INeo-Epp: T-cell HLA class I immunogenic or neoantigenic epitope prediction via random forest algorithm based on sequence related amino acid features

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

[1] College of Food Science and Technology, Shanghai Ocean University, Shanghai, China

[2] Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai, China

[3] School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai, China

[4] Key Laboratory of Carcinogenesis and Cancer Invasion, Ministry of Education; Key Laboratory of Carcinogenesis, National Health and Family Planning Commission, Xiangya Hospital, Central South University, Changsha, China.

* Corresponding author

E-mail: yzhao@ shou.edu.cn (YZ)

E-mail: yong_lynn@163.com (YL)

E-mail: luxiex2017@outlook.com (LX)

¶These authors contributed equally to this work.

Abstract

In silico T-cell epitope prediction plays a key role in immunization experiments design and vaccine preparation. In this study, classification models based on random forests algorithm were trained by use of experimental human leukocyte antigen class I (HLA-I) presenting T-cell peptides data, in which several characteristics were constructed as immunogenicity features, including amino acid sequence characteristics, peptide entropy, eluted ligand likelihood percentile rank (EL %Rank) score and score of immunogenic peptide. The classification result for the antigen epitopes outperformed the previous research (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 epitopes named 'neoantigens'. Based on mutation information and sequence related amino acid characteristics, a prediction model of neoantigen was established as well (AUC=0.78). Further, a web-based tool was developed for the prediction of either human antigen epitope or neoantigen epitope (http://www.biostatistics.online/INeo-Epp/antigen.php). Overall, by analyzing amino acid distribution in T-cell receptor (TCR) contact sites, we found that TCR prefers to recognize the hydrophobic amino acids. This work may provide a new insight for T-cell recognition of antigen peptides.

Author summary

Currently, most epitope prediction researches focus on peptides processing and presenting, such as proteasomal cleavage, transporter associated with antigen processing (TAP) and major histocompatibility complex (MHC) combination. To date, however, the immunogenicity mechanism of epitopes remains unclear. It is generally agreed upon that T-cell immunogenicity may be influenced by foreignness, accessibility, molecular weight, molecular structure, molecular conformation, chemical properties and physical properties of target peptides in different degrees. Here, we first collected quite an amount of experimental HLA-I T-cell peptides data, as well as the potential immunogenic amino acid features. Subsequently, based on the random forest algorithm, we successfully constructed the separate prediction models for T cell immunogenic HLA-I presenting antigen and neoantigen epitopes. Furthermore, we built a web-based tool to facilitate the prediction of HLA-I T-cell immunogenic epitopes.

Introduction

An antigen is consisted of several epitopes, which can be recognized either by B-

or T-cells and/or molecules of the host immune system. However, usually, a few amino acid residues that comprise an epitope are sufficient to elicit an immune response [1]. MHC-I (HLA-I in human) antigen peptides are processed and presented as follows: (1) cytosolic and nuclear proteins are cleaved to short peptides by intracellular proteinases; (2) some are selectively transferred to endoplasmic reticulum (ER) by TAP transporter, and subsequently are treated by endoplasmic reticulum aminopeptidase; (3) antigen presenting cells (APCs) present peptides possessed to 8-11 AA (amino acid) residues on MHC class I molecules to CD8+ T cells [2]. So far, several software have been developed to predict the antigen processing and presentation, including NetChop [3], NetCTL [4], NetMHCpan [5], MHCflurry [6]. However, statistically, approximately only 1% of the predicted binding peptide-MHC complexes (p-MHC) can eventually cause immunogenicity [7]. Although the recognition and amplification of T-cells may benefit from the development of T-cell receptor (TCR) sequencing, the cycle of vaccine development and immunization research is extended. Thus, an effective identification method follow-up the above software is urgently needed to shorten the whole cycle.

Nowadays, many experimental human epitopes may be acquired from the immune epitope database (IEDB) [8], which makes it feasible to mathematically predict human epitopes. Even if IEDB provides us a wide range of information on T cell epitopes, a high degree of MHC polymorphism brings forward a severe challenge for T-cell epitope prediction. HLA molecules have hundreds of different variants [9].

Experimentally, many infrequent HLA subtypes peptides (*e.g.* B55, B63) with uneven positive and negative distributions are not conducive to analyze the potential deviation existed in TCR recognition owing to various HLA presented peptides. A general analysis of all HLA presented peptides, ignoring the pattern of TCR recognition of specific HLA, may result in a lower prediction.

Due to the intensive study on HLA, HLA supertype has been proposed. Sette *et al.* [10] classified, for the first time, overlapping peptide binding repertoires into nine major functional HLA supertypes (A1, A2, A3, A24, B7, B27, B44, B58, B62). In 2008, John Sidney *et al* [11] made a further supplement, 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 for analyzing.

Screening of mutant and abnormally expressed epitopes are crucial in tumor immunotherapy. In 2017, Ott PA *et al.* [12] and Sahin *et al* [13]. confirmed that peptides and RNA vaccines made up of neoantigens in melanoma can stimulate and proliferate CD8+ and CD4+ T cells. Neoantigen vaccination not only can expand the existing specific T cells, but also induce a wide range of novel T-cell specificity in cancer patients and enhance tumor suppression [14]. 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 [15-16]. However, a considerable amount of identified candidate neoantigens in the process of sequencing recognition of somatic cell mutations were false positive, which would fail to stimulate TCR recognition and immune response. This is undoubtedly a disadvantage for designing vaccines against neoantigens.

In this study, based on the collection of the validated HLA-I T-cell peptides, including antigens and neoantigens, we discovered several effective classification features and successfully constructed the classification models for antigens and neoantigens, respectively. Furthermore, a web-based tool, INeo-Epp (immunogenic and neoangtigenic epitope prediction), was built for separate prediction of human antigen and neoantigen epitopes.

Results

Immunogenic and non-immunogenic epitopes

Peptides that can promote cytokines proliferation are considered as immunogenic epitopes. However, non-immunogenic epitopes may result from the following reasons: a) p-MHC truly unrecognized by TCR; b) peptides unpresented by MHC (quantitatively expressed as %rank>2); c) negative selection/clonal presentation induced by excessive similarity with autologous peptides [17]. In this work, to further study the recognition preferences of T cells, >2 %rank and 100% matching human GRCh38 peptide sequences were removed from the definition of non-immunogenic peptides.

Data statistics

In this study, 11,297 validated epitopes and non-epitopes with the length of 8-11 amino acids were collected from IEDB. T-cell responses include activation, cytotoxicity, proliferation, IFN-γ release, TNF release, granzyme B release, IL-2 release, IL-10 release. Seventeen different HLA alleles were collected (Fig 1A), and the detailed antigen lengths distribution are shown in (Fig 1B). Besides, we also collected the neoantigen data from 12 publications, including 2837 non-epitopes and 164 epitopes (Fig 1C), and the detailed neoantigen lengths distribution are shown in (Fig 1D).

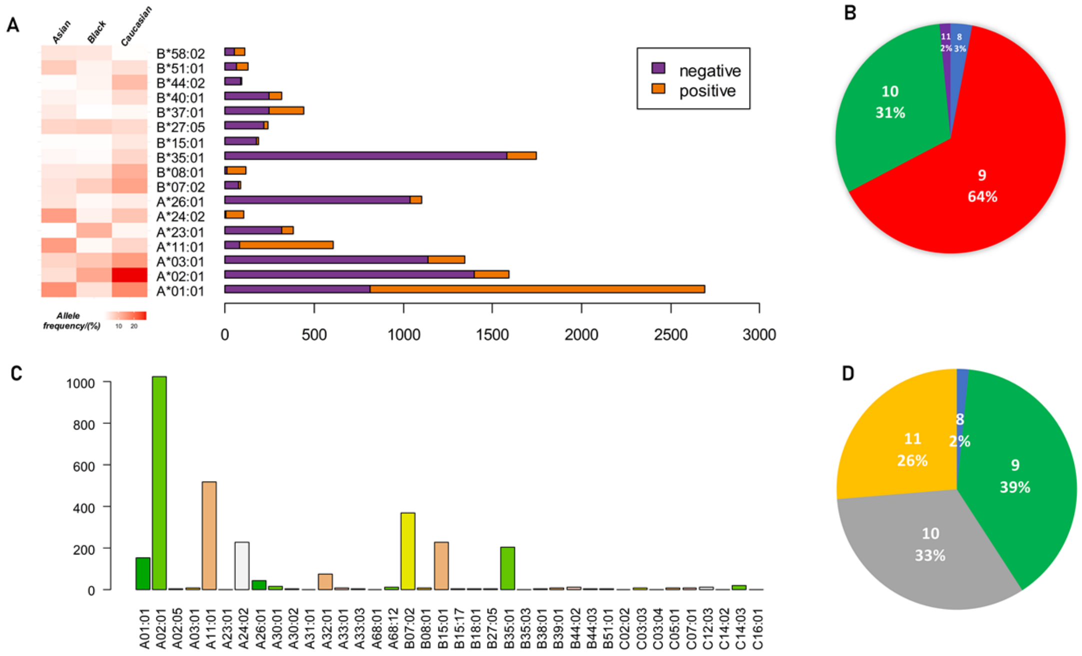


Figure1： Epitope peptides composition and amino acid lengths distribution.

(A) Detailed seventeen HLA alleles of antigen peptides data distribution and each HLA allele positive and negative data proportion and the corresponding HLA frequency in Asian, Black, Caucasian. (B) Antigen peptides proportion of 8-11 AA lengths. (C) Distribution of HLA alleles of neoantigen peptides. (D) Neoantigen peptides proportion of 8-11 AA lengths.

Furthermore, we analyzed the position-related amino acid arrangement in antigen epitopes. The result showed that leucine was strongly preferred in all the positions of antigen epitope, however, tryptophan, histidine, cysteine were the least preferred (Fig2 A). TCR contact position plays a crucial role in the analysis of immunogenicity. As TCRs might be more sensitive to some amino acids, the amino acids preference in antigen epitope peptide and antigen non-epitope peptide was further analyzed after excluding anchor sites. We found that a TCR tends to identify the hydrophobic amino acids (Fig 2B). For example, 70% of amino acids that occur more frequently in immunogenicity epitopes are hydrophobic (W, P, A, V, L). Charged amino acids (*e.g.* D, K) are enriched in non-epitopes, and amino acids with more complex R group structure frequently occur in non-epitopes. Based on the above, the amino acid distribution difference at the TCR contact sites was regarded by us as one of the immunogenicity features (*i.e.* score for immunogenic peptide (C22)).

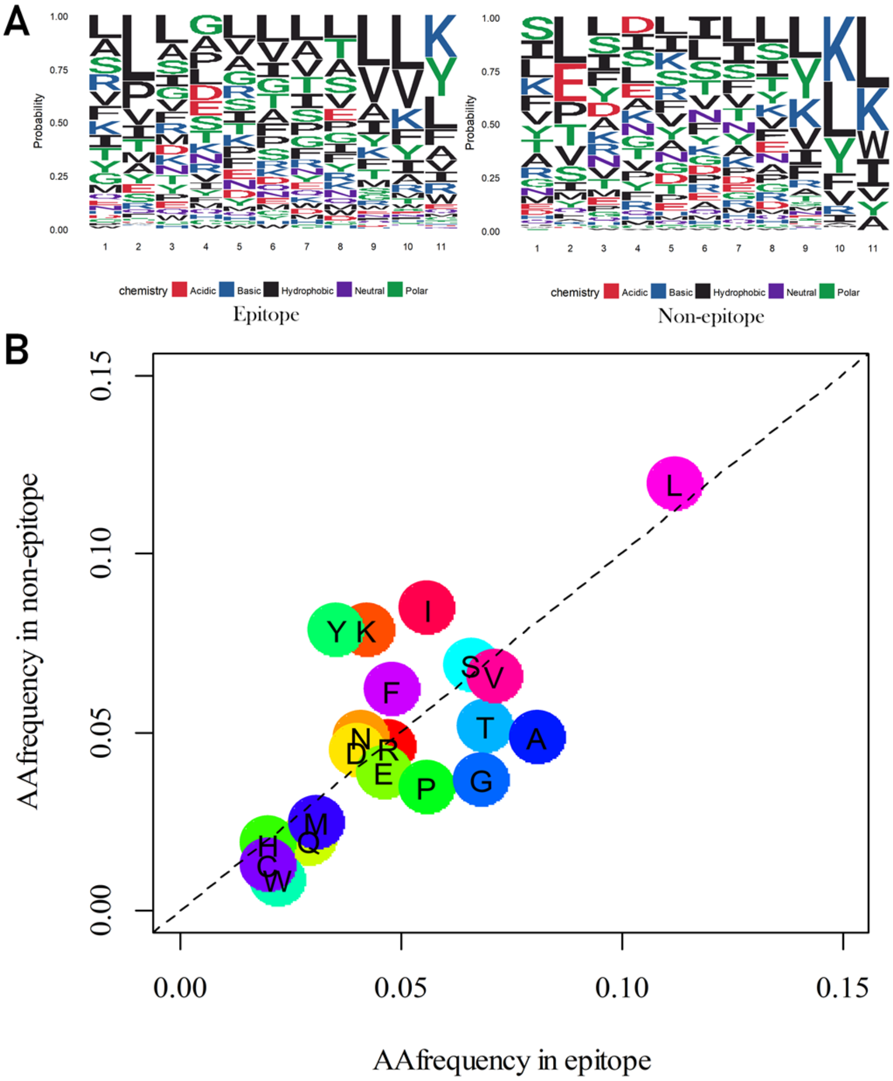


Figure 2: Antigen epitope amino acid distribution difference in P1-P11, and amino acid distribution frequency in TCR contact site of antigen epitope and non-epitope.

(A) The proportion of amino acids at each position of epitope and non-epitope peptides in antigen peptides, and the higher position the more frequency. (B) Frequency distribution of amino acids at solvent-exposed positions in antigen epitope and non-epitope peptides, and the amino acids below the dotted line are preferred by the epitope.

Classification prediction model for antigen epitopes

We constructed the features of peptides on the basis of the characteristics of amino acids (see Materials and Methods section: Characteristics Calculation of peptides based on amino acids). All amino acid characteristics were selected from Protscale [18] in ExPASy (SIB bioinformatics resource portal). The 21 involved features are as follows: Kyte–Doolittle numeric hydrophobicity scale (C1) [19], molecular weight (C2), bulkiness (C3) [20], polarity (C4) [21], recognition factors (C5) [22], hydrophobicity (C6) [23]， retention coefficient in HPLC (C7) [24], ratio hetero end/side (C8) [21], average flexibility (C9) [25], beta-sheet (C10) [26], alpha-helix (C11) [27]，beta-turn (C12) [27]， relative mutability (C13) [28], number of codon(s) (C14), refractivity (C15) [29], transmembrane tendency (C16) [30]， %accessible residues (C17) [31]， average area buried (C18) [32]，conformational parameter for coil (C19) [27]， total beta-strand (C20) [33]，parallel beta-strand (C21) [33] (see Table S4 in detail). Also, score for immunogenic peptide (C22), peptide entropy (C23) [34] and %rank (C24) were also taken into consideration. Together, 24 immunogenic features were collected, and all features were retained for antigen epitopes prediction after screening using R package Buroat [35]. Compared to other characteristics, score for immunogenic peptide and %rank have higher impacts, suggesting they have more significant power on antigen epitopes classification (Firure3 A).

The receiver operator characteristic (ROC) curve of models are shown in Fig 4. The five-fold cross validation AUC was 0.81 in the prediction model for antigen epitope (line in red Fig3 B) and the externally validated AUC was 0.75 (line in purple Fig4 C). Here, we tried to remove HLA supertypes (not included in training set) data 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 Fig4 C). This, to some extent, verifies our conjecture about TCR specific recognition of different HLA alleles presenting peptides.

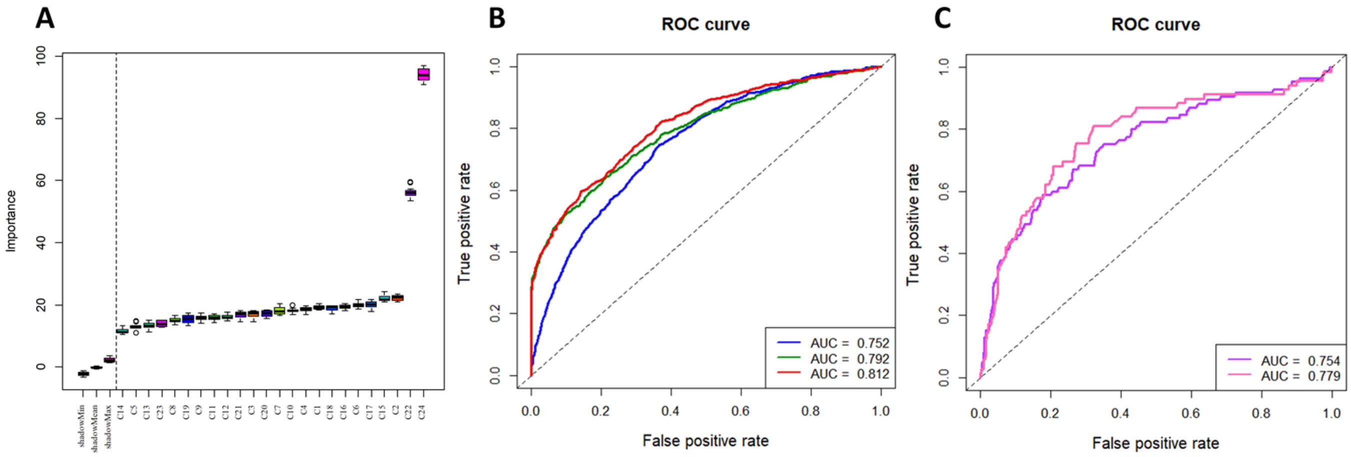


Figure 3: Feature selection in antigen epitopes and ROC curves of antigen epitopes classification. (A) Twenty four features were screened and retained, the features on the right of the dotted line are effective. (B) The line in blue represents antigen epitopes without screening; the line in green represents selection with the deletion of %rank>2 non-epitope; and the line in red represents selection with the deletion of the non-epitopes 100% matching human GRCh38 peptides sequence. (C) The ROC curves of external verification set, line in purple represents modeling using antigen epitopes without filtering, the line in pink represents using antigen epitopes removing non-

epitopes %rank>2 and HLA supertypes (not encountered in training set).

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, hydrophobic difference before and after mutation (C25), differential agretopicity index (DAI, C26) [36] and whether the mutation position was anchored (C27). Finally, 27 features were selected for the neoantigen model. However, only 25 neoantigen related features were retained after running Buroat, because C25 and C27 were removed. Also, %rank showed a marked effect (Fig 4A). in the five-fold cross-validation of the prediction model for neoantigen epitopes, AUC was 0.78 (Fig 4B).

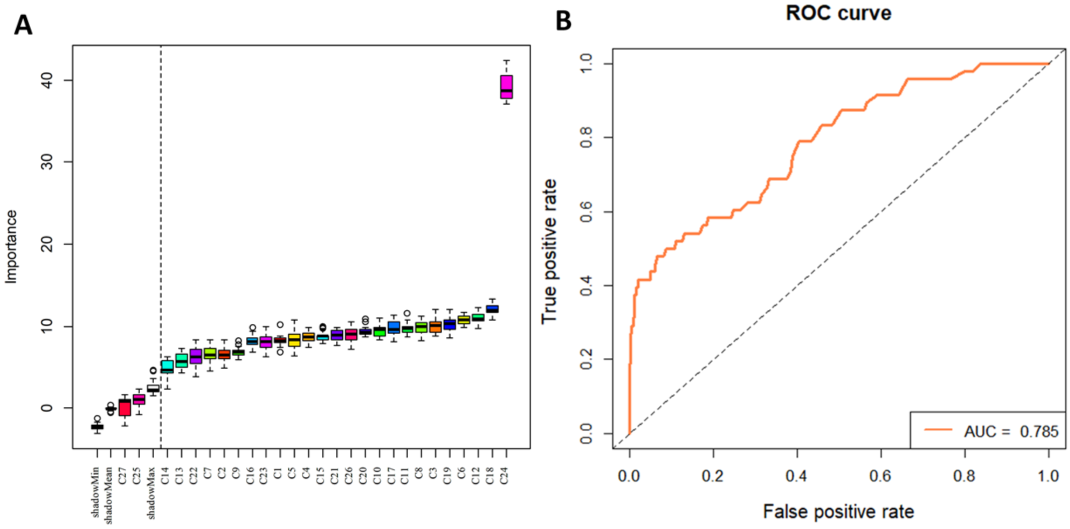


Figure 4: Feature selection in neoantigen epitopes and ROC curves of neoantigen epitopes classification. (A) Twenty seven features were screened and the 25 features on the right of the dotted line were reserved for modeling in random forest algorithm. (B) ROC curves of neoantigen epitopes classification.

Web server for TCR epitope prediction

Based on these above-mentioned validated features, we established a web server for TCR epitope prediction, named INeo-Epp. This tool can be used to predict both immunogenic antigen and neoantigen epitopes. For antigen, 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, C-terminal were treated as anchored sites by default. A predictive value greater than 0.5 is considered as positive immunogenicity (P). Please make sure that HLA-subtype must match your peptides. When HLA-subtype mismatches, the different %rank value may strongly influence the results. Additionally, the neoantigen model requires providing wild 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. You can use example option to test the INeo-Epp (http://www.biostatistics.online/INeo-Epp/antigen.php).

Discussion

Because of the complexity of antigen presenting and TCR binding, the mechanism of TCR recognition has not been clearly revealed. In 2013, J. A. Calis [37] developed a tool for epitope identification of mice and humans (AUC = 0.68). Although mice and human beings are highly homologous, the murine epitopes may very likely cause deviation in identifying human epitopes. Inspired by J. A. Calis, our research focused on human beings' epitopes and were conducted in a larger data set. In our study, the TCR recognized immunogenic epitope prediction AUC is increased to 0.81.

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, the prediction results of specified HLA-presenting peptides may be better. Recently, Céline M. Laumont [38] demonstrated that noncoding regions aberrantly expressed tumor-specific antigens (aeTSAs) may represent ideal targets for cancer immunotherapy. These epitopes can also be studied in the future.

However, for neoantigens prediction, the positive prediction rate is not as good (AUC is 0.78 and no external validation), because relevant and available experimental data of TCR recognized neoepitopes are limited. The immunogenic neoantigen prediction model remains to be improved as more data will be gathered. Besides, a TCR sequencing database would be needed to study the relationship between TCRs and epitopes from a deeper structure. More relevant amino acid properties and structural features may remain to be discovered for further mathematical analysis. We believe that in the age of biological systems data explosion, mathematical calculation is a good way to derive biological significance. With the development of machine learning and deep learning, we expect the prediction of neoantigen immunogenicity will be continually improved.

Neoantigen prediction is the most important step in the preparation of neoantigen vaccine. Bioinformatics methods can be used to extract tumor mutant peptides and predict neoantigens. Most current strategies end in presenting peptides predictions and among the results of these predictions, in the end, less than 10 neoantigens might be discovered, but it is time-consuming and costly to experimentally eliminate the false positively predicted peptides. Our methods in this study and the INeo-Epp tool may help eliminate a large number of false positive antigen/neoantigen peptides, and greatly reduce the amount of candidates to be verified by experiments.

In summary, this study provides an inference from the immunogenicity classification prediction of antigens to neoantigens, and the INeo-Epp can be applied not only to identify putative antigens, but also to identify putative neoantigens.

Materials and Methods

Generation of data sets

Antigen epitope data were collected from IEDB (Linear epitope, Humans, T cell assays, MHC class I, any disease were chosen). Data collection criteria: each HLA subtype quantity >50 and HLA frequency >0.5% (refer to allele frequency database [39]) (Table 1, check Table S1 for detailed information).

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		

The validation dataset was collected from seven published independent human antigen studies [40-46], consisting of 577 non-immunogenic epitopes and 85 immunogenic epitopes (Table 2, S2 Table)

TABLE2 | validated peptides data included in this study

Publication time	PMID	Author	non-epitopes	epitopes
2013	23580623	Weiskopf et al	477	42
2018	29397015	Hendrik Luxenburger et al	100	26
2018	30260541	Youchen Xia et al	-	1
2018	30487281	Hawa Vahed et al	-	4
2018	30518652	Atefeh Khakpoor et al	-	2
2018	30587531	Alina Huth et al	-	4
2018	30815394	Solomon Owusu Sekyere et al	-	6
Total			577	85
Remove negative %rank >2 and HLA supertypes (not in training set)			321	69

The neoantigen data were collected from 11 publications [15,48-57] and IEDB mutational epitopes, and 13 published data sets collected by Anne-Mette B in one publication [47] in 2017, see Table 3, S3 Table for details.

TABLE 3| Neoantigen data included in this study

Publication time	PMID	Author	Tumor Type	Non-immunogenic neo-epitopes	Immunogenic neo-epitopes	T-cell assay
2013-12	24323902	Darin A. W et al.	Ovarian Cancer	—	1	ELISPOT
2015-9	26359337	Eliezer M et al.	Melanoma	—	18	Clinical benefit
2015-11	26752676	Takahiro K et al.	Lung adenocarcinoma	—	4	—
2016-1	26901407	Alena Gros et al.	Melanoma	12	14	ELISPOT
2016-5	27198675	Erlend Strønen et al.	Melanoma	1134	16	CTL clone
2016-12	28405493	Annika Nelde et al.	Lymphoma	—	2	ELISPOT
2017-6	28619968	Xiuli Zhang et al.	Breast cancer	—	4	Flow cytometry
2017-10	29104575	Markus M et al.	Melanoma	10	16	—
2017-11	29187854	Anne-Mette B et al.	Polytype	1874	42	ELISPOT et al.
2017-11	29132146	Vinod P. B et al.	pancreatic	—	10	Flow Cytometry
2018-5	29720506	Tatsuo Matsuda et al.	Ovarian Cancer	—	3	ELISPOT
2018-12	29409514	Sonntag et al.	pancreatic ductal carcinoma	—	3	Flow Cytometry
2018-10	30357391	Randi Vita et al.	—	6	35	—
Total				3030	168	
Remove duplication				2837	164	
Remove negative %rank>2 and human 100% similar				1697	164	

Feature calculation

Characteristics calculation of peptides based on amino acid sequences. The formula for calculating peptide characteristics is shown in (1). P_N, P_2, P_C are considered to be embedded in HLA molecules and no contact with TCRs, so they're not evaluated.

$$P_c = \{ \sum_{x \in Pos(P)}^{x \notin (N,2,C)} P_{A_c} \} \Big/ (len(P) - 3) \quad (1)$$

P, peptide. ***c***, characteristic. Where P_c represents characteristics of peptides. ***A***, amino acid. ***N***, N-terminal in a peptide. ***C***, C-terminal in a peptide. Pos, amino acid position in peptide. Where P_{Ac} represents characteristics of amino acids in peptides.

Score for immunogenic peptide (C22). Amino acid distribution frequency differences between immunogenicity and non-immunogenic peptides at TCR contact sites were considered as a feature (2).

$$P_{score} = \sum_{x \in Pos(P)}^{x \notin (N,2,C)} \{ P_{ie^+}(f_A^{'}) - P_{ie^-}(f_A^{'}) \} \quad (2)$$

$P_{ie}{}^{+}$, immunogenic peptides. $P_{ie}{}^{-}$, non-immunogenic peptides. f'_A, amino acid frequency in TCR contact position. Where P_{ie+} (f'_A) represents frequency of amino acids in immunogenic peptides at TCR contact sites.

Calculating peptide entropy (C23). peptide entropy [58] was used as a feature (3).

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

$P_{H,}$ peptide entropy. $f_{A,}$ amino acid frequency in human GRCh38 peptides. Where P_{fA} represents the frequency in human GRCh38 peptides of amino acids in epitope peptides.

%rank score (C24). HLA binding prediction were run by netMHCpan4.0 in which %rank was recommended as evaluation standard， %rank<0.5 as strong binders, 0.5<%rank<2 as weak binders, %rank>2 as no binders.

Cross-validation, feature selection, random forests and ROC generation.

The cross-validation were generated in R using the package caret [59] (method = "repeatedcv", number = 5, repeats = 3). The feature screening result were generated in R using the package Buroat (a feature selection method). R package randomForest [60] was used for training data (mtry=14 for antigen epitope, mtry=15 for neoantigen, the remaining parameter use default values). R package ROCR was used [61] for drawing ROC.

Analysis and statistics

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

Acknowledgments

This work was funded by the National Natural Science Foundation of China (No. 31870829), Shanghai Municipal Health Commission, and Collaborative Innovation Cluster Project (No. 2019CXJQ02). We sincerely thank Menghuan Zhang, Guangrong Qin, Hong Li and Qibing Leng for their valuable advice.

Author Contributions

Conceptualization: Lu Xie, Guangzhi Wang

Funding acquisition: Lu Xie

Formal Analysis: Guangzhi Wang

Investigation: Guangzhi Wang

Methodology: Guangzhi Wang

Software: Huihui Wan, Ouyang Jan, Guangzhi Wang

Supervision: Yuyu Li, Xiaoxiu Tan, Yong Xu, Yong Zhao, Yong Lin

Writing – original draft: Guangzhi Wang**.**

Writing – review & editing: Lu Xie, Xingxing Jian**.**

Competing interests

The authors have declared that no competing interests exist.

References

1. Desai DV, Kulkarni-Kale U. T-cell epitope prediction methods: an overview. Methods in molecular biology (Clifton, NJ). 2014;1184:333-64.

2. Goldberg AL, Rock KL. Proteolysis, proteasomes and antigen presentation. Nature. 1992;357(6377):375-9.

3. Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. Protein engineering. 2002;15(4):287-96.

4. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, 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. 2005;35(8):2295-303.

5. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. Journal of immunology (Baltimore, Md : 1950). 2017;199(9):3360-8.

6. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher

J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell systems. 2018;7(1):129-32.e4.

7. Kristensen VN. The Antigenicity of the Tumor Cell - Context Matters. The New England journal of medicine. 2017;376(5):491-3.

8. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic acids research. 2019;47(D1): D339-d43.

9. Reche PA, Reinherz EL. Definition of MHC supertypes through clustering of MHC peptide-binding repertoires. Methods in molecular biology (Clifton, NJ). 2007;409: 163-73.

10. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics. 1999;50(3-4):201-

11. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and updated classification. BMC immunology. 2008;9:1.

12. Ott PA, Hu Z, Keskin DB, Shukla SA, Sun J, Bozym DJ, et al. An immunogenic personal neoantigen vaccine for patients with melanoma. Nature. 2017;547(7662):217-21.

13. Sahin U, Derhovanessian E, Miller M, Kloke BP, Simon P, Lower M, et al. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature. 2017;547(7662):222-6.

14. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nature reviews Immunology. 2018;18(3):168-82.

15. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. Science (New York, NY). 2015;350(6257):207-11.

16. Efremova M, Finotello F, Rieder D, Trajanoski Z. Neoantigens Generated by Individual Mutations and Their Role in Cancer Immunity and Immunotherapy. Frontiers in immunology. 2017;8:1679.

17. Klein L, Hinterberger M, Wirnsberger G, Kyewski B. Antigen presentation in the thymus for positive selection and central tolerance induction. Nature reviews Immunology. 2009;9(12):833-44.

18. Ramsby M, Makowski G. The Proteomics Protocols Handbook2005. 37-48 p.

19. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology. 1982;157(1):105-32.

20. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. Journal of theoretical biology. 1968;21(2):170-201.

21. Grantham R. Amino acid difference formula to help explain protein evolution. Science (New York, NY). 1974;185(4154):862-4.

22. Fraga SJCJoC. Theoretical prediction of protein antigenic determinants from amino acid sequences. 1982;60(20):2606-10.

23. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. Journal of molecular biology. 1983;171(4):479-88.

24. Meek JL. 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. 1980;77(3):1632-6.

25. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MHJS. Hydrophobicity of amino acid residues in globular proteins. 1985;229(4716):834-8.

26. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Advances in enzymology and related areas of molecular biology.

1978;47:45-148.

27. Deléage G, Roux BJPE. An algorithm for protein secondary structure prediction based on class prediction. 1987;1(4):289.

28. Hersh RTJJoMC. Atlas of Protein Sequence and Structure, 1966. 1965;13(2):337-.

29. Jones DD. Amino acid properties and side-chain orientation in proteins: a cross correlation appraoch. Journal of theoretical biology. 1975;50(1):167-83.

30. Zhao G, London E. Strong correlation between statistical transmembrane tendency and experimental hydrophobicity scales for identification of transmembrane helices. The Journal of membrane biology. 2009;229(3):165-8.

31. Janin J. Surface and inside volumes in globular proteins. Nature. 1979;277(5696):491-2.

32. Green JR, Korenberg MJ, David R, Hunter IW. Recognition of adenosine triphosphate binding sites using parallel cascade system identification. Annals of biomedical engineering. 2003;31(4):462-70.

33. Lifson S, Sander C. Antiparallel and parallel beta-strands differ in amino acid residue preferences. Nature. 1979;282(5734):109-11.

34. Shannon CEJBSTJ. A mathematical theory of communication. 1948;27(4):623-56.

35. Kursa MB, Rudnicki WRJJoSS. Feature Selection with Boruta Package. 2010;36(11):1-13.

36. Duan F, Duitama J, Al Seesi S, Ayres CM, Corcelli SA, Pawashe AP, et al. Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. The Journal of experimental medicine. 2014;211(11):2231-

37. Calis JJ, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, et al. Properties of MHC class I presented peptides that enhance immunogenicity. PLoS

computational biology. 2013;9(10): e1003266.

38. Laumont CM, Vincent K, Hesnard L, Audemard E, Bonneil E, Laverdure JP, et al. Noncoding regions are the main source of targetable tumor-specific antigens. Science translational medicine. 2018;10(470).

39. Gonzalez-Galarza FF, McCabe A, Melo Dos Santos EJ, Takeshita L, Ghattaoraya G, Jones AR, et al. Allele Frequency Net Database. Methods in molecular biology (Clifton, NJ). 2018;1802:49-62.

40. Weiskopf D, Angelo MA, de Azeredo EL, Sidney J, Greenbaum JA, Fernando AN, 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. 2013;110(22):E2046-53.

41. Luxenburger H, Grass F, Baermann J, Boettler T, Marget M, Emmerich F, et al. Differential virus-specific CD8(+) T-cell epitope repertoire in hepatitis C virus genotype 1 versus 4. Journal of viral hepatitis. 2018;25(7):779-90.

42. Xia Y, Pan W, Ke X, Skibbe K, Walker A, Hoffmann D, 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. 2019;26(1):73-82.

43. Vahed H, Agrawal A, Srivastava R, Prakash S, Coulon PA, Roy S, et al. Unique Type I Interferon, Expansion/Survival Cytokines, and JAK/STAT Gene Signatures of Multifunctional Herpes Simplex Virus-Specific Effector Memory CD8(+) TEM Cells Are Associated with Asymptomatic Herpes in Humans. Journal of virology. 2019;93(4).

44. Khakpoor A, Ni Y, Chen A, Ho ZZ, Oei V, Yang N, et al. Spatiotemporal Differences in Presentation of CD8 T Cell Epitopes during Hepatitis B Virus Infection. Journal of virology. 2019;93(4).

45. Huth A, Liang X, Krebs S, Blum H, Moosmann A. Antigen-Specific TCR

Signatures of Cytomegalovirus Infection. Journal of immunology (Baltimore, Md : 1950). 2019;202(3):979-90.

46. Owusu Sekyere S, Schlevogt B, Mettke F, Kabbani M, Deterding K, Wirth TC, et al. HCC Immune Surveillance and Antiviral Therapy of Hepatitis C Virus Infection. Liver cancer. 2019;8(1):41-65.

47. Bjerregaard AM, Nielsen M, Jurtz V, Barra CM, Hadrup SR, Szallasi Z, et al. An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. Frontiers in immunology. 2017;8: 1566.

48. Wick DA, Webb JR, Nielsen JS, Martin SD, Kroeger DR, Milne K, et al. Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer. Clinical cancer research: an official journal of the American Association for Cancer Research. 2014;20(5):1125-34.

49. Karasaki T, Nagayama K, Kawashima M, Hiyama N, Murayama T, Kuwano H, et al. Identification of Individual Cancer-Specific Somatic Mutations for Neoantigen-Based Immunotherapy of Lung Cancer. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer. 2016;11(3):324-33.

50. Gros A, Parkhurst MR, Tran E, Pasetto A, Robbins PF, Ilyas S, et al. Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. Nature medicine. 2016;22(4):433-8.

51. Stronen E, Toebes M, Kelderman S, van Buuren MM, Yang W, van Rooij N, et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. Science (New York, NY). 2016;352(6291):1337-41.

52. Nelde A, Walz JS, Kowalewski DJ, Schuster H, Wolz OO, Peper JK, et al. HLA class I-restricted MYD88 L265P-derived peptides as specific targets for lymphoma

immunotherapy. Oncoimmunology. 2017;6(3): e1219825.

53. Zhang X, Kim S, Hundal J, Herndon JM, Li S, Petti AA, et al. Breast Cancer Neoantigens Can Induce CD8(+) T-Cell Responses and Antitumor Immunity. Cancer immunology research. 2017;5(7):516-23

54. Muller M, Gfeller D, Coukos G, Bassani-Sternberg M. 'Hotspots' of Antigen Presentation Revealed by Human Leukocyte Antigen Ligandomics for Neoantigen Prioritization. Frontiers in immunology. 2017;8: 1367.

55. Balachandran VP, Luksza M, Zhao JN, Makarov V, Moral JA, Remark R, et al. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. Nature. 2017;551(7681):512-6.

56. Matsuda T, Leisegang M, Park JH, Ren L, Kato T, Ikeda Y, 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. 2018;24(21):5357-67.

57. Sonntag K, Hashimoto H, Eyrich M, Menzel M, Schubach M, Docker D, 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. 2018;16(1):23.

58. Shannon CEJBSTJ. A mathematical theory of communication. 1948;27(4):623-56.

59. Kuhn MJASCL. Caret: Classification and regression training. 2015;129(1):291–295.

60. Liaw A, Wiener M. Classification and Regression by RandomForest.2001.

61. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics (Oxford, England). 2005;21(20):3940-1.

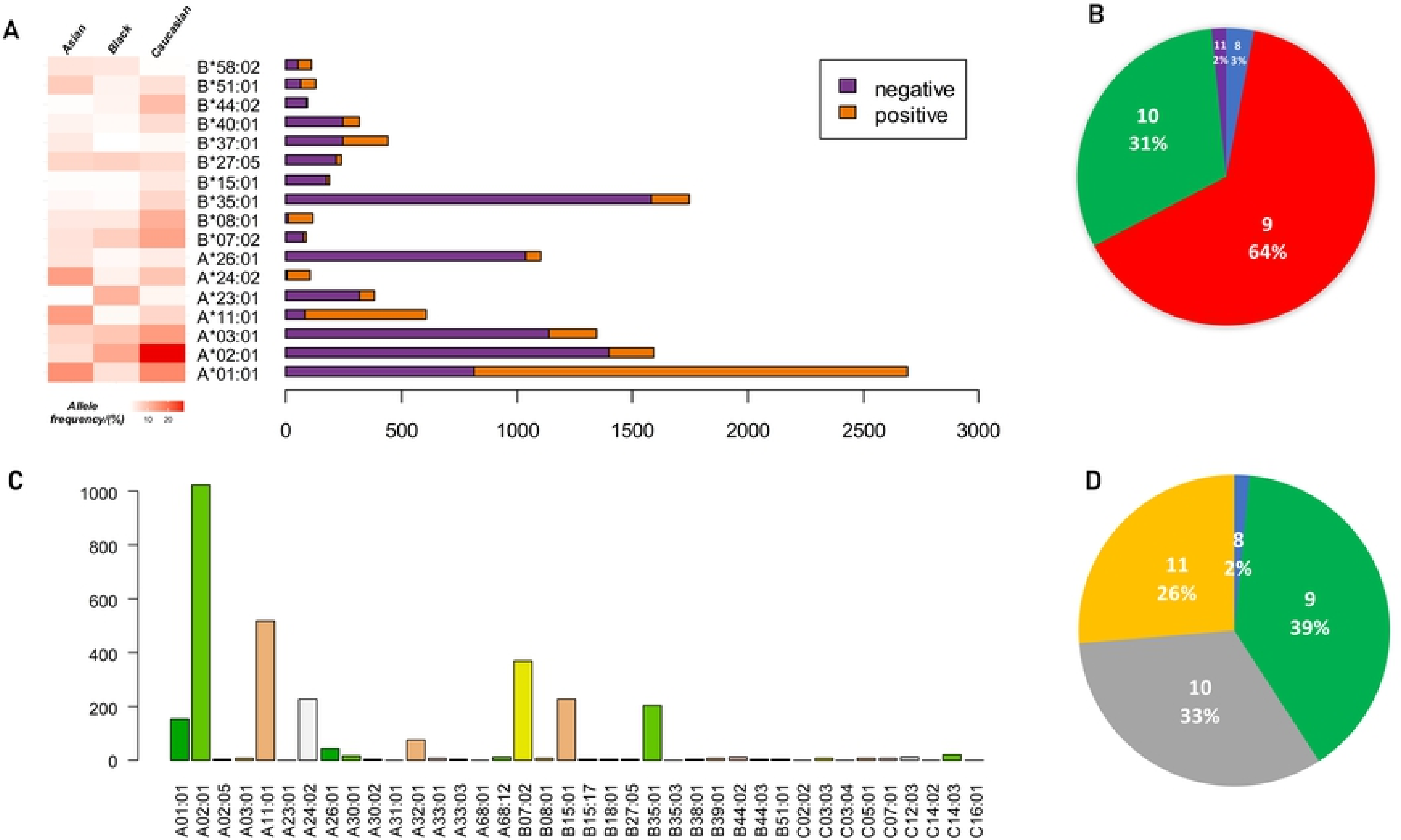
Supporting information captions

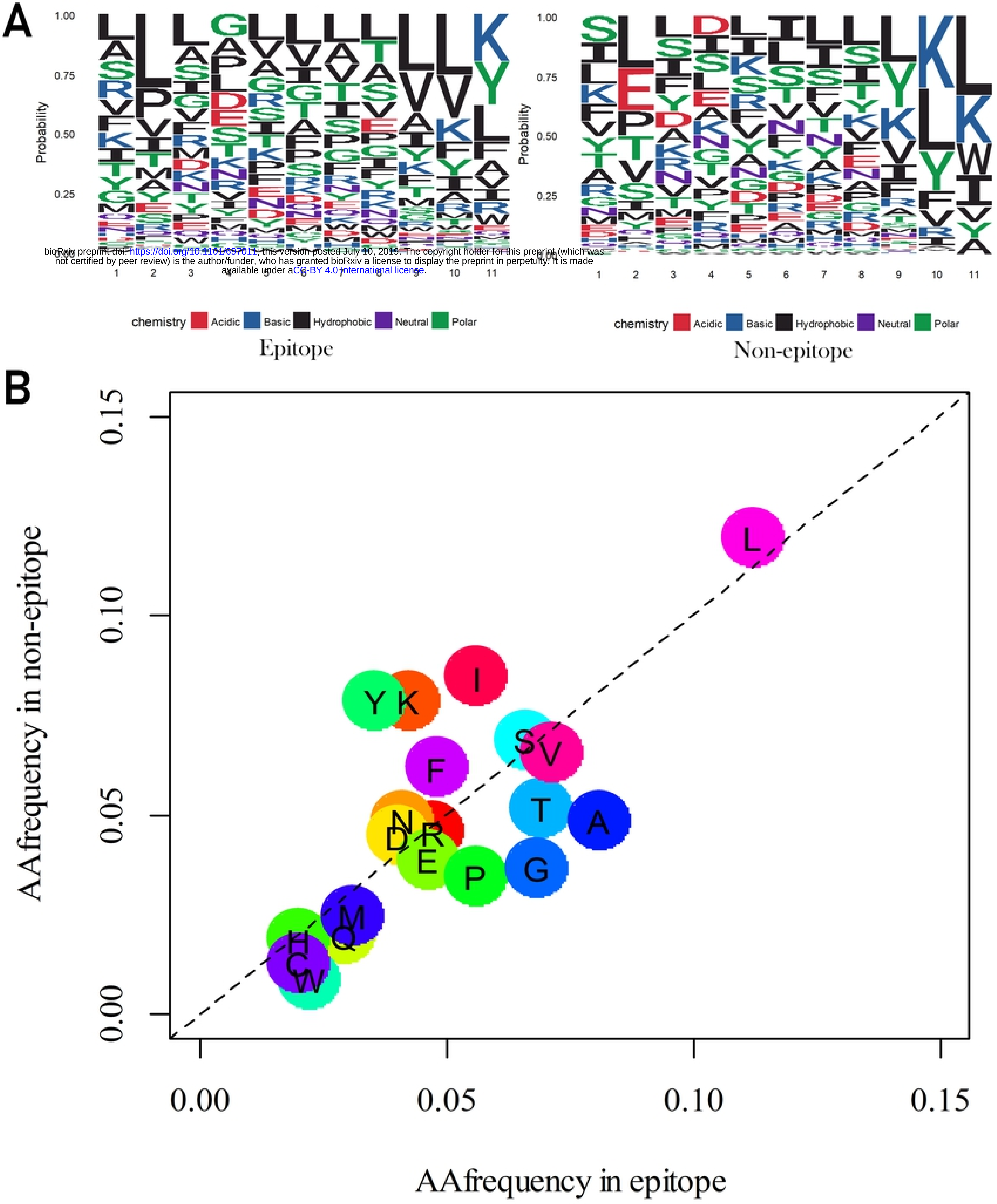
S1 Table **IEDB antigen epitopes summary.** Detailed description of 17 HLA molecules which collected from IEDB. (XLSX)

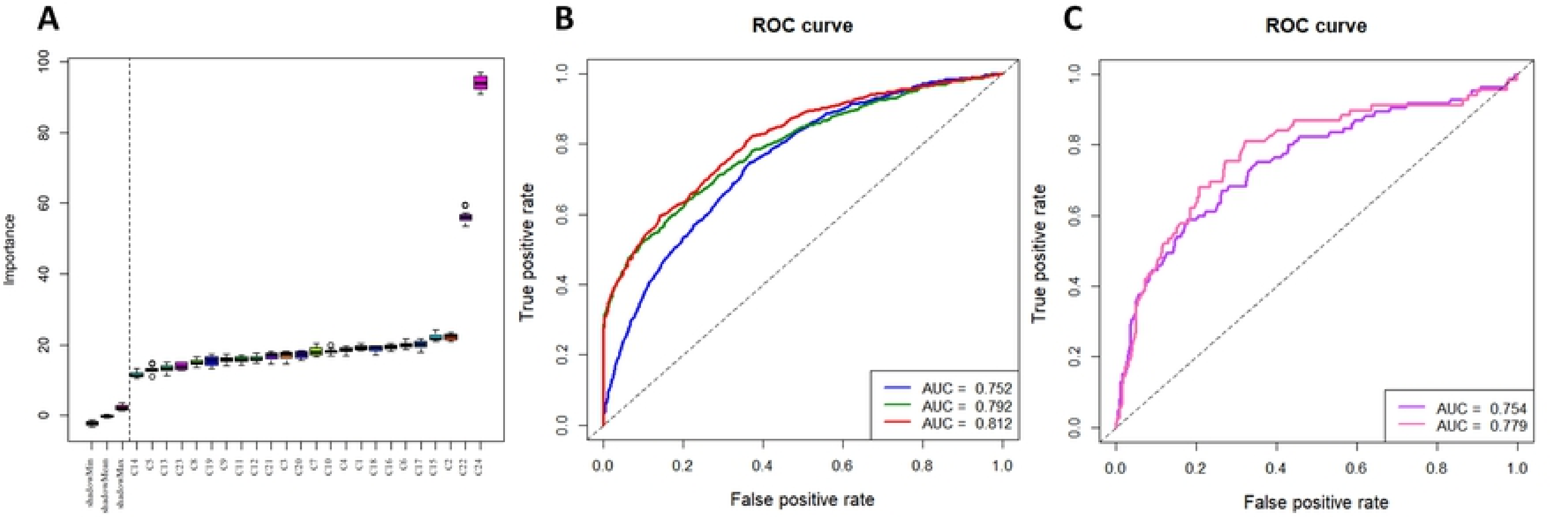
S2 Table **External validation antigen epitopes summary.** Epitope details of 7 publications. (XLSX)

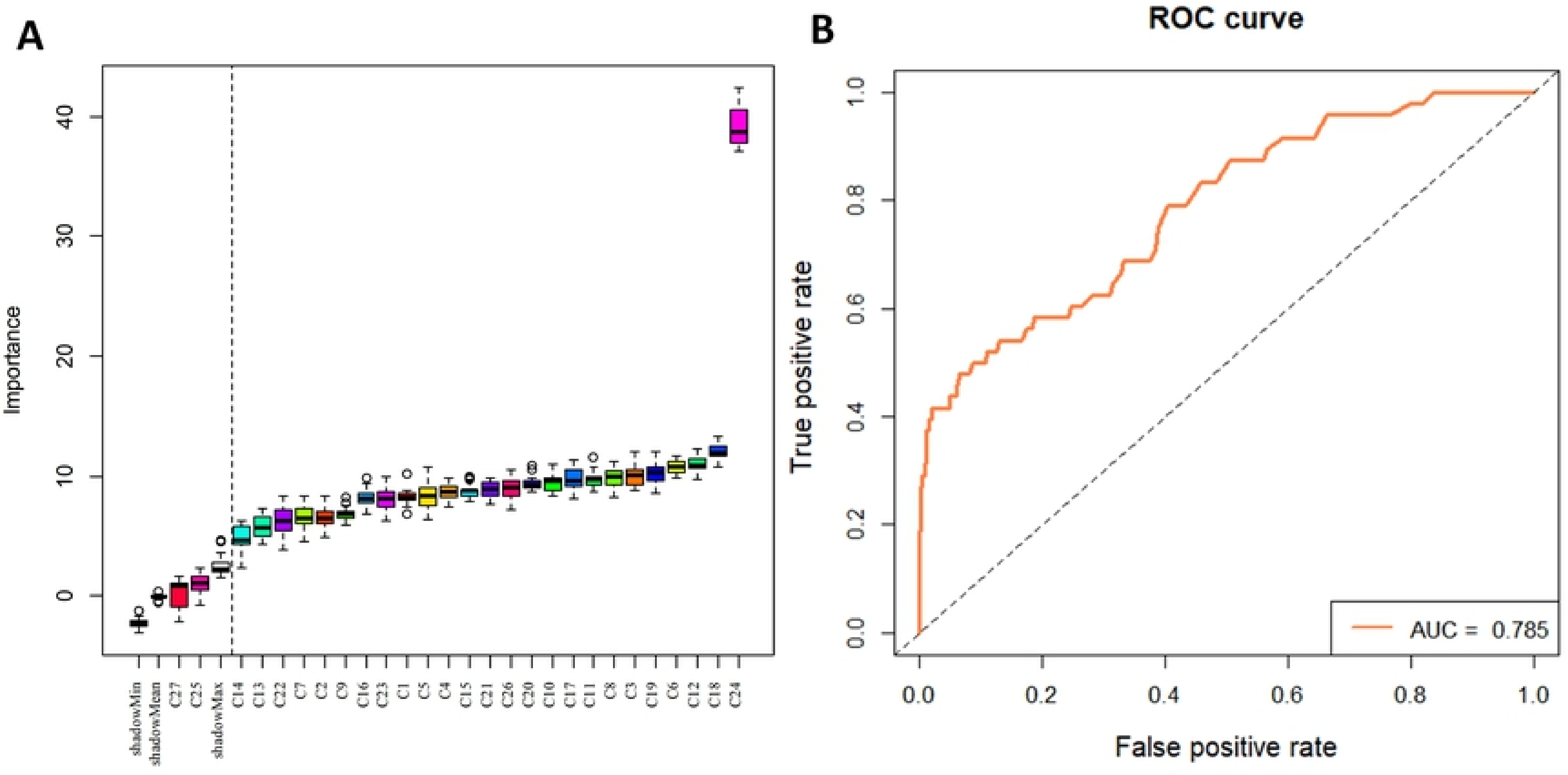
S3 Table **Neoantigen epitopes summary.** Epitope details of 13 publications. (XLSX)

S4 Table **Summary of amino acid characteristics.** For all amino acid characteristics (n=21) that are described in the ExPASy. (XLSX)


A
Asian
Black
Caucasian
B*58:02
B*51:01
B*44:02
B*40:01
B*37:01
B*27:05
B*15:01
B*35:01
B*08:01
B*07:02
A*26:01
A*24:02
A*23:01
A*11:01
A*03:01
A*02:01
A*01:01
Allele frequency(%)
10
20
negative
positive
0
500
1000
1500
2000
2500
3000
B
11 2%
8 3%
10 31%
9 64%
C
1000
800
600
400
200
0
A01:01
A02:01
A02:05
A03:01
A11:01
A23:01
A24:02
A26:01
A30:01
A30:02
A31:01
A32:01
A33:01
A33:03
A68:01
A68:12
B07:02
B08:01
B15:01
B15:17
B18:01
B27:05
B35:01
B35:03
B38:01
B39:01
B44:02
B44:03
B51:01
C02:02
C03:03
C03:04
C05:01
C07:01
C12:03
C14:02
C14:03
C16:01
D
8 2%
11 26%
9 39%
10 33%


A
1.00
0.75
0.50
0.25
0.00
Probability
1 2 3 4 5 6 7 8 9 10 11
chemistry 🟥 Acidic 🟦 Basic ⬛ Hydrophobic 🟪 Neutral 🟩 Polar
Epitope
1.00
0.75
0.50
0.25
0.00
Probability
1 2 3 4 5 6 7 8 9 10 11
chemistry 🟥 Acidic 🟦 Basic ⬛ Hydrophobic 🟪 Neutral 🟩 Polar
Non-epitope
B
0.00
0.05
0.10
0.15
AA frequency in non-epitope
0.00
0.05
0.10
0.15
AA frequency in epitope
L
I
Y
K
S
V
F
T
A
N
R
D
E
P
G
M
Q
H
C
W


A
Importance
100
80
60
40
20
0
shadowMin
shadowMean
shadowMax
C14
C5
C13
C23
C8
C19
C9
C11
C12
C21
C3
C20
C7
C10
C4
C1
C18
C16
C6
C17
C15
C2
C22
C24
B
ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.752
AUC = 0.792
AUC = 0.812
C
ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.754
AUC = 0.779


A
Importance
0
10
20
30
40
shadowMin
shadowMean
C27
C25
shadowMax
C14
C13
C22
C7
C2
C9
C16
C23
C1
C5
C4
C15
C21
C26
C20
C10
C17
C11
C8
C3
C19
C6
C12
C18
C24
B
ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.785