OXFORD

Antibody Therapeutics, 2024, 7(2), 177–186
https://doi.org/10.1093/abt/tbae011
Advance access publication date 17 June 2024
Research Article

An integrated database of experimentally validated major histocompatibility complex epitopes for antigen-specific cancer therapy

Satoru Kawakita[1], Aidan Shen[1], Cheng-Chi Chao[2], Zhaohui Wang[1], Siliangyu Cheng[3], Bingbing Li[4], Chongming Jiang[1,*]

[1]Department of Precision Medicine, Terasaki Institute for Biomedical Innovation, Los Angeles, CA 90024, United States
[2]Department of Pipeline Development, Biomap, Inc., Palo Alto, CA 94303, United States
[3]Quantitative and Computational Biology Department, University of Southern California, Los Angeles, CA 90089, United States
[4]Autonomy Research Center for STEAHM (ARCS), California State University Northridge, Northridge, CA 91324, United States
*Corresponding author. Department of Precision Medicine, Terasaki Institute for Biomedical Innovation, Los Angeles, CA 90024, United States.
E-mail: chongming.jiang@terasaki.org

Abstract

Cancer immunotherapy represents a paradigm shift in oncology, offering a superior anti-tumor efficacy and the potential for durable remission. The success of personalized vaccines and cell therapies hinges on the identification of immunogenic epitopes capable of eliciting an effective immune response. Current limitations in the availability of immunogenic epitopes restrict the broader application of such therapies. A critical criterion for serving as potential cancer antigens is their ability to stably bind to the major histocompatibility complex (MHC) for presentation on the surface of tumor cells. To address this, we have developed a comprehensive database of MHC epitopes, experimentally validated for their MHC binding and cell surface presentation. Our database catalogs 451 065 MHC peptide epitopes, each with experimental evidence for MHC binding, along with detailed information on human leukocyte antigen allele specificity, source peptides, and references to original studies. We also provide the grand average of hydropathy scores and predicted immunogenicity for the epitopes. The database (MHCepitopes) has been made available on the web and can be accessed at https://github.com/jcm1201/MHCepitopes.git. By consolidating empirical data from various sources coupled with calculated immunogenicity and hydropathy values, our database offers a robust resource for selecting actionable tumor antigens and advancing the design of antigen-specific cancer immunotherapies. It streamlines the process of identifying promising immunotherapeutic targets, potentially expediting the development of effective antigen-based cancer immunotherapies.

Statement of Significance: Current peptide-based cancer immunotherapy is challenged with the limited availability of immunogenic epitopes. To facilitate the discovery of immunogenic peptides, we developed a new, comprehensive database that contains both experimental and theoretical data on experimentally-verified MHC-binding epitopes.

Keywords: major histocompatibility complex; epitopes; cancer immunotherapy; immunogenicity

Introduction

Over the last decade, there has been an increasing interest in cancer immunotherapy as a strategy to treat cancer, which remains one of the leading causes of death worldwide [1]. Cancer immunotherapy works by reactivating the patient's immune system to trigger anticancer mechanisms targeting tumor cells [2, 3]. Cancer immunotherapy takes on various forms including but not limited to antibody, small molecule, adoptive cell therapy, oncolytic virus, and vaccine. While substantial progress has been made in enhancing the potency of immunotherapy in combating cancer, immunotherapeutic efficacy is still limited due to the scarcity of highly immunogenic antigens and the lack of complete understanding of various immune-resistant mechanisms. Therefore, discovering new immunogenic antigens and deciphering the immune mechanisms are of utmost importance to advance the current cancer immunotherapy.

Typically, peptide-based cancer vaccines employ cancer antigens, which can trigger B-cell and T-cell-mediated immune responses against cancer [4–6]. Upon recognition of cancer antigens, B cells can secrete antibodies that bind to cancer antigens, whereas T cells, particularly $CD8^+$ T cells and $CD4^+$ T cells, engage Class I and Class II major histocompatibility complex (MHC)-bound [i.e. human leukocyte antigen (HLA)-bound for humans] epitopes, respectively, which are presented on the cell surface. This can lead to either indirect or direct killing of target cancer cells. Sources of cancer antigens include but are not limited to mutated or overexpressed proteins, differentiation antigens, altered glycolipids and glycoproteins, and oncogenic viruses [7]. Also, cancer antigen can be broadly categorized into tumor-associated antigen (TAA) and tumor-specific antigen (TSA) with the latter being expressed only on cancer cells and not in normal tissues [8]. With the recent scientific advances such as

Received: January 26, 2024. **Revised:** April 18, 2024

the discovery of new immune-checkpoint blockades to enhance the efficacy of cancer vaccines [9], there has been rekindling and increasing interest in developing peptide-based, therapeutic immunotherapies [10].

To facilitate the development of anti-cancer strategies targeting cancer antigens, there is a need for a comprehensive database on cancer epitopes that are presented on cancer cell surfaces by MHC Class I or Class II molecules. Currently, one of the largest databases on cancer epitopes is the Cancer Epitope Database and Analysis Resource (CEDAR), which was established in January 2023 [11]. It serves as a companion site to Immune Epitope Database (IEDB), which has comprehensive epitope data for a variety of diseases ranging from allergy and transplantation to infectious disease, autoimmunity, and cancer [12]. Unlike IEDB, CEDAR's data include cancer-specific epitopes curated from the literature. While CEDAR is likely one of the main cancer epitope databases publicly accessible, there are other databases on cancer epitopes that have been made available on the web, including TANTIGEN 2.0 [13], SYFPEITHI [14], MHCBN 4.0 [15], and EPIMHC [16]. Undoubtedly, all of these databases have contributed substantially to the current understanding of cancer peptides and continue to serve as important sources of knowledge for cancer peptide research. The construction of each database follows a similar but different approach in which information is curated from literature, existing databases, or combination of both with different screening criteria. In addition to the curation method, the time of publication varies substantially from 1999 for SYFPEITHI to the recent introduction of CEDAR in 2023. As a result, there are likely both redundant and distinct sets of cancer peptides cataloged in the databases. However, the benefit of combining these databases to improve the comprehensiveness of cancer peptide database remains elusive. Moreover, despite the aforementioned available resources, prediction and selection of cancer-specific epitopes remain a major challenge for the immunotherapy field because the prediction accuracy is often low, exacerbated by the often-differing data given by the disparate databases.

To facilitate epitope-dependent immunotherapy development, we set out to amalgamate all of the aforementioned databases and also manually curate from literature to create an up-to-date, comprehensive and centralized MHC-binding cancer peptide database, the first of this kind. We compiled MHC epitopes from the data sources, supplemented with relevant recent literature on cancer MHC epitopes, and firstly calculated immunogenicity scores and hydropathy values for the cataloged epitopes to augment their utility. In this work, we describe the data curation and processing processes for the newly developed database and present a landscape of the curated cancer epitopes with summary and descriptive statistics. The final curated database with immunogenicity scores and hydropathy values is accessible via MHCepitopes (https://github.com/jcm1201/MHCepitopes.git). We anticipate that the database will serve as a knowledge base and resources for computational applications and the research communities to identify potential target candidates to develop effective cancer immunotherapies.

Materials and methods

Data curation

We screened and extracted peptide data from studies related to MHC peptides available on PubMed and Web of Science using the following keywords: peptid*, cancer/tumo*r, human, HLA, which related to the experiments validated MHC peptides. Additionally, we obtained MHC-binding peptide data from multiple public datasets, including two large repositories hosted by the National Institutes of Health including IEDB and CEDAR using data up to 20 November 2023. Peptide search on IEDB was performed by applying the following search criteria: linear peptide, human as the host, positive for MHC ligand assay, MHC Class I or II, and cancer as the disease. A similar set of filters were applied to search for relevant epitopes on CEDAR while including all epitope sources and any cancer type. Additional data on MHC-binding epitopes were curated from web-accessible databases that have been developed by academic institutions: TANTIGEN 2.0 [13] (http://projects.met-hilab.org/tadb/index.php), SYFPEITHI [14] (http://www.syfpeithi.de/), MHCBN 4.0 [15] (https://webs.iiitd.edu.in/raghava/mhcbn/index.html), and EPIMHC [16] (http://bio.med.ucm.es/epimhc/).

Data processing and database development

Extracted data were then processed further to eliminate any missing or poor data entries for any of the following variables: HLA class, epitope sequence, and data source. In cases where there were overlapping results between different data sources, the data source with the most complete information was selected for the final database construction. Finally, a total of 451 065 T-cell epitopes with multiple peptide features were registered in our library, which has been made accessible at MHCepitopes (https://github.com/jcm1201/MHCepitopes.git).

Prediction of immunogenicity scores

Theoretical immunogenicity scores for HLA Class I epitopes were predicted using a previously described method [17]. Briefly, the algorithm calculates immunogenicity scores based on the enrichment of amino acids in immunogenic versus non-immunogenic epitopes as determined by machine learning as well as the importance scores of different amino acid positions within the MHC Class I sequence. Mathematically, the immunogenicity score (S) of a peptide ligand (L) for that HLA molecule (H) is computed as follows:

$$S(H,L) = \sum_{p=1}^{9} E_{A(L,p)} \times I_p \times M(H,p)$$

where p is a position in the ligand (L), E represents the log enrichment score for the amino acid at that position $A(L, p)$, I_p is the importance of that position, and $M(H, p)$ is the masking status of anchor positions on that HLA (i.e. 0 if masked otherwise 1).

For HLA Class II peptide ligands, we employed a method described by Dhanda *et al.* [18], in which immunogenicity scores are calculated using an artificial neural network model trained on HLA binding affinity data. For Class I, higher values indicate greater immunogenicity whereas for Class II, lower values mean more immunogenic. For stratification purposes, high immunogenicity groups were defined as those higher than zero for Class I and lower than 50 for Class II.

Grand average of hydropathicity index (Gravy score) calculation

Gravy scores are calculated by taking the sum of the hydropathy values (i.e. hydrophilicity and hydrophobicity) of all the amino acids divided by the sequence length [19]. A positive value indicates that the peptide is hydrophobic while a negative value means that the peptide is hydrophilic.

Epitope sequence visualization via sequence logos

Sequence logos display patterns in conserved residues in amino acid sequences in case of protein molecules [20]. These highly conserved elements often indicate structural or functional

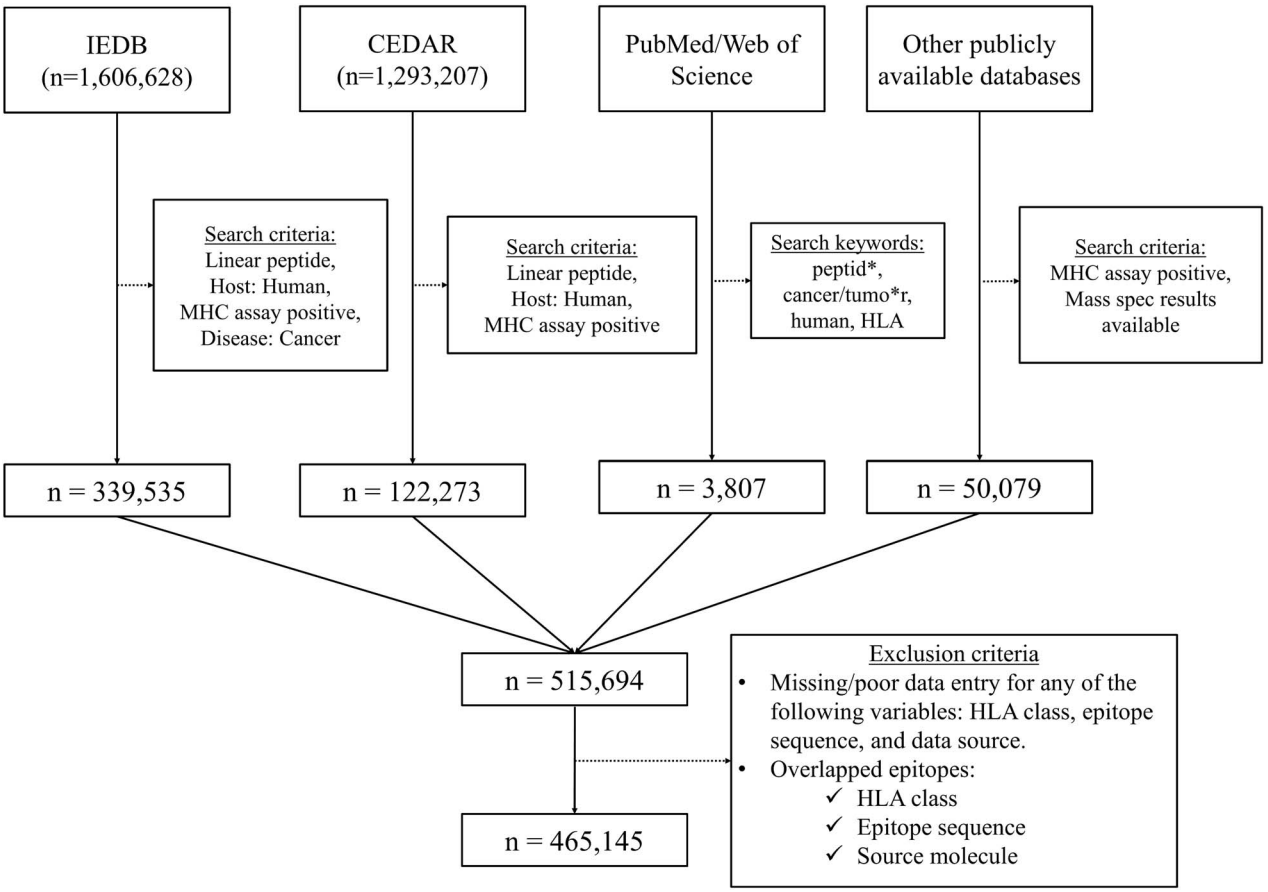

Figure 1. The data curation process for the construction of the MHC peptide library.

importance. The overall height (measured in bits) represents the sequence conservation at that position while the height of each symbol corresponds to the relative frequency of the amino acid at that particular position. The symbols are color-coded to represent the chemical properties of amino acids: polar (green), basic (blue), acidic (red), and hydrophobic (black).

Statistical analysis

Python 3.9.7 and R version 4.2.2 (2022-10-31 ucrt) were used for all data processing and statistical analysis. Mean and median values of two groups were compared using independent-sample t-tests and Wilcoxon tests, respectively. Pearson correlation analysis was performed to assess correlation between two continuous variables with the correlation coefficient (R) and associated P-value. The statistical tests used are specified in the associated text or figure legends. Statistical significance was determined using a P-value of <0.05 as a cut-off value.

Results

Construction of an MHC-peptide database

To develop a well-structured, comprehensive database of MHC-binding epitopes, we curated data from multiple massive data sources (Fig. 1). IEDB is a freely available resource that catalogs experimental data on antibodies and T cell epitopes studied in humans in the context of infectious disease, allergy, autoimmunity, and transplantation. CEDAR is a companion site to IEDB, but unlike IEDB, its focus centers around cancer, providing a comprehensive collection of cancer-specific epitopes curated from literature. More data on MHC epitopes were also obtained from other databases that have been curated by research groups at academic institutions and made freely accessible on the web. In addition to the aforementioned data sources, we performed a systematic literature search using a set of search keywords on PubMed and Web of Science to find relevant literature and manually parsed the text for MHC epitopes. Once compiled, the dataset containing 525 694 records underwent a rigorous filtering process using a set of exclusion criteria to exclude any missing values or poor data entry and overlaps between different data sources. As expected, we observed a significant number of overlapped peptides between the databases, and in such cases, we kept the data source with the most complete information. It is similarly important to mention that there were those epitopes uniquely found in each of the data sources, which verifies the importance of compiling the data sources to build a comprehensive database. The final dataset resulted in 465 145 epitopes with their key information and characteristics including amino acid sequence, HLA allele, HLA class, source molecule, data source, antigen type, and epitope length (Fig. 1). Furthermore, we augmented the library with immunogenicity and gravy scores. The final dataset has been made accessible under MHCepitopes (https://github.com/jcm1201/MHCepitopes.git), and also a prototype version of search engine system for this database is available at https://c2acs761.caspio.com/dp/3c01b00082e291b8cb0146e4b49a.

Profiling of the MHC-binding epitopes

Next, we characterized the epitopes included in our MHC-peptide database (Fig. 2). Of the 465 145 epitopes, 337 334 were HLA Class

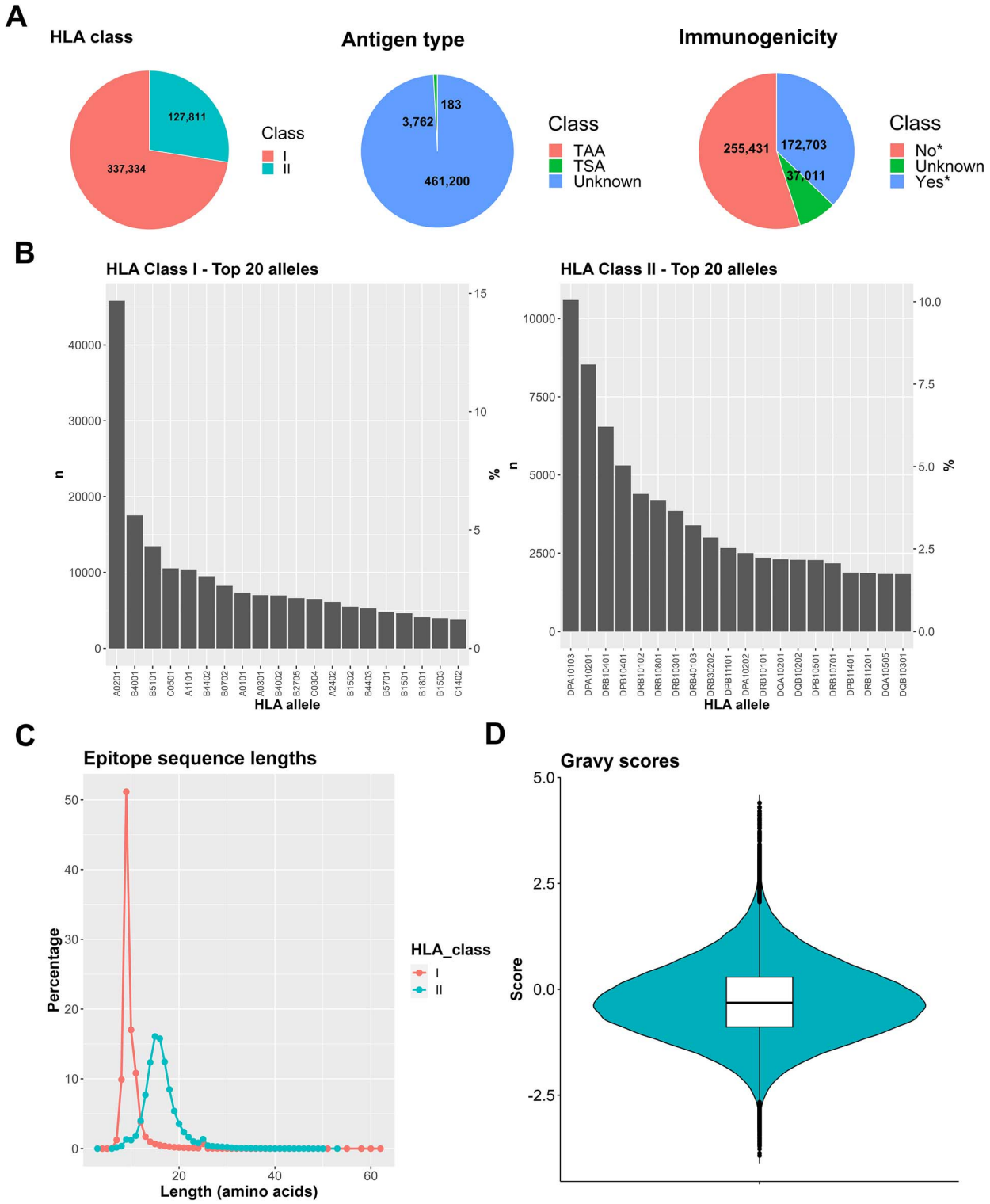

Figure 2. The landscape of the MHC epitope database. (A) Proportional distribution of HLA class, antigen type, and immunogenicity. (B) 20 most common HLA alleles for Classes I and II. The dual *y*-axes represent the count and percentage (%). (C) Line plot showing the distribution of sequence lengths for each HLA class. (D) Violin plot overlaid with a boxplot showing the distribution of Gravy scores.

I and 127 811 were Class II. The epitopes were also classified into either TAA (n = 3762) or TSA (n = 183), while the information was not available in most of the cases (n = 461 200). Immunogenicity scores of the epitopes were calculated separately for HLA Classes I and II and used to categorize them into high vs. low immunogenicity group using the appropriate threshold values (see the Materials and Methods section). The immunogenicity scores for HLA Class I had a normal distribution in contrast to those for HLA Class II where the distribution appeared skewed toward the upper limit of 100 (Supplementary Fig. S1A). The difference in the distribution of immunogenicity scores between the two HLA classes could be attributed to the two different methods used for immunogenicity

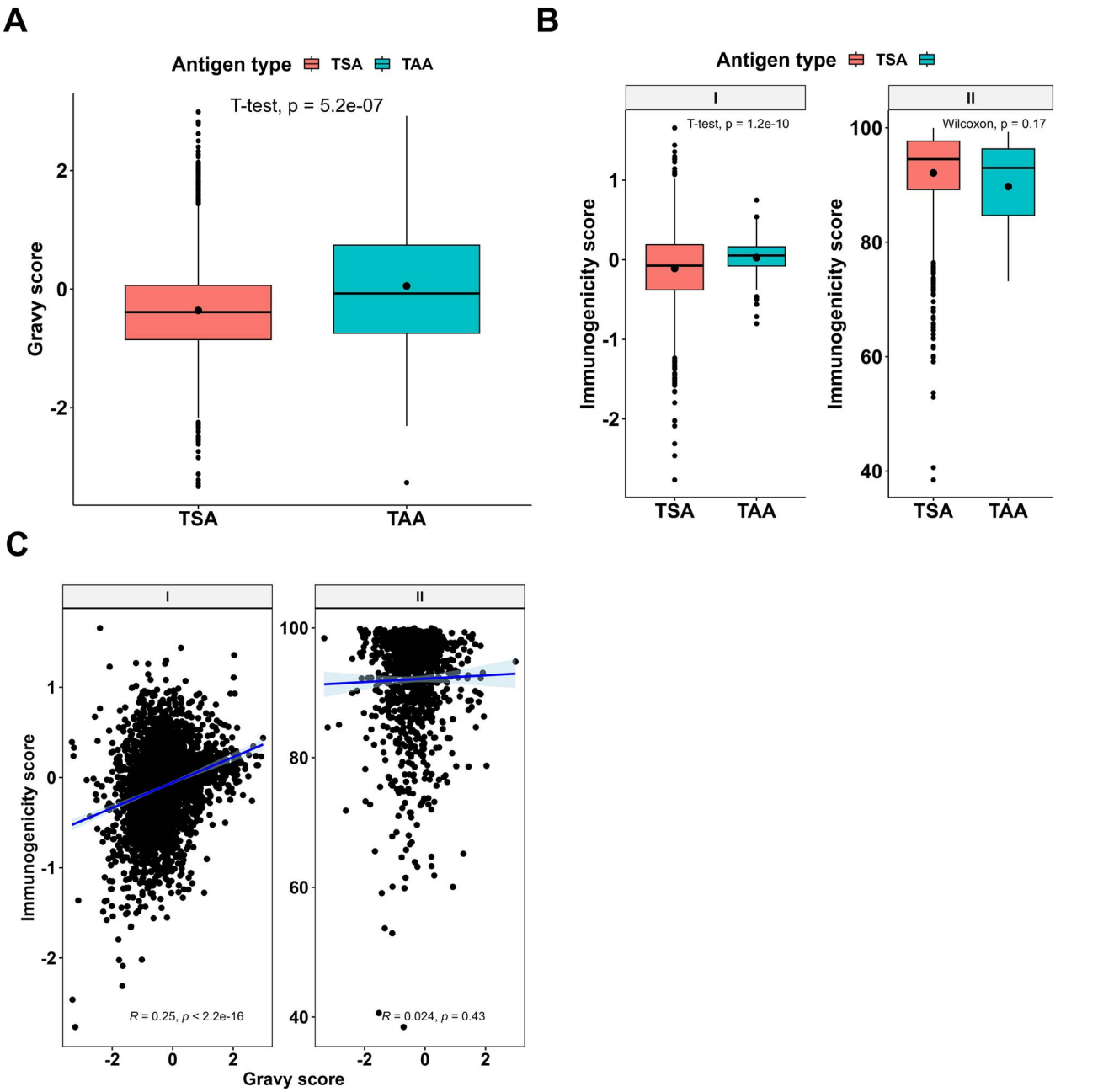

Figure 3. Immunogenicity and hydropathy of MHC epitopes. (A) Difference in the Gravy scores between TAA and TSA. (B) Differences in immunogenicity between TAA and TSA for HLA Classes I and II. (C) Correlation between the Gravy and immunogenicity scores for overall MHC epitopes. Gravy scores were weakly correlated with the immunogenicity of MHC epitopes. Two-group comparisons of their mean or median values were done via the independent-sample *t*-test or Wilcoxon test. Pearson correlation analysis was performed to test for correlation between the two variables *R* and *p*, indicating the correlation coefficient and *P*-value, respectively. *P*-values < 0.05 were considered statistically significant.

prediction (see Materials and Methods for detail). Moreover, MHC Class I and Class II molecules are structurally different and play distinct roles in eliciting immune response while being involved in different T cell activation pathways (i.e. Class I for $CD8^+$ T cells and Class II for $CD4^+$ T cells). 255 431 epitopes belonged to the low immunogenicity group, while 172 703 were considered highly immunogenic. 37 011 had unknown immunogenicity due to a limitation of the method for calculating immunogenicity scores for HLA Class II epitopes with 15 amino acids or less. We identified 20 most common HLA Class I and II alleles for the epitopes (Fig. 2B). The three most common HLA alleles included A*02:01 (n = 45 831; 14.59%), B*40:01 (n = 17 566; 5.59%), and B*51:01 (n = 13 453; 4.28%), for Class I and DPA1*01:03 (n = 10 599; 9.97%), DPA1*02:01 (n = 8529; 8.02%), and DRB1*04:01 (n = 6534; 6.14%) for Class II. HLA Class I and Class II epitopes are known to have different sequence length distributions, and this was also the case for our library (Fig. 2C). The most frequently observed peptide lengths were 9 for Class I (n = 172 612; 51.20%) and 15 for Class II (n = 20 543; 16.10%). Gravy scores are often used to assess hydropathy of epitopes as they are essential for studying the chemical, physical, and structural properties of epitopes to understand associated biological processes. We evaluated the hydropathy values for the MHC epitopes (Fig. 2D). The Gravy scores were roughly evenly distributed along the zero axis with a mean value of −0.340 and a standard deviation of 0.799. Looking at the source molecules of the epitopes, the most common protein

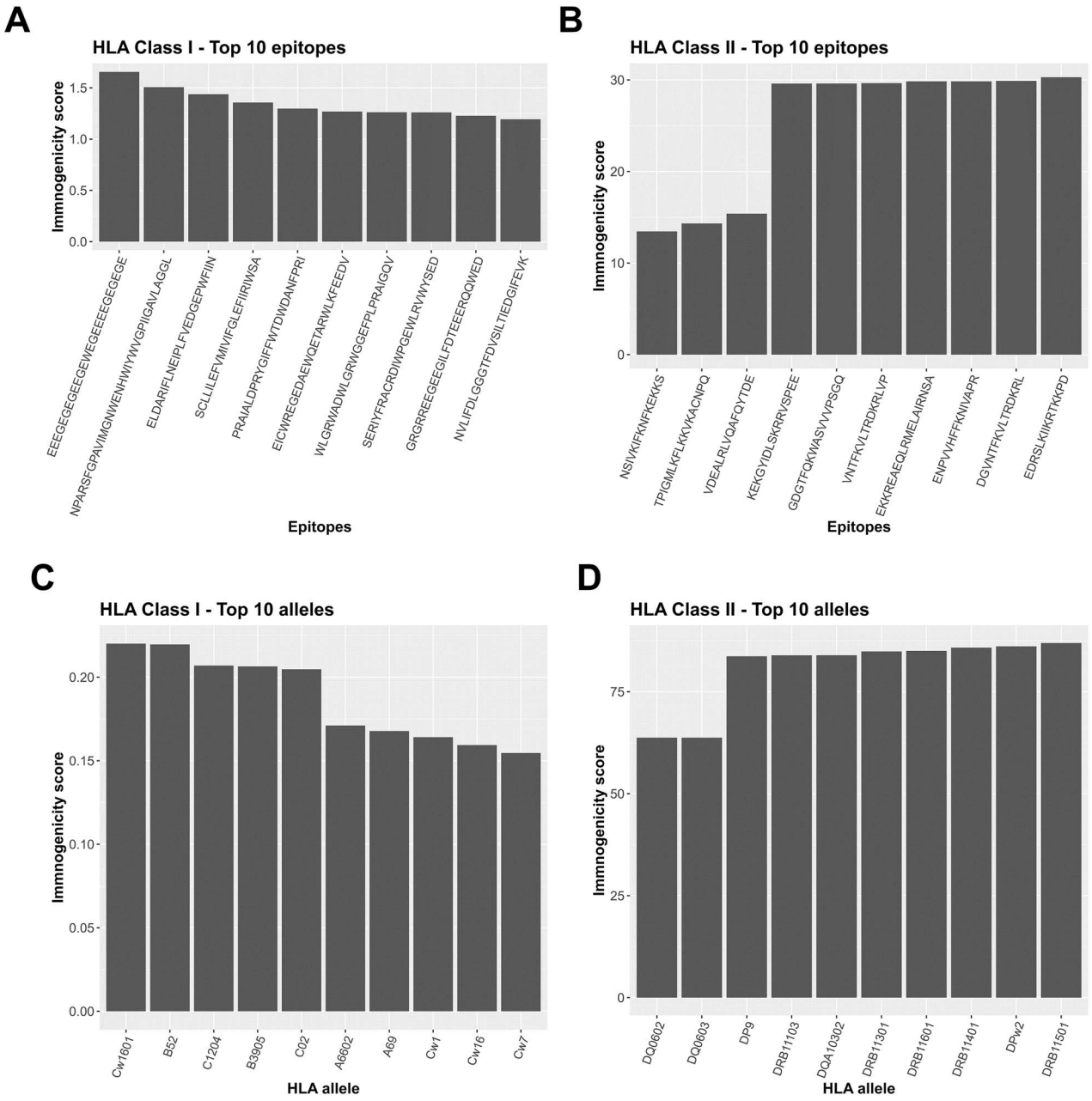

Figure 4. Top 10 HLA alleles and MHC epitopes with the highest immunogenicity scores. (A) Top 10 HLA Class I epitopes. (B) Top 10 HLA Class II epitopes. (C) Top 10 HLA Class I alleles. (D) Top 10 HLA Class II alleles.

was found to be Transferrin receptor protein 1 (n = 1759; 0.38%) followed by Apolipoprotein B-100 (n = 1172; 0.25%). The 10 most common source molecules are listed in Supplementary Fig. S1B.

Hydrophobicity and immunogenicity of TSA versus TAA

Next, we investigated how the hydropathy values and immunogenicity of the epitopes differ between TSA and TAA (Fig. 3). Comparisons of the Gravy scores demonstrated that TAAs were significantly more hydrophobic than TSA with a $P < 0.001$ (Fig. 3A). Similarly, when the immunogenicity scores were significantly higher for TAA compared with TSA for HLA Class I epitopes ($P < 0.001$), while no statistical significance was observed for Class II epitopes ($P = 0.23$). We further performed correlation analysis between the Gravy and immunogenicity scores for each HLA class (Fig. 3C). As expected, while the correlation was not strong for both, it was more statistically significant for Class I (Pearson's correlation coefficient, $R = 0.25$, $P < 0.001$) compared to Class II (Pearson's correlation coefficient, $R = 0.024$, $P = 0.43$). It is important to mention that the relative position and hydrophobicity of individual amino acid residues reportedly play a crucial role in determining immunogenicity [21], while the Gravy scores represent the hydrophobicity of the overall peptide sequences. Therefore, while our finding here sheds light into potential correlation between hydrophobicity and immunogenicity of MHC-binding peptides, it will require future validation studies to investigate the influence of arrangements and distribution of charges across the amino acid sequences of peptides while accounting for various factors such as HLA class and allele specificity.

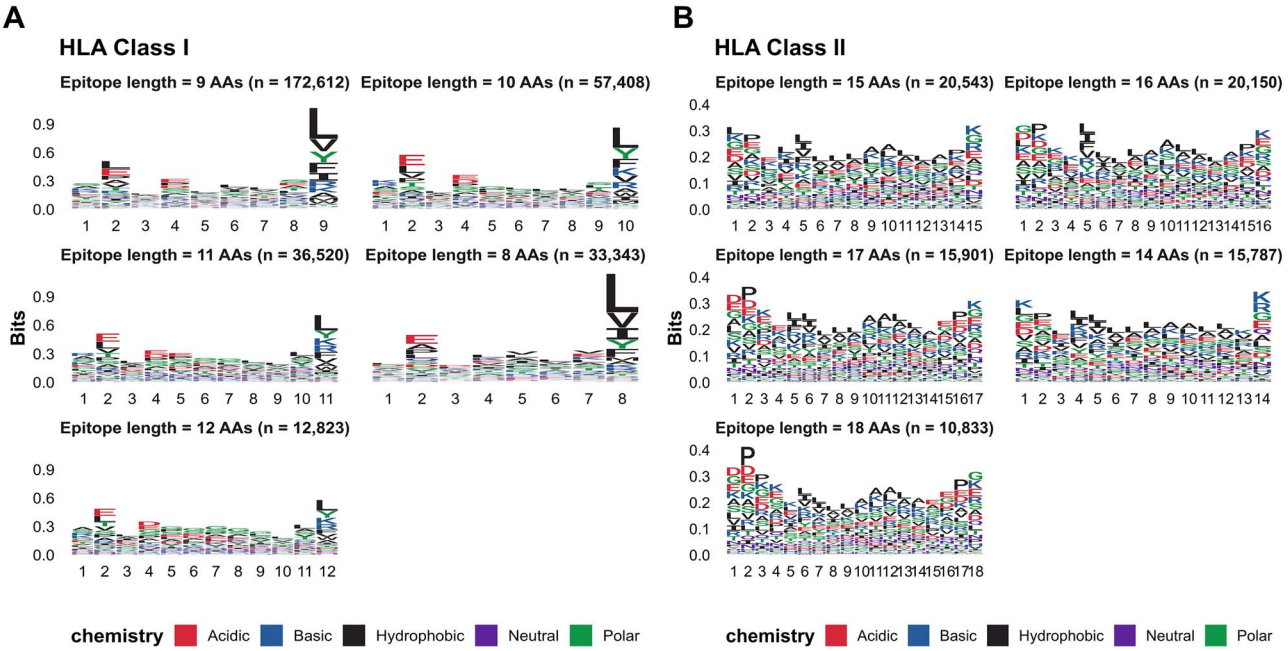

Figure 5. Chemical profiles of amino acid sequences of MHC epitopes. Sequence logos depicting the motifs of MHC epitopes with five most common protein sequence lengths for (A) Class I and (B) Class II. The x-axis displays the amino acid positions for multiple sequence alignment. The y-axis represents the bit scores.

Immunogenic epitopes and HLA alleles

Based on the predicted immunogenicity scores, we ranked each epitope and found those that are potentially highly immunogenic (Fig. 4A and B). The top three epitopes included EEEGEGEGEEGEWEGEEEEGEGEGE (score = 1.66), NPARSFGPAVIMGNWENHWIYWVGPIIGAVLAGGL (score = 1.51), and ELDARIFLNEIPLFVEDGEPWFIIN (score = 1.44) for Class I and NSIVKIFKNFKEKKS (score = 13.5), TPIGMLKFLKKVKACNPQ (score = 14.3), and VDEALRLVQAFQYTDE (score = 15.4) for Class II. Likewise, we identified highly immunogenic HLA alleles based on their peptide ligands (Fig. 4C and D). In cases where there were multiple epitopes found for the same allele, the mean immunogenicity score was calculated and used to define the allele's immunogenic status. The five most immunogenic alleles for Class I were Cw*16:01 (score = 0.220), B*52 (score = 0.220), C*12:04 (score = 0.207), B*39:05 (score = 0.206), and C*02 (score = 0.205), whereas for Class II, they were DQ*06:02 (score = 63.7), DQ*06:03 (score = 63.7), DP*9 (score = 83.7), DRB1*11:03 (score = 83.9), and DQA1*03:02 (score = 84.0).

Profiles of amino acid sequences of MHC epitopes.

We then evaluated whether there are any patterns in the MHC-binding epitopes for Classes I and II by creating sequence logos of the amino acid sequences within a multiple sequence alignment to reveal sequence similarity and significant features using WebLogo (Fig. 5). We selected the five most frequently observed sequence lengths (Fig. 2C) for each class and plotted their amino acid sequences in a multiple sequence alignment fashion. For both Classes I and II, the amino acid residues appeared to be highly conserved across all peptide lengths. Specifically for Class I, the dominant presences of glutamic acid (E) and leucine (L) at position 2 and leucine (L), tyrosine (Y), valine (V), and Phenylalanine (F) at the last position of each peptide length can be appreciated. As expected, when comparing the two classes, they had very distinct sequence patterns. We then further stratified the epitopes into high vs. low immunogenicity groups to investigate whether there are any unique features related to immunogenicity (Fig. 6). The sequence patterns did not change substantially after the stratification for Class I (Fig. 6A and C). In contrast, the high immunogenicity group for MHC Class II showed distinctive amino acid sequence patterns (Fig. 6B), whereas the low immunogenicity group for MHC Class II (Fig. 6D) had only modest differences from all (Fig. 5B). The differences observed for Class II high immunogenicity group could be attributed to the unique motifs associated with the highly immunogenic epitopes; however, further validation is needed as the skewed distribution of highly immunogenic HLA Class II epitopes resulted in a small number of epitopes per class ($n < 100$). Additionally, we analyzed sequence patterns for epitopes with their immunogenicity verified and documented in clinical trial reports (Supplementary Fig. S2). Amino acid residues, particularly for those with a relatively larger sample size ($L = 9$ and 10) displayed similar characteristics when compared with all HLA Class I epitopes (Fig. 5A). There was a very minimal similarity between the rest of the HLA Class I epitopes as well as HLA Class II epitopes when compared with all HLA epitopes (Fig. 5B).

Discussion

Immunotherapy has gained substantial interest over the years as an effective treatment against cancer. However, its wide application and efficacy have been hindered by the lack of immunogenic antigens/peptides that sufficiently elicit the body's immune response specifically against cancer cells. To this end, our efforts have been made to construct databases such as CEDAR, cataloging cancer peptides/epitopes that may potentially serve as targets for cancer immunotherapy. However, currently these existing databases are disparate in nature, posing a significant challenge for the research communities to make the best use of all the data available. To address this problem, we curated epitope data on MHC-binding peptides from these disparate data

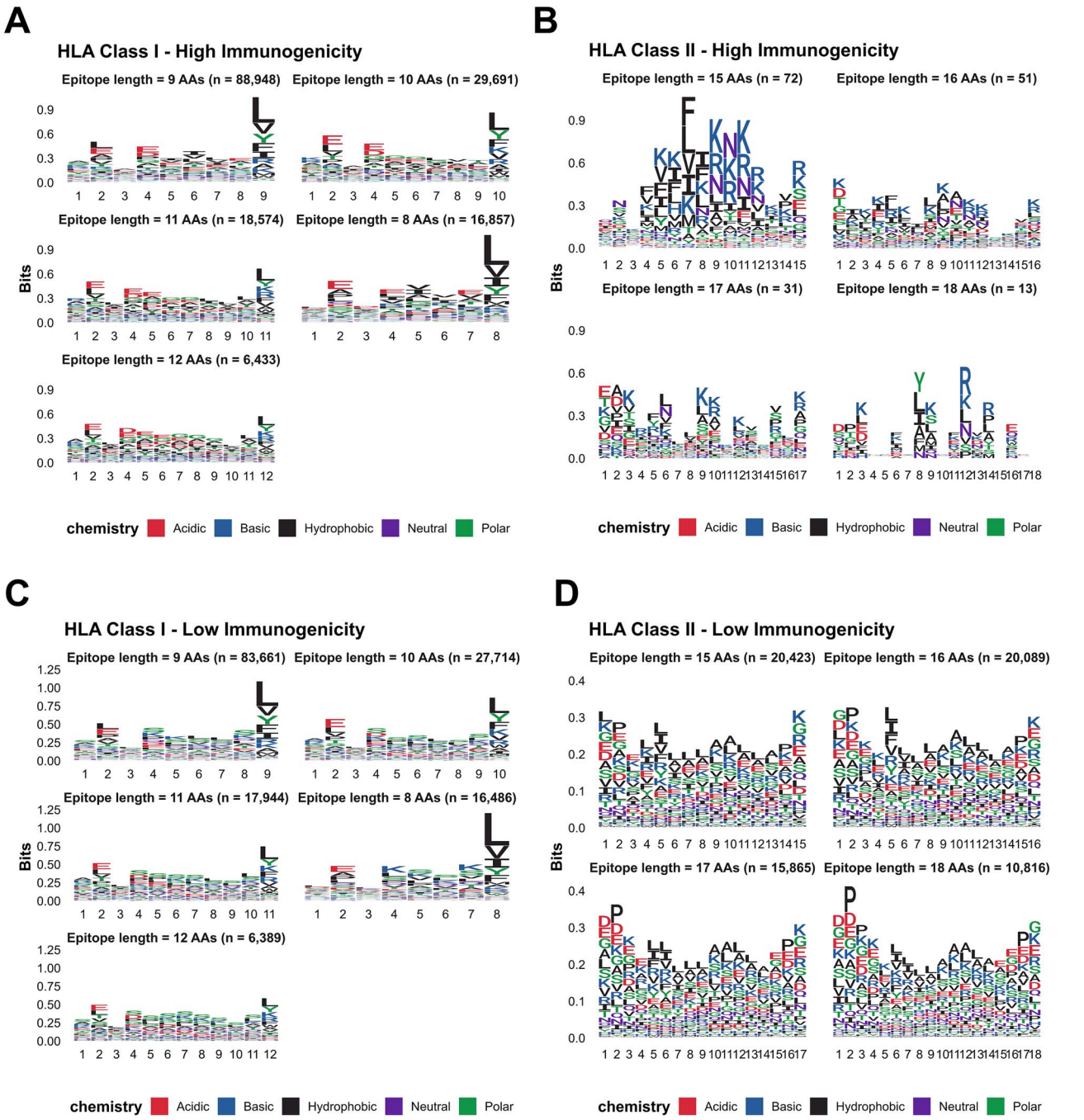

Figure 6. Amino acid sequences for MHC epitopes between high and low immunogenicity. Sequence logos depicting the motifs of MHC epitopes stratified by high vs. low immunogenicity. The x-axis displays the amino acid positions for multiple sequence alignment. The y-axis represents the bit scores. (A) HLA Class I epitopes with high immunogenicity. (B) HLA Class II epitopes with high immunogenicity. (C) HLA Class I MHC epitopes with low immunogenicity. (D) HLA Class II MHC epitopes with low immunogenicity.

sources and compiled them into one central database, where all the epitopes have been experimentally verified to bind MHC molecules. In addition to the basic epitope information and experimental results, immunogenicity scores and hydropathy values of the epitopes were added to further enhance the practical utility of the database.

While the developed database presents a highly comprehensive repertoire of MHC peptide ligands, there are a few limitations worth mentioning. One is that the immunogenicity scores of these MHC epitopes were predicted using the prediction algorithms available from IEDB Analysis Resource [17, 18], which may need to be further experimentally validated by T-cell assays. In fact, the current version of the database includes epitopes that may or may not trigger T-cell response as there is limited availability of pertinent results from T-cell assay data. It is also worth mentioning that there have been recent developments in the immunogenicity prediction methods for both Class I [22, 23] and Class II [24, 25]. However, currently there is no consensus as to what the best method is as their comparative performance remains elusive. Secondly, there was limited information available on the antigen type—whether it is TSA or TAA. Therefore, in the future, the database will likely undergo multiple updates as (i) a benchmarking study comparing the performance of existing immunogenicity prediction systems, (ii) more data on

T-cell activation results, and (iii) more information on antigen type classification become available. The addition of enhanced functionalities to the web application in order to improve user experience such as T-cell epitope information and epitope BLAST function is also forthcoming.

In summary, we developed a new, comprehensive database that contains both experimental and theoretical data on experimentally verified MHC-binding epitopes. We expect that as our database evolves, it will potentially facilitate the development of peptide-based cancer immunotherapies by allowing the researchers to (i) readily take on data-intensive applications such as building machine learning models and (ii) more efficiently identify and select therapeutic candidates.

Acknowledgements

We would like to thank IEDB, CEDAR for letting us use the sequencing data. We would like to give special thanks to Dr. Xiling Shen for his valuable discussions and critical feedback. We especially thank Xiuying Li for the valuable suggestions.

Author contributions

Satoru Kawakita (Data curation [Equal], Formal analysis [Equal], Investigation [Equal], Methodology [Equal], Validation [Equal], Writing—original draft [Equal]), Aidan Shen (Investigation [Equal], Writing—review & editing [Equal]), Cheng-chi Chao (Investigation [Supporting], Methodology [Supporting], Writing—review & editing [Equal]), Zhaohui Wang (Visualization [Supporting], Writing—review & editing [Equal]), Siliangyu Cheng (Investigation [Supporting], Writing—review & editing [Equal]), Bingbing Li (Writing—review & editing [Equal]), and Chongming Jiang (Conceptualization [Equal], Data curation [Lead], Funding acquisition [Equal], Methodology [Equal], Project administration [Lead], Resources [Equal], Software [Lead], Supervision [Lead], Validation [Equal], Visualization [Equal], Writing—original draft [Equal], Writing—review & editing [Lead]). Chongming Jiang conceived the project. Satoru Kawakita, Aidan Shen, and Chongming Jiang prepared and analyzed the results. Satoru Kawakita and Chongming Jiang evaluated the conclusions and wrote the manuscript. Chongming Jiang, Cheng-Chi Chao, Zhaohui Wang, Siliangyu Cheng, and Bingbing Li reviewed and revised the content. All authors read and approved the final manuscript.

Supplementary data

Supplementary data are available at *ABT* Online.

Conflict of interest

C.C.C. was employed by the company Biomap, Inc. The remaining authors declare that the research was conducted without commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work is supported by the National Institutes of Health, United States (NIH) (R01 DK119795 and R35 GM122465).

Ethics and consent statement

Not applicable.

Animal research statement

Not applicable.

Data availability

All data available in this study are publicly available. Please see section “Materials and Methods” for more details.

References

1. Mattiuzzi C, Lippi G. Current cancer epidemiology. *J Epidemiol Glob Health* 2019; **9**: 217–22. https://doi.org/10.2991/jegh.k.191008.001
2. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol* 2020; **20**: 651–68. https://doi.org/10.1038/s41577-020-0306-5
3. Esfahani K, Roudaia L, Buhlaiga N. *et al.* A review of cancer immunotherapy: from the past, to the present, to the future. *Curr Oncol* 2020; **27**: S87–s97. https://doi.org/10.3747/co.27.5223
4. Laumont CM, Banville AC, Gilardi M. *et al.* Tumour-infiltrating B cells: immunological mechanisms, clinical impact and therapeutic opportunities. *Nat Rev Cancer* 2022; **22**: 414–30. https://doi.org/10.1038/s41568-022-00466-1
5. Gupta SL, Khan N, Basu S. *et al.* B-cell-based immunotherapy: a promising new alternative. *Vaccines (Basel)* 2022; **10**: 879. https://doi.org/10.3390/vaccines10060879
6. Oliveira G, Wu CJ. Dynamics and specificities of T cells in cancer immunotherapy. *Nat Rev Cancer* 2023; **23**: 295–316. https://doi.org/10.1038/s41568-023-00560-y
7. Xie N, Shen G, Gao W. *et al.* Neoantigens: promising targets for cancer therapy. *Signal Transduct Target Ther* 2023; **8**: 9. https://doi.org/10.1038/s41392-022-01270-x
8. Smith CC, Selitsky SR, Chai S. *et al.* Alternative tumour-specific antigens. *Nat Rev Cancer* 2019; **19**: 465–78. https://doi.org/10.1038/s41568-019-0162-4
9. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer* 2012; **12**: 252–64. https://doi.org/10.1038/nrc3239
10. Hollingsworth RE, Jansen K. Turning the corner on therapeutic cancer vaccines. *NPJ Vaccines* 2019; **4**: 7. https://doi.org/10.1038/s41541-019-0103-y
11. Koşaloğlu-Yalçın Z, Blazeska N, Vita R. *et al.* The Cancer Epitope Database and Analysis Resource (CEDAR). *Nucleic Acids Res* 2023; **51**: D845–d852. https://doi.org/10.1093/nar/gkac902
12. Vita R, Mahajan S, Overton JA. *et al.* The Immune Epitope DataBase (IEDB): 2018 update. *Nucleic Acids Res* 2019; **47**: D339–d343. https://doi.org/10.1093/nar/gky1006
13. Zhang G, Chitkushev L, Olsen LR. *et al.* TANTIGEN 2.0: a knowledge base of tumor T cell antigens and epitopes. *BMC Bioinformatics* 2021; **22**: 40. https://doi.org/10.1186/s12859-021-03962-7
14. Rammensee HG, Bachmann J, Emmerich NPN. *et al.* SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999; **50**: 213–9. https://doi.org/10.1007/s002510050595
15. Lata S, Bhasin M, Raghava GPS. MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. *BMC Res Notes* 2009; **2**: 61. https://doi.org/10.1186/1756-0500-2-61
16. Reche PA, Zhang H, Glutting J-P. *et al.* EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* 2005; **21**: 2140–1. https://doi.org/10.1093/bioinformatics/bti269
17. Calis JJ, Maybeno M, Greenbaum JA. *et al.* Properties of MHC Class I presented peptides that enhance immunogenicity. *PLoS Comput Biol* 2013; **9**: e1003266. https://doi.org/10.1371/journal.pcbi.1003266

18. Dhanda SK, Karosiene E, Edwards L. *et al.* Predicting HLA CD4 immunogenicity in human populations. *Front Immunol* 2018; **9**: 1369. https://doi.org/10.3389/fimmu.2018.01369
19. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982; **157**: 105–32. https://doi.org/10.1016/0022-2836(82)90515-0
20. Crooks GE, Hon G, Chandonia JM. *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004; **14**: 1188–90. https://doi.org/10.1101/gr.849004
21. Chowell D, Krishna S, Becker PD. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc Natl Acad Sci U S A* 2015; **112**: E1754–62. https://doi.org/10.1073/pnas.1500973112
22. Albert BA, Yang Y, Shao XM. *et al.* Deep neural networks predict Class I major histocompatibility complex epitope presentation and transfer learn neoepitope immunogenicity. *Nat Mach Intell* 2023; **5**: 861–72. https://doi.org/10.1038/s42256-023-00694-6
23. Wang G, Wan H, Jian X. *et al.* INeo-Epp: a novel T-cell HLA class-I immunogenicity or neoantigenic epitope prediction method based on sequence-related amino acid features. *Biomed Res Int* 2020; **2020**: 5798356–12. https://doi.org/10.1155/2020/5798356
24. Xu S, Wang X, Fei C. A highly effective system for predicting MHC-II epitopes with immunogenicity. *Front Oncol* 2022; **12**: 888556. https://doi.org/10.3389/fonc.2022.888556
25. Wang G, Wu T, Ning W. *et al.* TLimmuno2: predicting MHC Class II antigen immunogenicity through transfer learning. *Brief Bioinform* 2023; **24**: bbad116. https://doi.org/10.1093/bib/bbad116