lab" measurements of gene transcription and protein translation in a model that contained sequence features did not consistently improve predictiveness (Figures S4F and S4G), indicating that sequence features and "wet lab" measurements may capture similar information. Together, these data show that gene and protein sequence features jointly provide a similar degree of information on HLA ligandome composition as experimentally obtained expression levels.

Sequence feature-based HLA ligand classifiers are generalizable

Next, we sought to understand to what extent the information value of sequence features in HLA ligandome predictions is generalizable to other tumor types, to different HLA class I alleles, and to independent datasets. To this purpose, we assessed the performance of the different classifiers on HLA ligandome data obtained from either mono-allelic B721.221 cell lines,[41] cultured T and B cells, meningeal tumor lines, ovarian carcinomas,[46] and lung adenocarcinomas.[47] In addition, to evaluate whether the predictive power of the sequence feature-based classifier could be further boosted by extending the size of the training dataset, an additional XGBoost model was trained using the HLA ligandome data generated here combined with HLA ligandome data obtained from two external studies[40,47] (termed A+SF 2.0). Application of these classifiers to data derived from 20 mono-allelic B721.221 lines showed that XGBoost models that included sequence feature information reproducibly outperformed models based only on affinity (Figures 4A and 4B). Furthermore, the extended A+SF 2.0 model performed better than the model trained on internal LC-MS data (A+SF$_T$) and showed increased performance relative to the RNA sequencing-based model by a small but significant margin. This improved performance of the A+SF 2.0 model was likewise observed in HLA ligandome data obtained from lung adenocarcinomas (Figure 4C), meningeal cancer lines (Figure 4D), Epstein-Barr virus (EBV) immortalized B cells, and tumor-infiltrating T cells (Figure 4E).

To further test the robustness of the sequence feature based XGBoost models, their performance was assessed on HLA ligandome data obtained from ovarian carcinoma cells cultured in the presence or absence of interferon γ (IFNγ),[46] and an HLA ligand dataset comprising non-canonical open reading frames (ORFs) detected in multiple B721.221 cell lines.[48] Despite the strong modulatory effects of IFNγ on the cellular proteome and the antigen-processing machinery, XGBoost models that included sequence features still resulted in improved performance compared to those solely based on predicted affinity (Figure S5A). The A+SF 2.0 model also performed significantly better than the affinity-based model in predicting ligands derived from non-canonical ORFs (Figure S5B). Taken together, the A+SF 2.0 XGBoost model is generalizable across various cell types, maintains performance upon cellular perturbation (IFNγ treatment), and enhances prediction of HLA ligands derived from non-canonical ORFs.

DISCUSSION

Gene and protein sequence features represent a class of "hard-coded" regulators of protein expression, influencing this process at many different levels. In this study, we leveraged a large set of such gene and protein features to assess their contribution to the composition of the HLA class I ligandome. We demonstrate that sequence features can augment HLA ligand predictions and that the predictive gain obtained in these models is equal to that of models that incorporate experimentally obtained gene expression and protein translation data.

While not formally assessed here, it is expected that at least some of the sequence features contribute to HLA ligand predictions by providing a proxy for protein abundance. This notion is supported by the observation that sequence features such as mRNA region length, GC content, and post-transcriptional

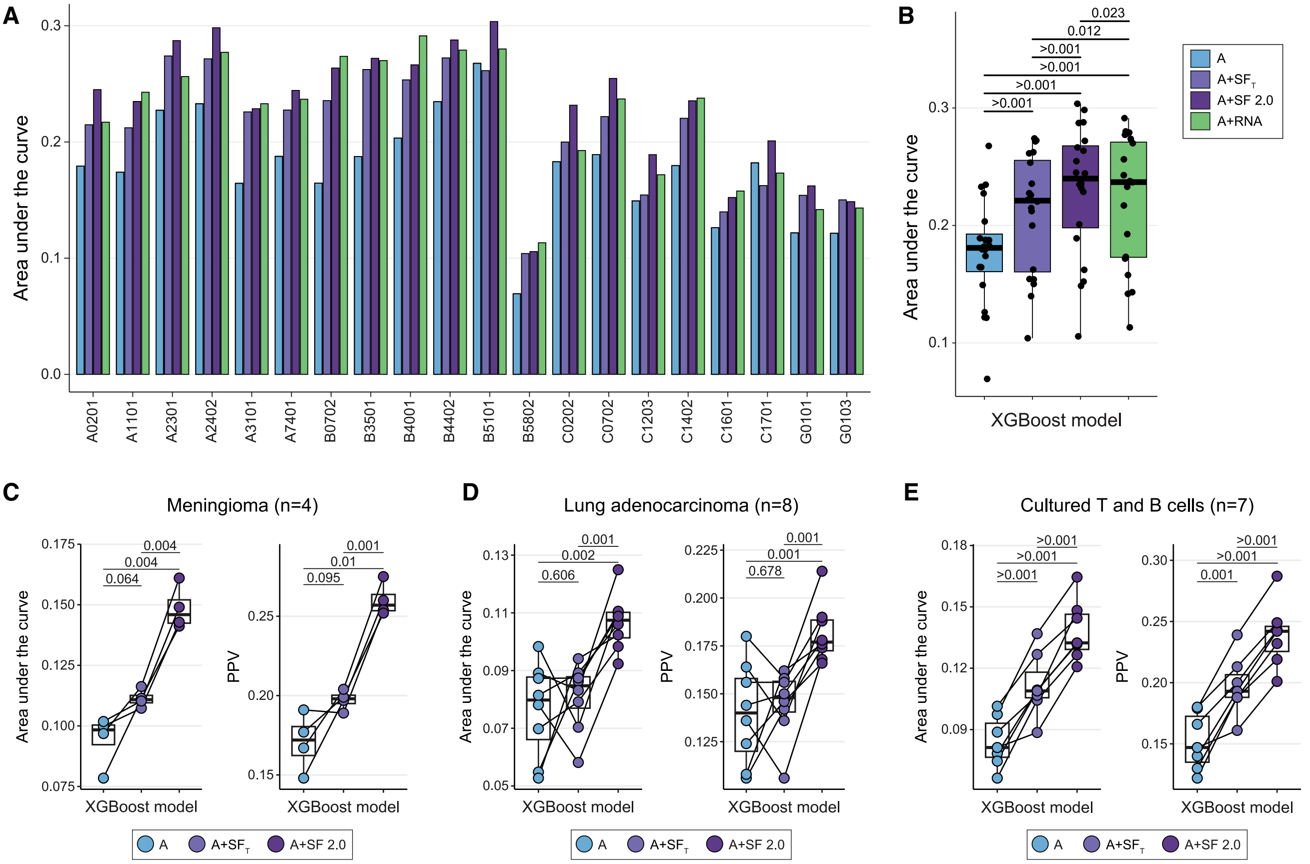


Figure 4. Sequence feature-based XGBoost models generalize to external data

XGBoost classifiers were validated using HLA ligandome data from 3 external datasets.

(A and B) Validation of the indicated XGBoost models on HLA ligandome data obtained from 20 mono-allelic B721.221 cell lines. (A) Bar charts depict the area under the curve calculated over the number of true HLA ligands observed in the top 0.1% of predicted peptides. (B) Boxplots summarizing the data shown in (A). Boxplots indicate group median and 25th and 75th percentiles, whiskers indicate the interquartile range multiplied by 1.5, and dots signify individual cell lines.

(C–E) Validation of the indicated XGBoost models on HLA ligandome data obtained from 4 meningeal cancer lines[46] (C), 8 lung adenocarcinomas[47] (D), 4 EBV immortalized B cell lines, and 3 tumor-infiltrating T cell cultures[46] (referred as cultured T and B cells; E). Boxplots depict the area under the curve calculated over the number of true HLA ligands observed in the top 0.1% of predicted peptides (left graphs) or the PPV within the top 0.1% of predicted peptides (right graphs). Boxplots indicate group median and 25th and 75th percentiles, whiskers indicate the interquartile range multiplied by 1.5, dots signify individual samples, and lines connect matched samples. Features used to build classifiers were predicted HLA class I affinity (A), transcript abundance (RNA), trimmed sequence feature library (SF$_T$), and sequence feature library 2.0 (SF 2.0). The *p* values represent results of two-sided paired t tests. The Holm-Bonferroni method was applied to correct for multiple testing.

See also Figure S5.

modifications can be used to help predict protein levels.[18] Furthermore, predicted RNA methylation sites were among the features that were most prominently associated with the presence of HLA ligands, an observation that may be explained by their known modulatory effect on both mRNA stability and translation efficiency.[27,28,37,49] In addition to features involved in mRNA regulation and translation, our data reveal that the predicted occurrence of several PTM sites informed on the presence of HLA ligands. For ubiquitination, the PTM that displayed the highest predictive value, its positive association with HLA ligand yield may be caused by an enhanced accessibility to proteasomal degradation.[21,33,50] For other PTMs that were predictive of HLA ligands, such as methylation and acetylation, their involvement in specific pathways is less well understood.[22,51,52] Our data provide correlative evidence that that these modifications influence availability of proteins to the antigen processing machinery but further work will be required to formally test this.

Improvement of HLA ligand prediction approaches remains an active field of research, with the aim to, for example, allow the more precise selection of cancer (neo)antigens for therapeutic purposes.[40,41,53,54] Because of its generalizable nature and lack of requirement for direct transcriptome measurements, we envision that the approach described here will be of value in these efforts.

Limitations of the study

The present study has the following limitations. (1) Predicted protein and RNA modification sites were found to influence HLA class I sampling. As our models rely on predicted modification sites, it cannot be concluded that the actual presence of these modifications influences HLA class I ligand sampling. Although we present an analysis in this work that suggests an association between experimentally observed ubiquitination and acetylation and HLA class I ligand yield, further work will be required to validate this observation. (2) This study largely focused on the rules that govern HLA class I ligand selection in transformed cell lines and cancer tissues. While the presented models are applicable across various transformed tissues and cell lines of different origin, it is possible that non-transformed cell types and specialized cell types (e.g., antigen-presenting cells) may require the training of dedicated prediction models. (3) Inclusion of a larger set of HLA ligandome data significantly improved the predictive power of the sequence feature-based XGBoost model, and it remains to be determined whether an additional expansion of the training data can further boost the performance of these models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Patient-derived melanoma cell lines
- METHOD DETAILS
  - Patient-derived melanoma cell lines
  - Cell culture
  - HLA class I peptide isolation and LC-MS/MS
  - HLA class I peptide analysis
  - mRNA sequencing
  - Ribosome profiling
  - Characterization of LC-MS detected peptides
  - Peptide database construction
  - Feature library construction
  - Importance assessment of sequence feature classes
  - XGBoost classifiers
  - External HLA ligandome data
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.114325.

ACKNOWLEDGMENTS

We would like to thank Mireille Toebes for helpful input during generation of the HLA class I ligandome dataset. We would like to thank Prof. Dr. Daniel Peeper for kindly providing the SK-MEL-95 and M026.X1 cell lines. This work was supported by Institute for Chemical Immunology (ICI) grant 003 (to T.N.S.).

AUTHOR CONTRIBUTIONS

K.B., conceptualization, methodology, validation, formal analysis, investigation, data curation, writing – original draft, writing – review & editing, visualization; B.P.N., conceptualization, methodology, formal analysis, writing – review & editing; A.J., formal analysis, methodology, and investigation; W.D., data curation; F.L.-P., investigation, methodology, and resources; R.A., methodology, resources, and supervision; A.J.R.H., conceptualization, methodology, resources, supervision, and funding acquisition; M.C.W., methodology, writing – review & editing, and supervision; T.N.S., conceptualization, methodology, writing – review & editing, supervision, and funding acquisition.

DECLARATION OF INTERESTS

T.N.S. is an advisor for Allogene Therapeutics, Asher Bio, Merus, Neogene Therapeutics, and Scenic Biotech; a stockholder in Allogene Therapeutics, Asher Bio, Cell Control, Celsius, Merus, and Scenic Biotech; and a venture partner at Third Rock Ventures; all outside of the current work.

Received: October 20, 2023
Revised: April 22, 2024
Accepted: May 22, 2024
Published: June 11, 2024

REFERENCES

1. Tran, E., Robbins, P.F., and Rosenberg, S.A. (2017). Final common pathway' of human cancer immunotherapy: targeting random somatic mutations. Nat. Immunol. *18*, 255–262.
2. Gubin, M.M., Zhang, X., Schuster, H., Caron, E., Ward, J.P., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C.D., Krebber, W.J., et al. (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature *515*, 577–581.
3. Wells, D.K., van Buuren, M.M., Dang, K.K., Hubbard-Lucey, V.M., Sheehan, K.C.F., Campbell, K.M., Lamb, A., Ward, J.P., Sidney, J., Blazquez, A.B., et al. (2020). Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. Cell *183*, 818–834.e13.
4. Antoniou, A.N., Powis, S.J., and Elliott, T. (2003). Assembly and export of MHC class I peptide ligands. Curr. Opin. Immunol. *15*, 75–81.
5. Kloetzel, P.M. (2001). Antigen processing by the proteasome. Nat. Rev. Mol. Cell Biol. *2*, 179–187.
6. Rao, X., Costa, A.I.C.A.F., van Baarle, D., and Kesmir, C. (2009). A comparative study of HLA binding affinity and ligand diversity: implications for generating immunodominant CD8+ T cell responses. J. Immunol. *182*, 1526–1532.
7. Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. *48*, W449–W454.
8. Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S.L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. *12*, 1007–1017.
9. Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517.
10. Kisselev, A.F., Callard, A., and Goldberg, A.L. (2006). Importance of the different proteolytic sites of the proteasome and the efficacy of inhibitors varies with the protein substrate. J. Biol. Chem. *281*, 8582–8590.
11. Gomez-Perosanz, M., Ras-Carmona, A., Lafuente, E.M., and Reche, P.A. (2020). Identification of CD8+ T cell epitopes through proteasome cleavage site predictions. BMC Bioinf. *21*, 484.
12. Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics *57*, 33–41.
13. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. Mol. Cell. Proteomics *14*, 658–673.
14. Juncker, A.S., Larsen, M.V., Weinhold, N., Nielsen, M., Brunak, S., and Lund, O. (2009). Systematic characterisation of cellular localisation and

expression profiles of proteins containing MHC ligands. PLoS One *4*, e7448.

15. Milner, E., Barnea, E., Beer, I., and Admon, A. (2006). The turnover kinetics of major histocompatibility complex peptides of human cancer cells. Mol. Cell. Proteomics *5*, 357–365.

16. Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. Nat. Rev. Genet. *21*, 630–644.

17. Cuadrado, E., van den Biggelaar, M., de Kivit, S., Chen, Y.Y., Slot, M., Doubal, I., Meijer, A., van Lier, R.A.W., Borst, J., and Amsen, D. (2018). Proteomic Analyses of Human Regulatory T Cells Reveal Adaptations in Signaling Pathways that Protect Cellular Identity. Immunity *48*, 1046–1059.e6.

18. Nicolet, B.P., and Wolkers, M.C. (2020). *Limited but Gene-Class Specific Correlation of mRNA and Protein Expression in Human CD8* $^+$ *T Cells*. https://doi.org/10.1101/2020.04.21.053884. http://biorxiv.org/lookup/doi/10.1101/2020.04.21.053884.

19. Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

20. Vogel, C., Abreu, R.d.S., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Penalva, L.O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. *6*, 400.

21. Komander, D., and Rape, M. (2012). The ubiquitin code. Annu. Rev. Biochem. *81*, 203–229.

22. Narita, T., Weinert, B.T., and Choudhary, C. (2019). Functions and mechanisms of non-histone protein acetylation. Nat. Rev. Mol. Cell Biol. *20*, 156–174.

23. Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.

24. Andreatta, M., Alvarez, B., and Nielsen, M. (2017). GibbsCluster: unsupervised clustering and alignment of peptide sequences. Nucleic Acids Res. *45*, W458–W463.

25. Kaabinejadian, S., Barra, C., Alvarez, B., Yari, H., Hildebrand, W.H., and Nielsen, M. (2022). Accurate MHC Motif Deconvolution of Immunopeptidomics Data Reveals a Significant Contribution of DRB3, 4 and 5 to the Total DR Immunopeptidome. Front. Immunol. *13*, 835454.

26. Nicolet, B.P., Zandhuis, N.D., Lattanzio, V.M., and Wolkers, M.C. (2021). Sequence determinants as key regulators in gene expression of T cells. Immunol. Rev. *304*, 10–29.

27. Li, X., Xiong, X., Zhang, M., Wang, K., Chen, Y., Zhou, J., Mao, Y., Lv, J., Yi, D., Chen, X.W., et al. (2017). Base-Resolution Mapping Reveals Distinct m1A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. Mol. Cell *68*, 993–1005.e9.

28. Malbec, L., Zhang, T., Chen, Y.S., Zhang, Y., Sun, B.F., Shi, B.Y., Zhao, Y.L., Yang, Y., and Yang, Y.G. (2019). Dynamic methylome of internal mRNA N7-methylguanosine and its regulatory role in translation. Cell Res. *29*, 927–941.

29. Nicolet, B.P., Jurgens, A.P., Bresser, K., Guislain, A., and Wolkers, M.C. (2023). Learning the Sequence Code for mRNA and Protein Abundance in Human Immune Cells. https://doi.org/10.1101/2023.09.01.555843. http://biorxiv.org/lookup/doi/10.1101/2023.09.01.555843.

30. Karagkouni, D., Paraskevopoulou, M.D., Chatzopoulos, S., Vlachos, I.S., Tastsoglou, S., Kanellos, I., Papadimitriou, D., Kavakiotis, I., Maniou, S., Skoufos, G., et al. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. Nucleic Acids Res. *46*, D239–D245.

31. Caron, C., Boyault, C., and Khochbin, S. (2005). Regulatory cross-talk between lysine acetylation and ubiquitination: role in the control of protein stability. Bioessays *27*, 408–415.

32. Jeong, J.W., Bae, M.K., Ahn, M.Y., Kim, S.H., Sohn, T.K., Bae, M.H., Yoo, M.A., Song, E.J., Lee, K.J., and Kim, K.W. (2002). Regulation and destabilization of HIF-1alpha by ARD1-mediated acetylation. Cell *111*, 709–720.

33. Wilkinson, K.D., Tashayev, V.L., O'Connor, L.B., Larsen, C.N., Kasperek, E., and Pickart, C.M. (1995). Metabolism of the polyubiquitin degradation signal: structure, mechanism, and role of isopeptidase T. Biochemistry *34*, 14535–14546.

34. Mirihana Arachchilage, G., Hetti Arachchilage, M., Venkataraman, A., Piontkivska, H., and Basu, S. (2019). Stable G-quadruplex enabling sequences are selected against by the context-dependent codon bias. Gene *696*, 149–161.

35. Rao, Y.S., Wang, Z.F., Chai, X.W., Nie, Q.H., and Zhang, X.Q. (2013). Relationship between 5' UTR length and gene expression pattern in chicken. Genetica *141*, 311–318.

36. Abelin, J.G., Bergstrom, E.J., Rivera, K.D., Taylor, H.B., Klaeger, S., Xu, C., Verzani, E.K., Jackson White, C., Woldemichael, H.B., Virshup, M., et al. (2023). Workflow enabling deepscale immunopeptidome, proteome, ubiquitylome, phosphoproteome, and acetylome analyses of sample-limited tissues. Nat. Commun. *14*, 1851.

37. Zhang, L.-S., Liu, C., Ma, H., Dai, Q., Sun, H.L., Luo, G., Zhang, Z., Zhang, L., Hu, L., Dong, X., and He, C. (2019). Transcriptome-wide Mapping of Internal N7-Methylguanosine Methylome in Mammalian mRNA. Mol. Cell *74*, 1304–1316.e8.

38. Courel, M., Clément, Y., Bossevain, C., Foretek, D., Vidal Cruchez, O., Yi, Z., Bénard, M., Benassy, M.N., Kress, M., Vindry, C., et al. (2019). GC content shapes mRNA storage and decay in human cells. Elife *8*, e49708.

39. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM), pp. 785–794. https://doi.org/10.1145/2939672.2939785.

40. Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. Immunity *46*, 315–326.

41. Sarkizova, S., Klaeger, S., Le, P.M., Li, L.W., Oliveira, G., Keshishian, H., Hartigan, C.R., Zhang, W., Braun, D.A., Ligon, K.L., et al. (2020). A large peptidome dataset improves HLA class I epitope prediction across most of the human population. Nat. Biotechnol. *38*, 199–209.

42. Garcia Alvarez, H.M., Koşaloğlu-Yalçın, Z., Peters, B., and Nielsen, M. (2022). The role of antigen expression in shaping the repertoire of HLA presented ligands. iScience *25*, 104975.

43. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

44. Blevins, W.R., Tavella, T., Moro, S.G., Blasco-Moreno, B., Closa-Mosquera, A., Díez, J., Carey, L.B., and Albà, M.M. (2019). Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. Sci. Rep. *9*, 11005.

45. Koşaloğlu-Yalçın, Z., Lee, J., Greenbaum, J., Schoenberger, S.P., Miller, A., Kim, Y.J., Sette, A., Nielsen, M., and Peters, B. (2022). Combined assessment of MHC binding and antigen abundance improves T cell epitope predictions. iScience *25*, 103850.

46. Chong, C., Marino, F., Pak, H., Racle, J., Daniel, R.T., Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2018). High-throughput and Sensitive Immunopeptidomics Platform Reveals Profound Interferonγ--Mediated Remodeling of the Human Leukocyte Antigen (HLA) Ligandome. Mol. Cell. Proteomics *17*, 533–548.

47. Kraemer, A.I., Chong, C., Huber, F., Pak, H., Stevenson, B.J., Müller, M., Michaux, J., Altimiras, E.R., Rusakiewicz, S., Simó-Riudalbas, L., et al. (2023). The immunopeptidome landscape associated with T cell infiltration, inflammation and immune editing in lung cancer. Nat. Cancer *4*, 608–628.

48. Ouspenskaia, T., Law, T., Clauser, K.R., Klaeger, S., Sarkizova, S., Aguet, F., Li, B., Christian, E., Knisbacher, B.A., Le, P.M., et al. (2022). Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. Nat. Biotechnol. *40*, 209–217.

49. Wang, X., Zhao, B.S., Roundtree, I.A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. Cell *161*, 1388–1399.

50. Ravid, T., and Hochstrasser, M. (2008). Diversity of degradation signals in the ubiquitin-proteasome system. Nat. Rev. Mol. Cell Biol. *9*, 679–690.

51. Pang, C.N.I., Gasteiger, E., and Wilkins, M.R. (2010). Identification of arginine- and lysine-methylation in the proteome of Saccharomyces cerevisiae and its functional implications. BMC Genom. *11*, 92.

52. Lee, J.M., Lee, J.S., Kim, H., Kim, K., Park, H., Kim, J.Y., Lee, S.H., Kim, I.S., Kim, J., Lee, M., et al. (2012). EZH2 generates a methyl degron that is recognized by the DCAF1/DDB1/CUL4 E3 ubiquitin ligase complex. Mol. Cell *48*, 572–586.

53. Jarchum, I. (2018). Putting a number on neoepitope quality. Nat. Biotechnol. *36*, 151.

54. (2017). The problem with neoantigen prediction. Nat. Biotechnol. *35*, 97. https://doi.org/10.1038/nbt.3800.

55. Kemper, K., Krijgsman, O., Kong, X., Cornelissen-Steijger, P., Shahrabi, A., Weeber, F., van der Velden, D.L., Bleijerveld, O.B., Kuilman, T., Kluin, R.J.C., et al. (2016). BRAF(V600E) Kinase Domain Duplication Identified in Therapy-Refractory Melanoma Patient-Derived Xenografts. Cell Rep. *16*, 263–277.

56. Kelderman, S., Heemskerk, B., Fanchi, L., Philips, D., Toebes, M., Kvistborg, P., van Buuren, M.M., van Rooij, N., Michels, S., Germeroth, L., et al. (2016). Antigen-specific TIL therapy for melanoma: A flexible platform for personalized cancer immunotherapy. Eur. J. Immunol. *46*, 1351–1360.

57. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

58. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

59. Wang, H., McManus, J., and Kingsford, C. (2016). Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. Bioinformatics *32*, 1880–1882.

60. Elek, A., Kuzman, M., and Vlahovicek, K. (2018). coRdon: Codon Usage Analysis and Prediction of Gene Expressivity. https://doi.org/10.18129/B9.bioc.coRdon.

61. Marino, F., Mommen, G.P.M., Jeko, A., Meiring, H.D., van Gaans-van den Brink, J.A.M., Scheltema, R.A., van Els, C.A.C.M., and Heck, A.J.R. (2017). Arginine (Di)methylated Human Leukocyte Antigen Class I Peptides Are Favorably Presented by HLA-B*07. J. Proteome Res. *16*, 34–44.

62. Giudice, G., Sánchez-Cabo, F., Torroja, C., and Lara-Pezzi, E. (2016). ATtRACT—a database of RNA-binding proteins and associated motifs. Database *2016*, baw035. https://doi.org/10.1093/database/baw035.

63. Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. *48*, D127–D131.

64. Juzenas, S., Venkatesh, G., Hübenthal, M., Hoeppner, M.P., Du, Z.G., Paulsen, M., Rosenstiel, P., Senger, P., Hofmann-Apitius, M., Keller, A., et al. (2017). A comprehensive, cell specific microRNA catalogue of human peripheral blood. Nucleic Acids Res. *45*, 9290–9301.

65. Luo, X., Li, H., Liang, J., Zhao, Q., Xie, Y., Ren, J., and Zuo, Z. (2021). RMVar: an updated database of functional variants involved in RNA modifications. Nucleic Acids Res. *49*, D1405–D1412.

66. Huang, K.-Y., Lee, T.Y., Kao, H.J., Ma, C.T., Lee, C.C., Lin, T.H., Chang, W.C., and Huang, H.D. (2019). dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. Nucleic Acids Res. *47*, D298–D308.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
mouse monoclonal IgG2a antibody W6/32	In house; Heck Lab	N/A
Chemicals, peptides, and recombinant proteins		
Cycloheximide	Sigma-Aldrich	Cat#C7698-1G
Ambion^{TM} RNase I	Invitrogen	Cat#AM2294
proteinase K	Merck	Cat#3115836001
TRIzol reagent	Invitrogen	Cat# 15596026
T4 RNA ligase 1	NEB	Cat# M0204S
T4 polynucleotide kinase	NEB	Cat# M0201S
Critical commercial assays		
RNeasy Mini Kit	Qiagen	Cat#74104
TruSeq Stranded mRNA Library Prep	Illumina	Cat#20020595
Deposited data		
mRNA sequencing data melanoma lines	This paper	GEO: GSE210999
Ribosome profiling data melanoma lines	This paper	GEO: GSE210998
Source data and analyses	This paper	Zenodo: https://doi.org/10.5281/zenodo.11151263
Experimental models: Cell lines		
SK-MEL-95	Memorial Sloan Kettering Cancer Center	RRID:CVCL_6064
M026.X1	Netherlands Cancer Institute, Daniel Peeper Group	Ref. [55]
NKIRTIL006	Netherlands Cancer Institute, Ton Schumacher Group	Ref. [56]
Software and algorithms		
STAR aligner	Ref. [57]	https://github.com/alexdobin/STAR
Salmon	Ref. [58]	https://github.com/COMBINE-lab/salmon
Ribomap	Ref. [59]	https://github.com/Kingsford-Group/ribomap
Proteome Discoverer 1.4	Thermo Fisher	https://www.thermofisher.com/nl/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html
netMHCpan 4.1	Ref. [7]	https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/
coRdon	Ref. [60]	https://github.com/BioinfoHR/coRdon

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ton Schumacher (t.schumacher@nki.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Transcriptomic and ribosome profiling data presented in this manuscript have been deposited to GEO and can be accessed under the accession number GSE211000. Mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD036277. All statistical source data of the figures presented in this study are provided with this paper. Transcriptomic data of the Sarkizova study[41] was accessed from GEO under the accession number GSE131267. HLA ligands from the Sarkizova study[41] were downloaded from the publisher's website.
- This paper does not report original code.
- Any information required to reanalyze the data reported in this work, and source data underlying the figures, have been uploaded to Zenodo: https://doi.org/10.5281/zenodo.11151263. All scripts used to perform the analyses included in this manuscript have been uploaded to GitHub: https://github.com/kasbress/HLA_Ligandome_Analyses. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Patient-derived melanoma cell lines

Patient-derived melanoma cell lines were cultured in RPMI (Gibco) supplemented with 8% fetal calf serum (FCS, Sigma), 100 U/mL penicillin (Gibco) and 100 μg/mL streptomycin (Gibco) at 37°C and 5% CO_2. SK-MEL-95 and M026.X1[55] were a kind gift from Daniel Peeper (Netherlands Cancer Institute). SK-MEL-95 was originally established in the Memorial Sloan Kettering Cancer Center (RRID: CVCL_6064). M026.X1 was originally established in the lab of Daniel Peeper as a xenograft-derived melanoma cell line. NKIRTIL006 was established in house.[56]

METHOD DETAILS

Patient-derived melanoma cell lines

SK-MEL-95 and M026.X1[55] were a kind gift from Daniel Peeper (Netherlands Cancer Institute). SK-MEL-95 was originally established in the Memorial Sloan Kettering Cancer Center. NKIRTIL006 was established in house.[56]

Cell culture

Patient-derived melanoma cell lines were cultured in RPMI (Gibco) supplemented with 8% fetal calf serum (FCS, Sigma), 100 U/mL penicillin (Gibco) and 100 μg/mL streptomycin (Gibco) at 37°C and 5% CO_2. For mRNA sequencing and ribosome profiling, cell lines were cultured to a density of 70–90% on 150mm Corning tissue-culture treated culture dishes (Merck). For HLA ligandome LC-MS, approximately $1 \cdot 10^9$ cells were cultured in Corning CELLSTACK Culture Chambers (Corning, 05-539-096).

HLA class I peptide isolation and LC-MS/MS

HLA class I-associated peptides were isolated by immunoprecipitation of HLA class I complexes using the mouse monoclonal IgG2a antibody W6/32, as described previously.[61] Peptides were eluted from HLA class I protein molecules using a 10% acetic acid (v/v) solution, and subsequently separated using a 10 kDa molecular weight cutoff filter. Obtained solution was then desalted into 3 fractions using in-house made c18 STAGE (STop And Go Extraction) tips, eluted with 20%, 30% and 50% acetonitrile, respectively. The resulting fractions were injected on an Agilent 1290 system using a 120-min gradient coupled to an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific). Fractions 1 and 2 were injected in triplicate, whereas fraction 3 was injected in duplicate. The LC system comprised of a 20 × 0.1 mm i.d. trapping column (Reprosil C18, 3 μm; Dr. Maisch) and a 50 × 0.005 cm i.d. analytical column (Poroshell 120 EC-C18; 2.7 μm). An LC resolving gradient of 13–43% Solvent B (80% acetonitrile, 20% water, 0.1% formic acid) was used. The Top Speed method was enabled for fragmentation, where the most abundant precursor ions were selected in a 3 s cycle for data-dependent EThcD. MS1 and MS2 spectra were acquired at a resolution of 60,000 (FWHM at 400 *m/z*) and 15,000 (FWHM at 400 *m/z*), respectively. RF lens voltage was set to 60. Dynamic exclusion of 18s was used. Peptide precursors of charges 2 to 6 were fragmented, if accumulated to a minimum intensity of $4 \cdot 10^5$ within 50 ms. In MS2, a maximum injection time of 250ms was used with a minimum intensity filter of $5 \cdot 10^4$.

HLA class I peptide analysis

RAW data files were analyzed using the Proteome Discoverer 1.4 software package (Thermo Fisher Scientific). MS/MS scans were searched against the human Swissprot reviewed database (accessed in September 2015; 20,203 entries), with no enzyme specificity using the SEQUEST HT search engine. Precursor ion and MS/MS tolerances were set to 10 ppm and 0.05 Da. Methionine oxidation was set as variable modification. The peptides-to-spectrum matches were filtered for precursor tolerance 5 ppm, <1% FDR using Percolator, XCorr >1.7, and peptide rank 1. Peptides that were between 8 and 14 amino acid long were selected for further analysis. The mass spectrometry data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD036277.

Replicate injections displayed an overlap of approximately 70% (shared between at least 2 replicates). Consistent with their shared tissue origin, a large part of peptides detected across the melanoma lines mapped to a core group of proteins (47.6% shared between at least 2 lines). In contrast, the MS detected peptides exhibited a small degree of overlap (12.4% shared between at least 2 lines), in line with their difference in HLA haplotype.

mRNA sequencing

Cells were cultured to an approximate density of 80%, and $1 \cdot 10^7$ cells were subsequently dissociated using a cell-scraper in cold (4°C) PBS, centrifuged for 10 min at 300× *g*, and snap-frozen in liquid nitrogen. RNA was extracted from the frozen pellets using the RNeasy Mini Kit (Qiagen). Whole transcriptome sequencing samples were prepared using the TruSeq Stranded mRNA Kit (Illumina). Single-end 65 bp sequencing was performed on a HiSeq 2500 System (Illumina). Obtained reads were aligned to the GRCh38 reference (gencode release 21) using STAR aligner[57] (version 2.5.2b), and transcripts were quantified using Salmon[58] (version 0.7.0). Transcript counts belonging to a single consensus coding sequence were summed.

Ribosome profiling

Cells were cultured to an approximate density of 80%, and $5 \cdot 10^7$ cells were subsequently treated with 100 μg/mL cycloheximide for 5 min at 37°C. Cells were then washed once in cold (4°C) PBS containing 100 μg/mL cycloheximide, dissociated using a cell-scraper in cold (4°C) PBS supplemented with 100 μg/mL cycloheximide, centrifuged for 10 min at 300× *g*, and snap-frozen in liquid nitrogen. Frozen pellets were resuspended in lysis buffer (20 mM Tris–HCl, pH 7.8, 100 mM KCl, 10 mM $MgCl_2$, 1% Triton X-100, 2 mM DTT, 100 μg/mL cycloheximide, 1× Complete protease inhibitor), and incubated on ice for 20 min. Lysates were sheared using a 26G needle, centrifuged for 10 min at 1,300× *g*, and supernatants were transferred to a clean tube. Supernatants were treated with 2 U/μl of RNase I (Ambion) for 45 min at room temperature, with rotation. Next, lysates were fractionated on a linear sucrose gradient (7%–47%) using the SW-41Ti rotor (Beckman Coulter) at 221,633× *g* for 2 h at 4°C, without brake. Obtained sucrose gradients were then divided in 14 fractions, and fractions 7–10 (cytosolic ribosomes) were pooled and treated with PCR grade proteinase K (Roche) in 1% SDS to release ribosome protected fragments. The resulting fragments were subsequently purified using Trizol reagent (Invitrogen) and precipitated in the presence of glycogen, following the manufacturer's instructions. For library preparation, RNA was gel-purified on a denaturing 10% polyacrylamide urea (7 M) gel. A section corresponding to 25 to 36 nucleotides—the region that comprises the majority of the ribosome-protected RNA fragments—was excised, and purified through ethanol precipitation. RNA fragments were then 3′-dephosphorylated using T4 polynucleotide kinase (New England Biolabs) for 6 h at 37°C in 2-(N-morpholino)ethanesulfonic acid (MES) buffer (100 mM MES-NaOH pH 5.5, 10 mM $MgCl_2$, 10 mM β-mercaptoethanol, 300 mM NaCl). The 3′ adaptor was added using T4 RNA ligase 1 (New England Biolabs) for 2.5 h at 37°C. Ligation products were 5′-phosphorylated with T4 polynucleotide kinase for 30 min at 37°C, and the 5′ adaptor was added using T4 RNA ligase 1 for 2 h at 37°C. Sequencing was performed on a HiSeq 2500 System (Illumina). Ribosome occupancy was calculated using the Ribomap pipeline,[59] and was aligned to the GRCh38 reference (gencode release 21). Counts belonging to a single consensus coding sequence were summed.

Characterization of LC-MS detected peptides

For comparison of peptide length distributions, known melanoma HLA class I ligands were downloaded from the IEDB web-interface (https://www.iedb.org) in June 2021 using the following search filters: Epitope – Any; Assay Outcome – Positive; MHC restriction – Class I; Host – Human; Disease – Melanoma.

To assess the amino acid positional biases of the LC-MS detected peptides, the dataset was filtered for 9-meric species, and the occurrence of each amino acid on each peptide position was tallied. As a reference, all expressed proteins (TPM >0 in the mRNA-seq dataset) were selected for each melanoma line, and the number occurrences of each amino acid was calculated. Amino acid enrichment was then defined as the fraction by which an amino acid occurred at a certain position divided by the fraction by which that amino acid occurred in the reference. The positional bias was defined as the median of the absolute amino acid enrichment values for each peptide position.

For binding motif analyses, 9-meric peptide sequences from each melanoma line were clustered using GibbsCluster 2.0 (command line options set to: -g 3–7 -C -D 4 -I 1 -S 5 -T -j 2 -c 1 -k 25), with the number of clusters for each melanoma line set to the number of alleles expressed by that line. Sequence logos were generated using the R package ggseqlogo. To generate reference sequence logos, all known human 9mers for each of the shown HLA class I alleles were downloaded from IEDB in June 2021.

Peptide database construction

To investigate characteristics of HLA class I ligands, a database consisting of LC-MS detected peptides (i.e., true HLA ligands) and not-detected peptides (referred to as decoy peptides) was constructed. To this end, binding scores for the HLA class I alleles of each melanoma line were calculated for all 9-, 10-, and 11-mers that were detected by LC-MS using netMHCpan 4.1,[7] and to each peptide, the highest affinity rank score for the expressed HLA alleles was subsequently assigned. Separate databases were generated for each melanoma line assigning each peptide the expression level (TPM) of its source protein. Each database was supplemented with 'decoy peptides' at the indicated ratios, with decoy peptides being randomly sampled from the human proteome, at a length distribution that was equal to the set of LC-MS detected peptides.

Feature library construction

5′ UTR, coding region (CDS) and 3′ UTR nucleotide sequences were downloaded from ENSEMBL BiomaRt (release 104; accessed September 2021) for all protein-coding transcripts. RNA-binding protein motifs were acquired from ATtracT[62] (accessed June 2021) and filtered for human RBPs (142 RBPs; 2,271 motifs). In each transcript region (e.g., 5′ UTR, CDS, 3′ UTR), motifs were searched and counted using a custom script (see GitHub project), and GC content and nucleotide length were computed. Also included in the sequence feature library were: Codon usage (applying coRdon[60]), amino acid usage within the CDS, miR-DB[63] miRNA seed scores (accessed August 2021 and filtered for miRNA expressed immune cells, based on previous analysis by Juzenas et al.[64]), sequence homology between Human and Zebrafish (*Danio rerio*, obtained through Ensembl BiomaRt), predicted mRNA modification site occurrence per transcript region (obtained from the RMVar database,[65] accessed at https://rmvar.renlab.org/in September 2021), and predicted post-translational modification (Acetylation, Amidation, Hydroxylation, Malonylation, Methylation, N-linked_Glycosylation, O-linked_Glycosylation, Palmitoylation, Phosphorylation, S-nitrosylation, Succinylation, Sumoylation, Ubiquitination) site occurrence (obtained from the dbPTM database,[66] accessed at https://awi.cuhk.edu.cn/dbPTM/in February 2024).

Importance assessment of sequence feature classes

To assess the ability of sequence features to inform on HLA sampling, features belonging to five major classes (5′ UTR, CDS, 3′ UTR, miR binding and PTM) were extracted from the sequence feature library. The 5′ UTR, CDS, 3′ UTR classes were filtered based on their variance across the proteome using the nearZeroVar function in the caret R package (setting cutoffs at: freqRatio <500 and percentUnique >0.05). All putative miR binding sites and PTMs in the library were used in the analysis. The number of features left after filtering are shown in Figure 2A. 2,000 true HLA ligands and 4,000 decoy peptides were sampled from the peptide database of each melanoma line, and subsequently used to train individual Random Forest models for each melanoma line and each feature class to predict true HLA ligands (15 models in total). The Random Forest models were generated using the R packages randomForest and caret, using 10-fold cross validation optimizing the ROC metric. Number of trees in each forest was set to 5,000 and minimum terminal node size was set to 2. The mtry parameter was set to $\sqrt[2]{n\ features - 1} \times 1.5$. Feature importance (i.e., mean decrease in accuracy) was calculated using the varImp function from the R package caret.

Analyses examining HLA ligand enrichment potential of individual sequence features (Figures 2D–2F and S2B) were focused on the 10 most important features in each class (defined as the highest mean importance score of the models trained for that feature class), and were performed using 3,389 true HLA ligands and 13,556 decoy peptides per tumor line. For the analysis presented in Figures 2E and 2F, A custom enrichment metric was calculated. In brief, 30% of the data was sampled and peptides were ranked either by the occurrence of a sequence feature or at random. In both cases the total number of true HLA ligands within the top 50% ranked peptides was tallied. Next, the percentage increase in true HLA ligands was calculated comparing the sequence feature ranked case versus the randomly ranked case. This process was performed for all sequence features in the analyses, and was repeated 50 times.

XGBoost classifiers

The number of experimentally detected HLA ligands from each melanoma line was down-sampled to the number of HLA ligands in the smallest dataset to ensure each melanoma line had equal weight during the analyses. LC-MS detected peptides were randomly split into a training (80%) and a test (20%) set, and these sets were subsequently supplemented with a 4-fold and 1,000-fold excess of decoy peptides. XGBoost models were generated using the R packages xgboost and caret, using 2-times 10-fold cross validation optimizing the accuracy metric. Learning rate was set to 0.3, minimum loss reduction was set to 1.0, maximum tree depth was set to 1, sub-sampling ratio of features for each tree was set to 0.5, minimum sum of instance weight needed in a terminal leaf was set to 0.9, number of rounds was set to 1,000.

External HLA ligandome data

Transcriptomic data of the Sarkizova study[41] was accessed from the Gene Expression Omnibus (GEO) at GSE131267 and was aligned to the GRCh38 reference (gencode release 21) using Salmon (quasi-mapping mode, version 0.7.0). Mean transcript counts were calculated between replicates, and transcripts belonging to a single consensus coding sequence were summed. HLA ligands from the Sarkizova study[41] were downloaded from the publisher's website. 350 LC-MS detected peptides (comprising 9-, 10- and 11-mers) were sampled from each of the 20 indicated mono-allelic cell lines. HLA ligandome data obtained from lung adenocarcinoma,[47] meningeal cancer lines, EBV immortalized B cells and tumor-infiltrating T cells, IFNγ treated ovarian carcinoma lines,[46] and HLA ligands derived from non-canonical open-reading-frames found in B721.221 cells[48] were downloaded from the respective publisher's websites. Each of these datasets was down-sampled to 500–1,000 LC-MS detected peptides. A 1,000-fold excess of length-matched decoy peptides randomly sampled from the human proteome was added to each dataset. Predicted HLA-peptide affinity scores were calculated using netMHCpan 4.1[7] for all expressed HLA alleles, and to each peptide the highest affinity rank score for the expressed HLA alleles was subsequently assigned.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were performed using the rstatix package in the R programming language. All applied statistical tests, *p* values, and confidence invervals are reported in the figures or figure legends.