Hindawi
BioMed Research International
Volume 2020, Article ID 5798356, 12 pages
https://doi.org/10.1155/2020/5798356

Research Article

INeo-Epp: A Novel T-Cell HLA Class-I Immunogenicity or Neoantigenic Epitope Prediction Method Based on Sequence-Related Amino Acid Features

Guangzhi Wang,[1,2] **Huihui Wan**,[2,3] **Xingxing Jian**,[2,4] **Yuyu Li**,[1] **Jian Ouyang**,[2] **Xiaoxiu Tan**,[3] **Yong Zhao**,[1] **Yong Lin**,[3] **and Lu Xie**[1,2]

[1]*College of Food Science and Technology, Shanghai Ocean University, Shanghai 201306, China*
[2]*Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai 201203, China*
[3]*School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*
[4]*Key Laboratory of Carcinogenesis and Cancer Invasion, Ministry of Education and Key Laboratory of Carcinogenesis, National Health and Family Planning Commission, Xiangya Hospital, Central South University, Changsha 410008, China*

Correspondence should be addressed to Yong Zhao; yzhao@shou.edu.cn, Yong Lin; yong_lynn@163.com, and Lu Xie; luxiex2017@outlook.com

Received 26 March 2020; Accepted 23 May 2020; Published 16 June 2020

Guest Editor: Zhenguo Zhang

In silico T-cell epitope prediction plays an important role in immunization experimental design and vaccine preparation. Currently, most epitope prediction research focuses on peptide processing and presentation, e.g., proteasomal cleavage, transporter associated with antigen processing (TAP), and major histocompatibility complex (MHC) combination. To date, however, the mechanism for the immunogenicity of epitopes remains unclear. It is generally agreed upon that T-cell immunogenicity may be influenced by the foreignness, accessibility, molecular weight, molecular structure, molecular conformation, chemical properties, and physical properties of target peptides to different degrees. In this work, we tried to combine these factors. Firstly, we collected significant experimental HLA-I T-cell immunogenic peptide data, as well as the potential immunogenic amino acid properties. Several characteristics were extracted, including the amino acid physicochemical property of the epitope sequence, peptide entropy, eluted ligand likelihood percentile rank (EL rank(%)) score, and frequency score for an immunogenic peptide. Subsequently, a random forest classifier for T-cell immunogenic HLA-I presenting antigen epitopes and neoantigens was constructed. The classification results for the antigen epitopes outperformed the previous research (the optimal AUC = 0.81, external validation data set AUC = 0.77). As mutational epitopes generated by the coding region contain only the alterations of one or two amino acids, we assume that these characteristics might also be applied to the classification of the endogenic mutational neoepitopes also called "neoantigens." Based on mutation information and sequence-related amino acid characteristics, a prediction model of a neoantigen was established as well (the optimal AUC = 0.78). Further, an easy-to-use web-based tool "INeo-Epp" was developed for the prediction of human immunogenic antigen epitopes and neoantigen epitopes.

1. Introduction

An antigen consists of several epitopes, which can be recognized either by B- or T-cells and/or molecules of the host immune system. However, usually only a small number of amino acid residues that comprise a specific epitope are necessary to elicit an immune response [1]. The properties of these amino acid residues causing immunogenicity are unknown. HLA-I antigen peptides are processed and presented as follows: (a) cytosolic and nuclear proteins are cleaved to short peptides by intracellular proteinases; (b) some are selectively transferred to the endoplasmic reticulum (ER) by the TAP transporter, and subsequently are treated by endoplasmic reticulum aminopeptidase; and (c) antigen-presenting cells (APCs) present peptides containing 8-11 AA (amino acid) residues on HLA class I

molecules to CD8+ T-cells [2]. Researchers can now simulate antigen processing and presentation by computational methods to predict binding peptide-MHC complexes (p-MHC). Several types of software systems have been developed, including NetChop [3], NetCTL [4], NetMHCpan [5], and MHCflurry [6]. However, despite that the binding to MHC molecules of most peptides is predicted, only 10%~15% of those have been shown to be immunogenic [7–10]. For neoantigens, the result was approximately 5% (range: 1%-20%) due to central immunotolerance [11, 12]. As a result, the cycle for vaccine development and immunization research is extended. Here, we aim to develop a T-cell HLA class-I immunogenicity prediction method to further identify real epitopes/neoepitopes from p-MHC to shorten this cycle.

Many experimental human epitopes have been collected and summarized in the immune epitope database (IEDB) [13], which makes it feasible to mathematically predict human epitopes. However, there still exist two limitations: (i) a high level of MHC polymorphism produces a severe challenge for T-cell epitope prediction and (ii) there is an extremely unequal distribution of data to compare epitopes and nonepitopes. It is not conducive to analyze the potential deviation existing in TCR recognition owing to the presentation of different HLA peptides. A general analysis of all HLA-presented peptides, ignoring the specific pattern of TCR recognition of individual HLA-presented peptides, may result in a lower predictive accuracy.

With the advances in HLA research, Sette and Sidney [14] classified, for the first time, overlapping peptide binding repertoires into nine major functional HLA supertypes (A1, A2, A3, A24, B7, B27, B44, B58, and B62). In 2008, Sidney et al. [15] made a further refinement, in which over 80% of the 945 different HLA-A and B alleles can be assigned to the original nine supertypes. It has not been reported whether peptides presented by different HLA alleles influence TCR recognition. Hence, we collected experimental epitopes according to HLA alleles and assumed that epitopes belonging to the same HLA supertypes have similar properties.

Moreover, screening for endogenic mutational neoepitopes is one of the core steps in tumor immunotherapy. In 2017, Ott et al. [16] and Sahin et al. [17] confirmed that peptides and RNA vaccines made up of neoantigens in melanoma can stimulate and proliferate CD8+ and CD4+ T-cells. In addition, a recent research suggests that including neoantigen vaccination not only can expand the existing specific T-cells but also can induce a wide range of novel T-cell specificity in cancer patients and enhance tumor suppression [18]. Meanwhile, a tumor can be better controlled by the combination therapy of neoantigen vaccine and programmed cell death protein 1 (PD-1)/PD1 ligand 1 (PDL-1) therapy [19, 20]. Nevertheless, a considerable number of predicted candidate p-MHC from somatic cell mutations may be false positive, which would fail to stimulate TCR recognition and immune response. This is undoubtedly a challenge for designing vaccines against neoantigens.

In our study, based on HLA-I T-cell peptides collected from experimentally validated antigen epitopes and neoantigen epitopes, we aim to build a novel method to further reduce the range of immunogenic epitope screening based on predicted p-MHC. Finally, a simple web-based tool, INeo-Epp (immunogenic epitope/neoepitope prediction), was developed for prediction of human antigen and neoantigen epitopes.

2. Materials and Methods

The flow chart for "INeo-Epp" prediction is shown in Figure 1.

2.1. Construction of Immunogenic and Nonimmunogenic Epitopes. Peptides that can promote cytokine proliferation are considered to be immunogenic epitopes. However, nonimmunogenic epitopes may result from the following reasons: (a) p-MHC is truly unrecognized by TCR, (b) peptides are not presented by MHC (quantitatively expressed as rank (%) > 2, see Rank(%) Score (C24) for details), and (c) negative selection/clonal presentation is induced by excessive similarity to autologous peptides [21]. In this work, to further study the recognition preferences of T-cells, peptides with >2 rank(%) were regarded as not in contact with TCR, and sequences 100% matching the human reference peptides (ftp://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/pep/) were regarded as exhibiting immune tolerance. Hence, we removed these from the definition of nonimmunogenic peptides.

2.2. Construction of Data Sets: Epitopes, External Validation of Epitopes, and Neoepitopes. Antigen epitope data were collected from IEDB (linear epitope, human, T-cell assays, MHC class I, any disease was chosen). Data collection criteria accommodated for each HLA allele quantity > 50 and frequency > 0.5% (refer to allele frequency database [22]) (Table 1, check Table S1 for detailed information).

The external antigen epitope validation set was collected from seven published independent human antigen studies [23–29], consisting of 577 nonimmunogenic epitopes and 85 immunogenic epitopes (Table 2, S2 Table).

Here, we removed peptides for which HLA supertypes do not appear in the training set, because we assume peptides belonging to the same HLA supertypes to have similar properties. In the external validation set, some peptides bind to rare HLA supertypes. Their characteristics were not included in the training set. Hence, these peptides in the external validation data might lead to a classification bias.

The neoantigen data were collected from 11 publications [19, 30–39] and IEDB mutational epitopes, and 13 published data sets collected by Bjerregaard et al. in one publication [40] in 2017 (see Table 3, S3 Table for details) were also included.

2.3. Construction of Potential Immunogenicity Feature

2.3.1. Calculation of Peptide Characteristics Based on Amino Acid Sequences. The formula for calculating peptide characteristics is shown in (1). P_N, P_2, and P_C (N-terminal, position 2, C-terminal as anchored sites by default) are considered to

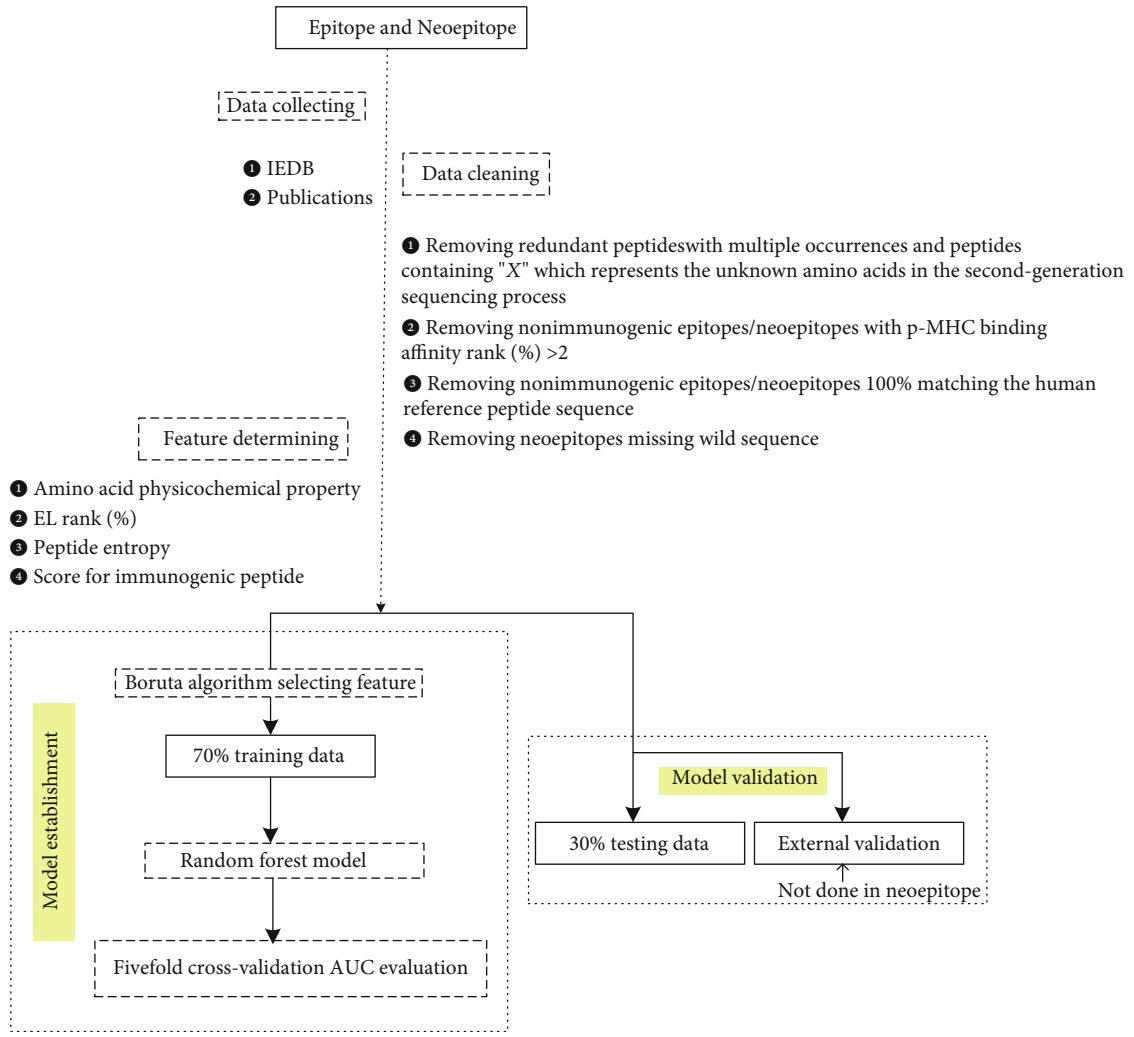


Figure 1: The flow chart for "INeo-Epp" prediction.

be embedded in HLA molecules and have no contact with TCRs; therefore, they were not evaluated.

$$Pc = \left\{ \sum_{\substack{x \notin (N,2,C) \\ x \in Pos(P)}} P_{A_C} \right\} / (len(P) - 3) \quad (1)$$

where P is peptide, c is characteristic. P_c represents the characteristics of peptides, A represents amino acids, N represents the N-terminal in a peptide, C represents the C-terminal in a peptide, Pos represents the amino acid position in a peptide, and P_{A_c} represents characteristics of amino acids in peptides.

2.3.2. Frequency Score for Immunogenic Peptide (C22). Amino acid distribution frequency differences between immunogenic and nonimmunogenic peptides at TCR contact sites (excluding anchor sites) were considered as a feature:

$$P_{score} = \sum_{\substack{x \notin (N,2,C) \\ x \in Pos(P)}} \left\{ P_{ie} + \left(f'_A \right) - P_{ie} - \left(f'_A \right) \right\} \quad (2)$$

where P^{+}_{ie} represents immunogenic peptides, P^{-}_{ie} represents nonimmunogenic peptides. f'_A represents amino acid frequency in the TCR contact position. $P^{+}_{ie}(f'_A)$ represents the frequency of amino acids in immunogenic peptides at TCR contact sites.

2.3.3. Calculating Peptide Entropy (C23). Peptide entropy [41] was used as a feature:

$$P_H = \left\{ -\sum_{\substack{x \notin (N,2,C) \\ x \in Pos(P)}} P_{f_A} * \log_2 \left(P_{f_A} \right) \right\} / (len(P) - 3) \quad (3)$$

where P_H represents peptide entropy. f_A represents amino acid frequency in the human reference peptide sequence. P_{f_A} represents the frequency in the human reference peptide sequence of amino acids in epitope peptides.

2.3.4. Rank(%) Score (C24). HLA binding prediction was performed using NetMHCpan 4.0. Rank(%) provides a robust filter for the identification of MHC-binding peptides, in

Table 1: Summary of IEDB epitope data.

HLA supertype	IEDB HLA data	Number		HLA allele frequency	Motif view
		Negative	Positive	Asian/Black/Caucasian	
A1	A01:01	811	103	0.154/0.046/0.164	1-2(ST)-3-4-5-6-7-8-9(Y)
	A26:01	83	19	0.041/0.014/0.030	1(DE)-2(ITV)-3-4-5-6-7-8-9(FMY)
A2	A02:01	1883	1580	0.049/0.123/0.275	1-2(LM)-3-4-5-6-7-8-9(ILV)-10(V)
A3	A11:01	196	174	0.139/0.014/0.060	1-2(IMSTV)-3-4-5-6-7-8-9(K)-10(K)
	A03:01	1400	169	0.063/0.083/0.139	1-2(ILMTV)-3-4-5-6-7-8-9(K)-10(K)
A24	A24:02	207	219	0.136/0.024/0.084	1-2(WY)-3-4-5-6-7-8-9(FIW)
	A23:01	1138	12	0.006/0.109/0.019	1-2(WY)-3-4-5-6-7-8-9-10(F)
B7	B35:01	63	248	0.062/0.068/0.055	1-2(P)-3-4-5-6-7-8-9(FMY)
	B07:02	523	244	0.034/0.005/0.0143	1-2(p)-3-4-5-6-7-8-9(FLM)
	B51:01	13	51	0.074/0.021/0.047	1-2(P)-3-4-5-6-7-8-9(IV)
B8	B08:01	317	195	0.036/0.037/0.114	1-2-3-4-5(HKR)-6-7-8-9(FILMV)
B27	B27:05	100	86	0.008/0.008/0.037	1(RY)-2(R)-3(FMLWY)-4-5-6-7-8-9
B44	B37:01	1036	10	0.034/0.005/0.014	—
	B40:01	67	65	0.022/0.012/0.052	—
	B44:02	73	66	0.008/0.020/0.095	1-2(E)-3-4-5-6-7-8-9(FIWY)
B58	B58:01	11	62	0.041/0.037/0.007	1-2(AST)-3-4-5-6-7-8-9(W)
B62	B15:01	3	70	0.016/0.010/0.060	1-2(LMQ)-3-4-5-6-7-8-9(FY)
Total		7924	3373		
Remove negative rank (%) > 2		5123	3373		
Remove negative human 100% similar		4943	3373		

Table 2: External data included in validation set.

Publication time	PMID	Author	Nonepitopes	Epitopes
2013	23580623	Weiskopf et al.	477	42
2018	29397015	Luxenburger et al.	100	26
2018	30260541	Xia et al.	—	1
2018	30487281	Vahed et al.	—	4
2018	30518652	Khakpoor et al.	—	2
2018	30587531	Huth et al.	—	4
2018	30815394	Sekyere et al.	—	6
Total			577	85
Remove negative with rank (%) > 2 and HLA supertypes (not appeared in training set)			321	69

which rank(%) was recommended as an evaluation standard, rank (%) < 0.5 as strong binders, 0.5 < rank (%) < 2 as weak binders, and rank (%) > 2 as no binders.

2.4. Fivefold Cross-Validation, Feature Selection, Random Forests, and ROC Generation. The 5-fold cross-validation was implemented in R using the caret package [42] (method = "repeatedcv," number = 5, repeats = 3). The feature screening results were generated in R using the package Boruta [43] (a novel random forest-based feature selection algorithm for finding all relevant variables, which provides unbiased and stable selection of important and nonimportant attributes from an information system). It iteratively removes the features which are proven by a statistical test to be less relevant than random probes. It uses Z score (computed by dividing the average loss by its standard deviation) as the importance measure, and it takes into account the fluctuations of the mean accuracy loss among trees in the forest. R package randomForest [44] was used for training data (the R language machine learning package caret provides automatic iteration selection of optimal parameters: mtry = 15 for antigen epitope and mtry = 14 for neoantigen epitope; the remaining parameters use default values). R package ROCR [45] was used for drawing ROC.

2.5. Web Tool Implementation. The front end of Ineo-Epp was constructed via HTML/JavaScript/CSS. The back end was written in PHP, connecting the web interface and

Table 3: Neoepitope data included in this study.

Publication time	PMID	Author	Tumor type	Nonimmunogenic neoepitopes	Immunogenic neoepitopes	T-cell assay
2013-12	24323902	D. A. Wick et al.	Ovarian cancer	—	1	ELISPOT
2015-9	26359337	E. M. Van Allen et al.	Melanoma	—	18	Clinical benefit
2015-11	26752676	T. Karasaki et al.	Lung adenocarcinoma	—	4	—
2016-1	26901407	A. Gros et al.	Melanoma	12	14	ELISPOT
2016-5	27198675	E. Strønen, et al.	Melanoma	1134	16	CTL clone
2016-12	28405493	A. Nelde et al.	Lymphoma	—	2	ELISPOT
2017-6	28619968	X. Zhang et al.	Breast cancer	—	4	Flow cytometry
2017-10	29104575	M. Markus et al.	Melanoma	10	16	—
2017-11	29187854	A.-M. Bjerregaard et al.	Polytype	1874	42	ELISPOT et al.
2017-11	29132146	V. P. Balachandran et al.	Pancreatic	—	10	Flow cytometry
2018-5	29720506	T. Matsuda et al.	Ovarian cancer	—	3	ELISPOT
2018-12	29409514	K. Sonntag et al.	Pancreatic ductal carcinoma	—	3	Flow cytometry
2018-10	30357391	V. Randi et al.	—	6	35	—
Total				3030	168	
Remove duplication				2837	164	
Remove negative rank (%) > 2 and human 100% similar				1697	164	

Apache web server. A python script was used for calculating peptide characteristics and extracting mutation information. Models were built using R.

3. Results

Ultimately, 11,297 validated epitopes and nonepitopes with lengths of 8-11 amino acids were collected from IEDB. T-cell responses included activation, cytotoxicity, proliferation, IFN-γ release, TNF release, granzyme B release, IL-2 release, and IL-10 release. Seventeen different HLA alleles were collected (Figure 2(a)), and the detailed antigen length distribution is shown in Figure 2(b). Additionally, we collected the neoantigen data from 12 publications, including 2837 nonneoepitopes and 164 neoepitopes (Figure 2(c)), and the detailed neoantigen length distribution is shown in Figure 2(d).

The TCR contact position plays a crucial role in the analysis of immunogenicity, as TCRs might be more sensitive to some amino acids; the amino acid preference in the antigen epitope peptide and the antigen nonepitope peptide was further analyzed after excluding anchor sites (N-terminal, position 2, and C-terminal) (Figure 3). We found that TCRs tend to identify hydrophobic amino acids. For example, 3/4 hydrophobic amino acids (L, W, P, A, V, and M) occur more frequently in immunogenicity epitopes. Charged amino acids (*e.g.*, D and K) are enriched in nonepitopes, whereas the rest of the charged amino acids (R, H, and E) show no difference. Based on the result in Figure 3, the amino acid distribution difference at the TCR contact sites was regarded by us as one of the immunogenicity features (*i.e.*, Frequency Score for Immunogenic Peptide (C22)).

3.1. Classification Prediction Model for Antigen Epitopes. We constructed the features of peptides on the basis of the characteristics of amino acids (see Calculation of Peptide Characterstics Based on Amino Acid Sequences). All amino acid characteristics were selected from ProtScale [46] in ExPASy (SIB Bioinformatics Resource Portal). The 21 involved features are as follows: Kyte-Doolittle numeric hydrophobicity scale (C1) [47], molecular weight (C2), bulkiness (C3) [48], polarity (C4) [49], recognition factors (C5) [50], hydrophobicity (C6) [51], retention coefficient in HPLC (C7) [52], ratio hetero end/side (C8) [49], average flexibility (C9) [53], beta-sheet (C10) [54], alpha-helix (C11) [55], beta-turn (C12) [55], relative mutability (C13) [56], number of codon(s) (C14), refractivity (C15) [57], transmembrane tendency (C16) [58], accessible residues (%) (C17) [59], average area buried (C18) [60], conformational parameter for coil (C19) [55], total beta-strand (C20) [60], and parallel beta-strand (C21) [61] (see Table S4 for details). Also, Frequency Score for Immunogenic Peptide (C22), Calculating Peptide Entropy (C23), and Rank(%) Score (C24) were also taken into consideration. Together, 24 immunogenic features were collected, and all features were retained for antigen epitope prediction after screening using the R package Boruta. Compared with other characteristics, the frequency score for immunogenic peptide and rank(%) have higher impact, suggesting that they have more significant influence on antigen epitope classification (Figure 4(a)).

The receiver operator characteristic (ROC) curve of models are shown in Figure 4. The fivefold cross-validation AUC was 0.81 in the prediction model for the antigen epitope (line in red, Figure 4(b)), and the externally validated (see Table 2) AUC was 0.75 (line in purple, Figure 4(c)). Here, we tried to remove peptides for which HLA supertypes did

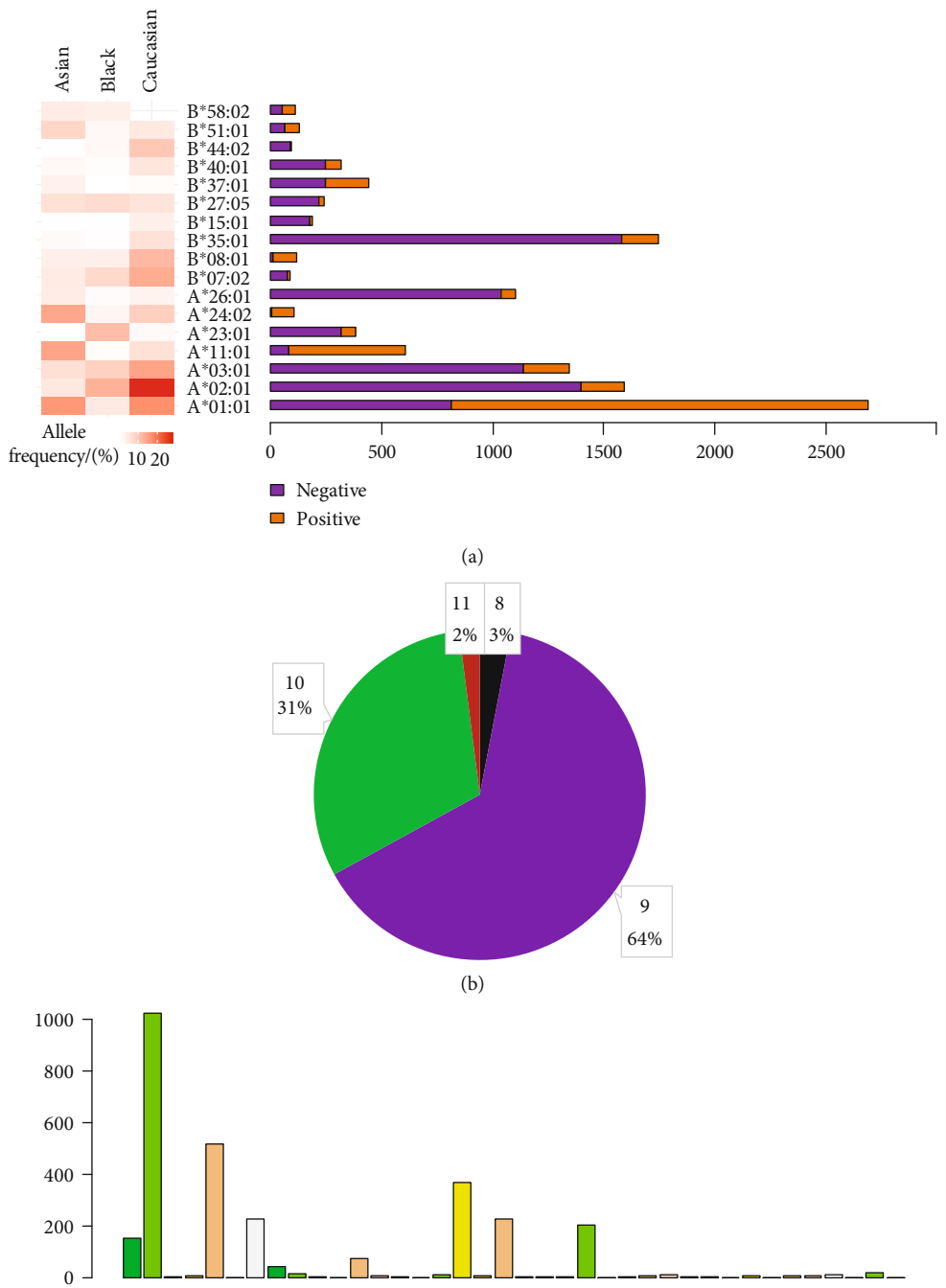
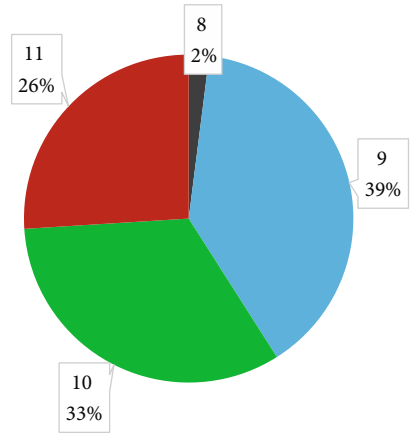


Figure 2: Continued.



(d)

Figure 2: Epitope/neoepitope peptide composition and amino acid length distribution. (a) Detailed data distribution of seventeen HLA alleles of antigen peptides, the proportion of each HLA allele (positive and negative) epitopes, and the corresponding HLA frequency in Asians, Blacks, Caucasians. (b) Proportion of antigen peptides with lengths of 8-11 AA. (c) Data distribution of HLA alleles of neoantigen peptides. (d) Proportion of neoantigen peptides with lengths of 8-11 AA.

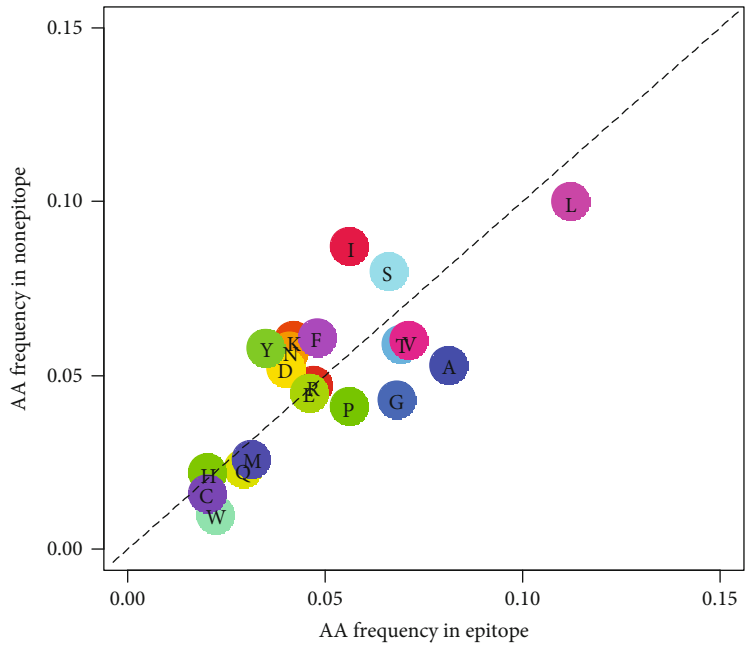


Figure 3: Antigen epitope amino acid distribution frequency in the TCR contact site of epitopes and nonepitopes. Frequency distribution of amino acids at TCR contact sites in antigen epitope and nonepitope peptides, and the amino acids below the dotted line are preferred by the epitope.

not appear in the training set from the externally validated antigen data, and the AUC, specificity, and sensitivity were increased to 0.78, 0.71, and 0.72, respectively (line in pink, Figure 4(c)). This, to some extent, verifies our conjecture about TCR specific recognition of different HLA alleles presenting peptides.

3.2. Classification Prediction Model for Neoantigen Epitopes. Neoantigens derived from somatic mutations are different from the wild peptide sequences. Therefore, some mutation-related characteristics were also taken into account. For instance, difference in hydrophobility before and after mutation (C25), differential agretopicity index (DAI, C26) [62], and whether the mutation position was anchored (C27). Finally, 27 features were selected for the neoantigen epitope prediction model. However, only 25 neoantigen-related features were retained after running Boruta, because C25 and C27 were removed. Also, rank(%) showed a marked effect (Figure 5(a)). In the fivefold cross-validation of the prediction model for neoantigen epitopes, AUC was 0.78 (Figure 5(b)).

3.3. Web Server for TCR Epitope Prediction. Based on the abovementioned validated features, we established a web server for TCR epitope prediction, named "INeo-Epp." This

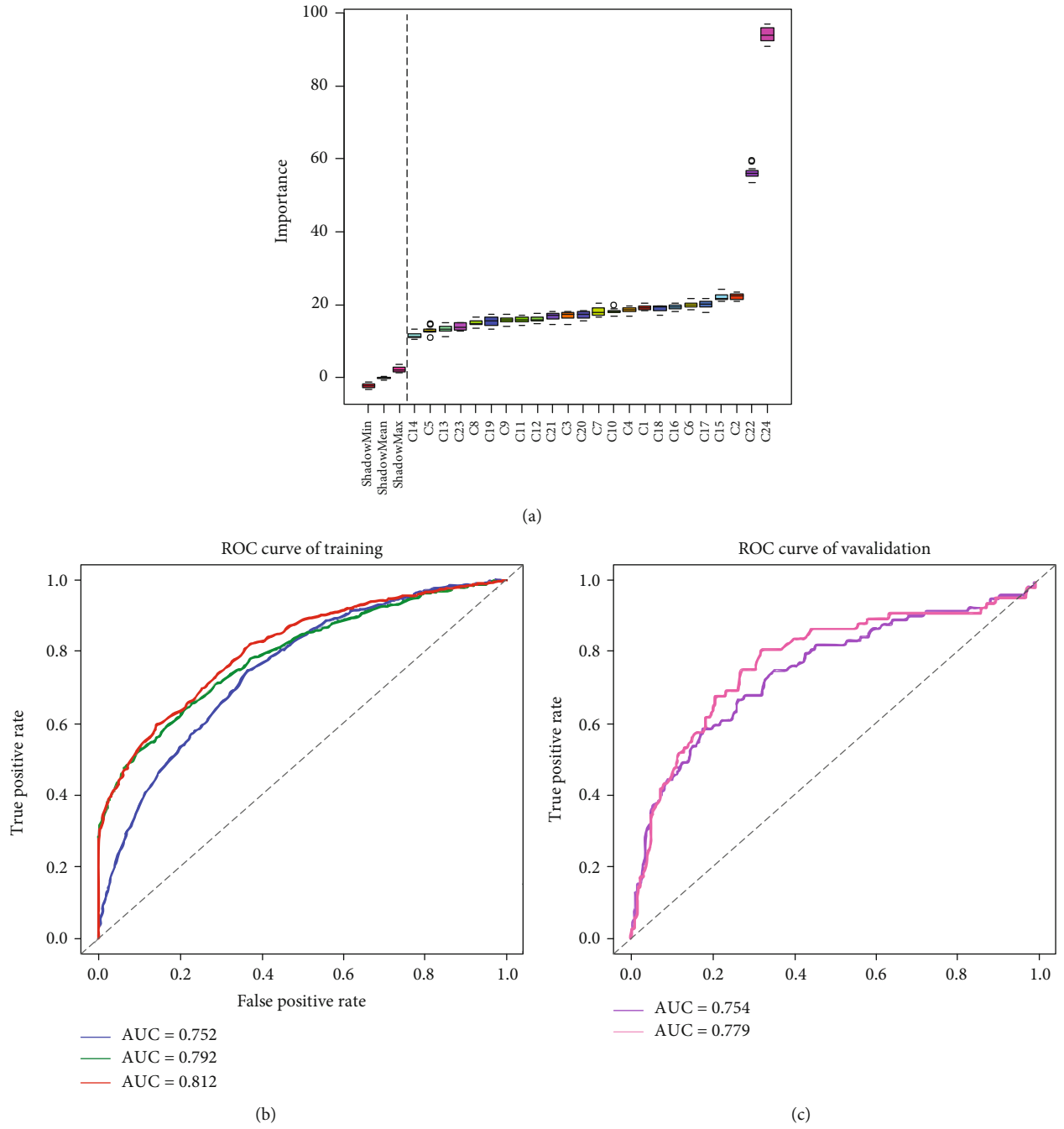


Figure 4: Feature selection in antigen epitopes and ROC curves of antigen epitope classification. (a) Peptide features: twenty-four features were screened, and we defined the features on the right of the dotted line as being effective. (b) Trained model: the line in blue represents antigen epitopes without screening; the line in green represents the selection with the deletion of the rank (%) > 2 nonepitope; the line in red represents the selection with the deletion of the nonepitopes 100% matching the human reference peptide sequence. (c) External validation: the ROC curves for the external verification set. The line in purple represents modeling using antigen epitopes without filtering, and the line in pink represents modeling using antigen epitopes removing nonepitopes with rank (%) > 2 and HLA for which supertypes did not appear in the training set.

tool can be used to predict both immunogenic antigen and neoantigen epitopes. For antigens, the nine main HLA supertypes can be used. We recommend the peptides with the lengths of 8-12 residues, but not less than 8. N-terminal, position 2, and C-terminal were treated as anchored sites by default. A predictive score value greater than 0.5 is considered as immunogenicity (positive-high), a score between 0.4 and 0.5 is considered as positive-low, and a score less than 0.4 is considered as negative-high. It is critical to make sure that the HLA-subtype must match your peptides (rank (%) < 2). Where HLA-subtypes mismatch, a large deviation of the rank(%) value may strongly influence the

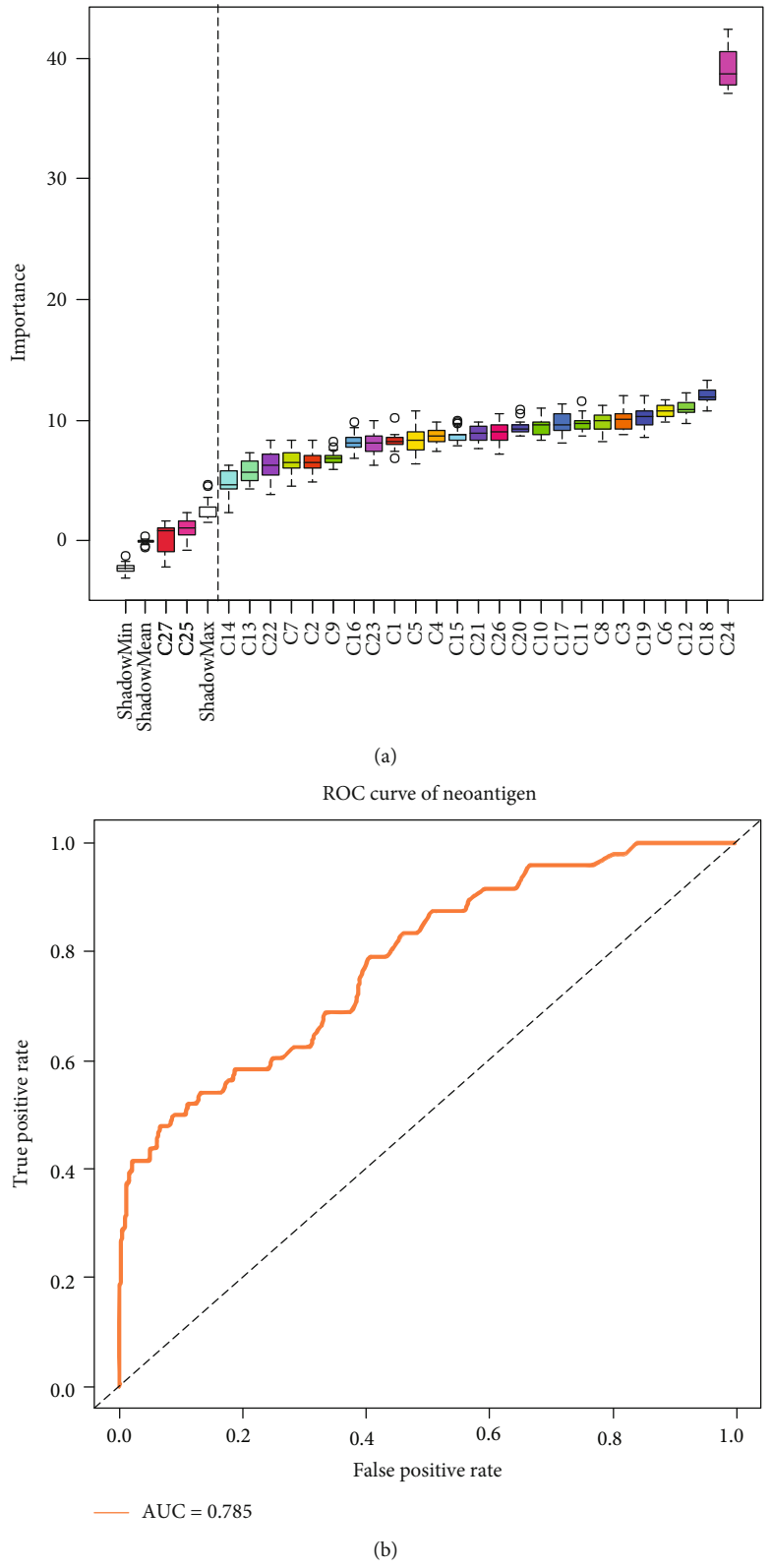


Figure 5: Feature selection in neoantigen epitopes and ROC curves of neoantigen epitope classification. (a) Twenty-seven features were screened, and the 25 features on the right of the dotted line were reserved for modeling using a random forest algorithm. (b) ROC curves of neoantigen epitope classification.

results. Additionally, the neoantigen model requires providing wild type and mutated sequences at the same time to extract mutation-associated characteristics, and currently only immunogenicity prediction for neoantigens of single amino acid mutations are supported. Users can choose example options to test the INeo-Epp (http://www.biostatistics.online/ineo-epp/neoantigen.php).

4. Discussion

Due to the complexity of antigen presenting and TCR binding, the mechanism of TCR recognition has not been clearly revealed. In 2013, Calis et al. [63] developed a tool for epitope identification for mice and humans (AUC = 0.68). Although mice and human beings are highly homologous, the murine epitopes may very likely cause limitations in identifying human epitopes. Inspired by J. A. Calis, our research here focused on human beings' epitopes and has been conducted in a larger data set.

By analyzing epitope immunogenicity from the perspective of amino acid molecular composition, we observed that TCRs do have a preference for hydrophobic amino acid recognition. For short peptides presented by different HLA supertypes, TCRs may have different identification patterns. The immunogenicity prediction based on all HLA-presenting peptides may affect the accuracy of the prediction results. That is, if the prediction could focus on specified HLA-presenting peptides, the results may improve. Therefore, in our work we used HLA supertypes to improve the prediction of HLA-presenting epitopes, including antigen epitopes and neoantigen epitopes, for a better recognition by TCRs. At present, neoantigen epitopes that can be collected in accordance with the standard for experimental verification are too few, the data of positive and negative neoantigens are unbalanced, and there is not enough data to be used for an external verification set. In the future, we will continue to refine and expand our training and verification datasets. Recently, Laumont et al. [64] demonstrated that noncoding regions aberrantly expressing tumor-specific antigens (aeTSAs) may represent ideal targets for cancer immunotherapy. These epitopes can also be studied in the future. Increased epitope data may also help empower the prediction of potentially immunogenic peptides or neopeptides.

5. Conclusions

Neoantigen prediction is the most important step at the start of preparation of a neoantigen vaccine. Bioinformatics methods can be used to extract tumor mutant peptides and predict neoantigens. Most current strategies aimed at and ended in presenting peptide predictions, and among the results of these predictions, probably only fewer than 10 neoantigens might be clinically immunogenic and produce effective immune response. It is time-consuming and costly to experimentally eliminate the false positively predicted peptides. Our methods as developed in this study and the INeo-Epp tool may help eliminate false positive antigen/-neoantigen peptides and greatly reduce the amount of candidates to be verified by experiments. We believe that in the age of biological system data explosion, computational approaches are a good way to enhance research efficiency and direct biological experiments. With the development of machine learning and deep learning, we expect that the prediction of epitope immunogenicity will be continually improved.

In summary, this study provides a novel T-cell HLA class-I immunogenicity prediction method from epitopes to neoantigens, and the INeo-Epp can be applied not only to identify putative antigens, but also to identify putative neoantigens.

It needs to be stated here that we published the preprint [65] of this article in July 2019. This is a modified version.

Data Availability

The data used to support the findings of this study are included within the supplementary information file(s).

Disclosure

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

We sincerely thank Drs. Menghuan Zhang, Hong Li, and Qibing Leng for our valuable discussion. We also acknowledge Dr. Michael Liebman for his critical reading and editing. This work was funded by the National Natural Science Foundation of China (No. 31870829), the Shanghai Municipal Health Commission, and the Collaborative Innovation Cluster Project (No. 2019CXJQ02).

Supplementary Materials

S1 Table IEDB antigen epitopes summary. Detailed description of 17 HLA molecules collected from IEDB. (XLSX) S2 Table External validation antigen epitopes summary. Epitope details of 7 publications. (XLSX) S3 Table Neoantigen epitopes summary. Epitope details of 13 publications. (XLSX) S4 Table Summary of amino acid characteristics. For all amino acid characteristics (n=21) that are described in the ExPASy. (XLSX). *(Supplementary Materials)*

References

[1] D. V. Desai and U. Kulkarni-Kale, "T-cell epitope prediction methods: an overview," *Methods in Molecular Biology*, vol. 1184, pp. 333–364, 2014.

[2] A. L. Goldberg and K. L. Rock, "Proteolysis, proteasomes and antigen presentation," *Nature*, vol. 357, no. 6377, pp. 375–379, 1992.

[3] C. Keşmir, A. K. Nussbaum, H. Schild, V. Detours, and S. Brunak, "Prediction of proteasome cleavage motifs by neural

networks," *Protein Engineering, Design and Selection*, vol. 15, no. 4, pp. 287–296, 2002.

[4] M. V. Larsen, C. Lundegaard, K. Lamberth et al., "An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions," *European Journal of Immunology*, vol. 35, no. 8, pp. 2295–2303, 2005.

[5] V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, and M. Nielsen, "NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data," *Journal of Immunology*, vol. 199, no. 9, pp. 3360–3368, 2017.

[6] T. J. O'Donnell, A. Rubinsteyn, M. Bonsack, A. B. Riemer, U. Laserson, and J. Hammerbacher, "MHCflurry: open-source class I MHC binding affinity prediction," *Cell Systems*, vol. 7, no. 1, pp. 129–132.e4, 2018.

[7] M. Wang, K. Lamberth, M. Harndahl et al., "CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening," *Vaccine*, vol. 25, no. 15, pp. 2823–2831, 2007.

[8] C. L. Pérez, M. V. Larsen, R. Gustafsson et al., "Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes," *Journal of Immunology*, vol. 180, no. 7, pp. 5092–5100, 2008.

[9] C. Lundegaard, I. Hoof, O. Lund, and M. Nielsen, "State of the art and challenges in sequence based T-cell epitope prediction," *Immunome Research*, vol. 6, Suppl 2, p. S3, 2010.

[10] J. L. Sanchez-Trincado, M. Gomez-Perosanz, and P. A. Reche, "Fundamentals and Methods for T- and B-Cell Epitope Prediction," *Journal of Immunology Research*, vol. 2017, 14 pages, 2017.

[11] V. N. Kristensen, "The Antigenicity of the Tumor Cell — Context Matters," *New England Journal of Medicine*, vol. 376, no. 5, pp. 491–493, 2017.

[12] K. Kiyotani, H. T. Chan, and Y. Nakamura, "Immunopharmacogenomics towards personalized cancer immunotherapy targeting neoantigens," *Cancer Science*, vol. 109, no. 3, pp. 542–549, 2018.

[13] J. Ponomarenko, N. Papangelopoulos, D. M. Zajonc, B. Peters, A. Sette, and P. E. Bourne, "IEDB-3D: structural data within the immune epitope database," *Nucleic Acids Research*, vol. 39, no. Database, pp. D1164–D1170, 2010.

[14] A. Sette and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," *Immunogenetics*, vol. 50, no. 3-4, pp. 201–212, 1999.

[15] J. Sidney, B. Peters, N. Frahm, C. Brander, and A. Sette, "HLA class I supertypes: a revised and updated classification," *BMC Immunology*, vol. 9, no. 1, p. 1, 2008.

[16] P. A. Ott, Z. Hu, D. B. Keskin et al., "An immunogenic personal neoantigen vaccine for patients with melanoma," *Nature*, vol. 547, no. 7662, pp. 217–221, 2017.

[17] U. Sahin, E. Derhovanessian, M. Miller et al., "Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer," *Nature*, vol. 547, no. 7662, pp. 222–226, 2017.

[18] Z. Hu, P. A. Ott, and C. J. Wu, "Towards personalized, tumour-specific, therapeutic vaccines for cancer," *Nature Reviews Immunology*, vol. 18, no. 3, pp. 168–182, 2018.

[19] E. M. Van Allen, D. Miao, B. Schilling et al., "Genomic correlates of response to CTLA-4 blockade in metastatic melanoma," *Science*, vol. 350, no. 6257, pp. 207–211, 2015.

[20] M. Efremova, F. Finotello, D. Rieder, and Z. Trajanoski, "Neoantigens generated by individual mutations and their role in cancer immunity and immunotherapy," *Frontiers in Immunology*, vol. 8, p. 1679, 2017.

[21] L. Klein, M. Hinterberger, G. Wirnsberger, and B. Kyewski, "Antigen presentation in the thymus for positive selection and central tolerance induction," *Nature Reviews Immunology*, vol. 9, no. 12, pp. 833–844, 2009.

[22] F. F. Gonzalez-Galarza, A. McCabe, E. J. Melo dos Santos et al., "Allele frequency net database," *Methods in Molecular Biology*, vol. 1802, pp. 49–62, 2018.

[23] D. Weiskopf, M. A. Angelo, E. L. de Azeredo et al., "Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 22, pp. E2046–E2053, 2013.

[24] H. Luxenburger, F. Graß, J. Baermann et al., "Differential virus-specific CD8+T-cell epitope repertoire in hepatitis C virus genotype 1 versus 4," *Journal of Viral Hepatitis*, vol. 25, no. 7, pp. 779–790, 2018.

[25] Y. Xia, W. Pan, X. Ke et al., "Differential escape of HCV from CD8+T cell selection pressure between China and Germany depends on the presenting HLA class I molecule," *Journal of Viral Hepatitis*, vol. 26, no. 1, pp. 73–82, 2019.

[26] H. Vahed, A. Agrawal, R. Srivastava et al., "Unique Type I Interferon, Expansion/Survival Cytokines, and JAK/STAT Gene Signatures of Multifunctional Herpes Simplex Virus-Specific Effector Memory CD8+TEMCells Are Associated with Asymptomatic Herpes in Humans," *Journal of Virology*, vol. 93, no. 4, p. e01882, 2019.

[27] A. Khakpoor, Y. Ni, A. Chen et al., "Spatiotemporal Differences in Presentation of CD8 T Cell Epitopes during Hepatitis B Virus Infection," *Journal of Virology*, vol. 93, no. 4, 2019.

[28] A. Huth, X. Liang, S. Krebs, H. Blum, and A. Moosmann, "Antigen-specific TCR signatures of cytomegalovirus infection," *The Journal of Immunology*, vol. 202, no. 3, pp. 979–990, 2019.

[29] S. O. Sekyere, B. Schlevogt, F. Mettke et al., "HCC immune surveillance and antiviral therapy of hepatitis C virus infection," *Liver Cancer*, vol. 8, no. 1, pp. 41–65, 2019.

[30] D. A. Wick, J. R. Webb, J. S. Nielsen et al., "Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer," *Clinical Cancer Research*, vol. 20, no. 5, pp. 1125–1134, 2014.

[31] T. Karasaki, K. Nagayama, M. Kawashima et al., "Identification of Individual Cancer-Specific Somatic Mutations for Neoantigen- Based Immunotherapy of Lung Cancer," *Journal of Thoracic Oncology*, vol. 11, no. 3, pp. 324–333, 2016.

[32] A. Gros, M. R. Parkhurst, E. Tran et al., "Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients," *Nature Medicine*, vol. 22, no. 4, pp. 433–438, 2016.

[33] E. Stronen, M. Toebes, S. Kelderman et al., "Targeting of cancer neoantigens with donor-derived T cell receptor repertoires," *Science*, vol. 352, no. 6291, pp. 1337–1341, 2016.

[34] A. Nelde, J. S. Walz, D. J. Kowalewski et al., "HLA class I-restrictedMYD88L265P-derived peptides as specific targets

for lymphoma immunotherapy," *OncoImmunology*, vol. 6, no. 3, p. e1219825, 2017.

[35] X. Zhang, S. Kim, J. Hundal et al., "Breast cancer neoantigens can induce CD8+T-cell responses and antitumor immunity," *Cancer Immunology Research*, vol. 5, no. 7, pp. 516–523, 2017.

[36] M. Müller, D. Gfeller, G. Coukos, and M. Bassani-Sternberg, "'Hotspots' of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization," *Frontiers in Immunology*, vol. 8, p. 1367, 2017.

[37] V. P. Balachandran, A. P. C. G. Initiative, M. Łuksza et al., "Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer," *Nature*, vol. 551, no. 7681, pp. 512–516, 2017.

[38] T. Matsuda, M. Leisegang, J.-H. Park et al., "Induction of neoantigen-specific cytotoxic T cells and construction of T-cell receptor-engineered T cells for ovarian cancer," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 24, no. 21, pp. 5357–5367, 2018.

[39] K. Sonntag, H. Hashimoto, M. Eyrich et al., "Immune monitoring and TCR sequencing of CD4 T cells in a long term responsive patient with metastasized pancreatic ductal carcinoma treated with individualized, neoepitope-derived multipeptide vaccines: a case report," *Journal of translational medicine*, vol. 16, no. 1, 2018.

[40] A.-M. Bjerregaard, M. Nielsen, V. Jurtz et al., "An analysis of natural T cell responses to predicted tumor neoepitopes," *Frontiers in Immunology*, vol. 8, 2017.

[41] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[42] M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.

[43] M. B. Kursa and W. R. Rudnicki, "Feature selection with the-BorutaPackage," *Journal of Statistical Software*, vol. 36, no. 11, 2010.

[44] I. Law and M. Wiener, *Classification and regression by randomForest*, vol. 2, no. 3, 2002R News, 2002.

[45] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940-3941, 2005.

[46] J. M. Walker, "The proteomics protocols handbook," *Biochemistry*, vol. 71, no. 6, pp. 696–696, 2006.

[47] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[48] J. M. Zimmerman, N. Eliezer, and R. Simha, "The characterization of amino acid sequences in proteins by statistical methods," *Journal of Theoretical Biology*, vol. 21, no. 2, pp. 170–201, 1968.

[49] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[50] S. Fraga, "Theoretical prediction of protein antigenic determinants from amino acid sequences," *Canadian Journal of Chemistry*, vol. 60, no. 20, pp. 2606–2610, 1982.

[51] R. M. Sweet and D. Eisenberg, "Correlation of sequence hydrophobicities measures similarity in three- dimensional protein structure," *Journal of Molecular Biology*, vol. 171, no. 4, pp. 479–488, 1983.

[52] J. L. Meek, "Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, no. 3, pp. 1632–1636, 1980.

[53] G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus, "Hydrophobicity of amino acid residues in globular proteins," *Science*, vol. 229, no. 4716, pp. 834–838, 1985.

[54] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence," *Advances in enzymology and related areas of molecular biology*, vol. 47, pp. 45–148, 1978.

[55] G. Deléage and B. Roux, "An algorithm for protein secondary structure prediction based on class prediction," *Protein Engineering*, vol. 1, no. 4, pp. 289–294, 1987.

[56] A. Burger, "Atlas of protein sequence and structure 1969," *Journal of Medicinal Chemistry*, vol. 13, no. 2, pp. 337–337, 1970.

[57] D. D. Jones, "Amino acid properties and side-chain orientation in proteins: a cross correlation approach," *Journal of Theoretical Biology*, vol. 50, no. 1, pp. 167–183, 1975.

[58] G. Zhao and E. London, "Strong correlation between statistical transmembrane tendency and experimental hydrophobicity scales for identification of transmembrane helices," *Journal of Membrane Biology*, vol. 229, no. 3, pp. 165–168, 2009.

[59] J. Janin, "Surface and inside volumes in globular proteins," *Nature*, vol. 277, no. 5696, pp. 491-492, 1979.

[60] J. R. Green, M. J. Korenberg, R. David, and I. W. Hunter, "Recognition of adenosine triphosphate binding sites using parallel cascade system identification," *Annals of Biomedical Engineering*, vol. 31, no. 4, pp. 462–470, 2003.

[61] S. Lifson and C. Sander, "Antiparallel and parallel β-strands differ in amino acid residue preferences," *Nature*, vol. 282, no. 5734, pp. 109–111, 1979.

[62] F. Duan, J. Duitama, S. al Seesi et al., "Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity," *Journal of Experimental Medicine*, vol. 211, no. 11, pp. 2231–2248, 2014.

[63] J. J. A. Calis, M. Maybeno, J. A. Greenbaum et al., "Properties of MHC class I presented peptides that enhance immunogenicity," *PLoS Computational Biology*, vol. 9, no. 10, article e1003266, 2013.

[64] C. M. Laumont, K. Vincent, L. Hesnard et al., "Noncoding regions are the main source of targetable tumor-specific antigens," *Science Translational Medicine*, vol. 10, no. 470, p. eaau5516, 2018.

[65] G. Wang, H. Wan, X. Jian et al., *INeo-Epp: T-cell HLA class I immunogenic or neoantigenic epitope prediction via random forest algorithm based on sequence related amino acid features*, bioRxiv, 2019.