Q:1,2,3,4,5

HLA Class I Supertype Classification Based on Structural Similarity

Q:6,7,8,9

Yue Shen,* Jerry M. Parks,† and Jeremy C. Smith*,†,‡

Q:11

HLA class I proteins, a critical component in adaptive immunity, bind and present intracellular Ags to CD8+ T cells. The extreme polymorphism of HLA genes and associated peptide binding specificities leads to challenges in various endeavors, including neoantigen vaccine development, disease association studies, and HLA typing. Supertype classification, defined by clustering functionally similar HLA alleles, has proven helpful in reducing the complexity of distinguishing alleles. However, determining supertypes via experiments is impractical, and current in silico classification methods exhibit limitations in stability and functional relevance. In this study, by incorporating three-dimensional structures we present a method for classifying HLA class I molecules with improved breadth, accuracy, stability, and flexibility. Critical for these advances is our finding that structural similarity highly correlates with peptide binding specificity. The new classification should be broadly useful in peptide-based vaccine development and HLA–disease association studies. *The Journal of Immunology*, 2023, 210: 1–12.

Human leukocyte Ag class I (HLA I) molecules are expressed on the surface of almost all nucleated cells and present intracellular antigenic peptides to CD8^{+} T lymphocytes (cytotoxic T cells), eliciting immune responses (1, 2). HLA I molecules are composed of two chains: the α-chain is polymorphic and functionally important, whereas the β-chain is the non-HLA–encoded β_2-microglobulin and is almost invariant. The HLA system is one of the most polymorphic regions within the human genome, which empowers the immune system to respond to a wide spectrum of epitopes (3). Three gene loci encode the classical HLA I α-chain: the HLA-A, HLA-B, and HLA-C genes. As of June 2022, the IPD-IMGT/HLA database includes 23,694 HLA I alleles, encoding 13,793 unique proteins. In addition, the diploid human genome creates a very large number of haplotypes, making the HLA system one of the most complex and diverse protein families.

The complexity arising from HLA polymorphism, in particular the largely unknown yet dissimilar functions, that is, peptide binding specificities, of alleles, is a challenging problem for researchers. For example, in peptide vaccine development, a key step is to find antigenic peptides that bind tightly to the specific HLA alleles carried by the patient. Because the frequencies of alleles and haplotypes vary greatly in populations, with many being very rare, most efforts were made focusing on populated alleles. However, only 150 alleles have experimentally characterized binding motifs and submotifs (4), and only 112 alleles have accurate allele-specific prediction models (5, 6). Still, uncharacterized alleles are carried by ~15% of the global population, as revealed by a simple calculation using the Immune Epitope Database population coverage tool (7), and some have been demonstrated to play a unique role in pathogenesis (8–11). In such cases, the rare and less common alleles need to be studied and cannot simply be neglected. Furthermore, the large number of alleles makes it difficult to identify associations between individual HLA phenotypes and disease susceptibility (12, 13).

Although HLA alleles do differ in peptide binding specificities, they are not always functionally distinct. Since the 1990s, studies have shown that some HLA alleles have largely overlapping peptide binding specificity (14–20). Accordingly, most HLA alleles could be clustered into supertypes and thus represented by a few typical alleles (21, 22), which greatly reduces the difficulty of discriminating between the huge number of HLA alleles.

Determining supertypes via experiments is very time- and effort-consuming, and thus it is impractical for classifying large numbers of alleles (23–25). As viable alternatives, several in silico classification methods have been proposed. Among these sequence-based approaches cluster alleles based on global (whole sequence) (26, 27) or local (binding groove or contact residues) (28–30) sequence similarity using sequence alignment. Prediction-based approaches calculate peptide binding specificity from predicted peptide–HLA affinities, instead of experimental data (31, 32). Structure-based methods use three-dimensional (3D) information derived from HLA structures, such as spatial similarity and molecular interaction fields (33), the number of hydrogen bonds, interface area, and the gap volume between HLA and peptide (34).

Current HLA I supertype classifications have proven to be helpful and are widely used but have limitations. Some methods cluster alleles based on sequence or structural similarities between the molecules. However, the correlation of sequence or structure with functional similarity is not well established, and thus the resulting supertypes are not guaranteed to include functionally similar alleles. Prediction-based methods are limited by the coverage and accuracy of peptide HLA affinity prediction methods. The accuracy of predictors relies heavily on training data, which are not abundant in general, leading to poor performance of widely used predictors on both rare and populated alleles (35, 36). Structure-based methods are potentially more informative than sequence-based methods but have been limited by the availability of high-quality HLA structures and the overall

*Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN; †Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; and ‡Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN

ORCIDs: 0000-0002-4148-6874 (Y.S.); 0000-0002-3103-9333 (J.M.P.).

Q:10

Received for publication September 13, 2022. Accepted for publication October 31, 2022.

The Compute and Data Environment for Science at the Oak Ridge National Laboratory was supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC05-00OR22725. This work used resources from The Compute and Data Environment for Science at the Oak Ridge National Laboratory.

Address correspondence and reprint requests to Dr. Jeremy C. Smith, Biosciences Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831. E-mail address: smithjc@ornl.gov

The online version of this article contains supplemental material.

Abbreviations used in this article: 3D, three-dimensional; HLA I, HLA class I; PD, peptide binding specificity distance; PDB, Protein Data Bank; RMSD, root mean square deviation; SC, silhouette coefficient; SD, structure distance; SSE, sum of squared error; TM, transmembrane.

www.jimmunol.org/cgi/doi/10.4049/jimmunol.2200685

complexity of structure analysis. In addition, due to the existence of interlocus interactions and coevolution of HLA genes, association analysis on the haplotype level has advantages over single-locus genotype approaches (37, 38), but most widely used supertype classifications are locus specific, and thus are not capable of describing functional relationships between alleles at different loci. These limitations indicate that advanced approaches to supertype classification are needed to further facilitate HLA studies.

In this study, we explore the use of 3D structural similarity to measure peptide binding specificity and cluster HLA alleles. By using the revolutionary structural modeling tool ColabFold (39) and an automated analysis pipeline, issues of structure-based methods in structure availability and analysis complexity were addressed, enabling the present method to be straightforwardly extended over the whole of the HLA I space and not be restricted to alleles with sufficient experimental data. Also, we establish that structural similarity between allele pairs is highly correlated with peptide binding specificity. Finally, based on structural similarity, 449 populated alleles are hierarchically clustered into 12 supertypes and 20 subtypes, giving flexibility in the resolution of describing epitope similarities between alleles. Compared to previous classifications, the present clustering method has better performance (cohesion), meaning that the classification better represents similarity in peptide binding specificity. Also, higher stability infers better confidence and extensibility, ensuring that users can add more custom alleles without perturbing the existing classification structure.

Materials and Methods

Structure modeling of HLA I molecules

Q:12 *3D structure of HLA I molecules.* HLA I protein molecules are heterodimers composed of two chains of different length: a polymorphic α-chain and a conserved β-chain (Fig. 1A). The α-chain contains the peptide binding domains (α1 and α2), the membrane-proximal domain (α3), the transmembrane (TM) region, and the cytoplasmic tail. Between the α1 and α2 domains, a binding groove is formed by two roughly parallel α helices on a β sheet surface (40). Residues of presented peptides occupy six characteristic binding pockets (named A–F) along the groove (Fig. 1B). Two deep pockets, B and F, contribute the most to the binding affinity and correspond to anchor residues on peptides (41).

HLA nomenclature. To maintain data integrity, HLA alleles are given unique names according to the nomenclature adopted by the WHO Nomenclature Committee for Factors of the HLA System (42). The allele name starts with the HLA prefix and gene name, followed by up to four sets of digits that identify the allotype group, specific protein, synonymous DNA variations within the coding region, and DNA variations in noncoding regions, respectively. For example, alleles HLA-A*01:01:01:01 and HLA-A*01:01:01:02 belong to the same gene locus and allotype group, encode the same proteins, and have the same DNA sequences in the coding regions, but they differ in the noncoding regions. For convenience, all alleles are referred to as unique proteins by the gene name and first two sets of digits, and the full names are used only when necessary.

HLA allele frequency analysis. Due to the large number of HLA alleles, protein structural modeling and supertype classification were limited to populated alleles. Allele frequency data were derived from the Allele Frequency Net Database (43). Alleles with a frequency >0.01 in any population with >50 samples were determined as populated alleles and selected for subsequent structural modeling and clustering. There were 128 HLA-A, 235 HLA-B, and 86 HLA-C alleles that met the above criterion.

Collecting HLA sequences and crystal structures. Protein sequences of populated HLA alleles were downloaded from the IPD-IMGT/HLA database (44). Crystal structures were downloaded from the IMGT/3Dstructure-DB (45). The allele name of each crystal structure was validated by extracting the HLA α-chain protein sequence and comparing it to the record in the IPD-IMGT/HLA database. There were 397 crystal structures of 40 alleles collected (Supplemental Table I). Structure cleaning was performed with PyMOL (version 2.5.2) (46), during which water molecules, ions, binding peptides, and the β-chain were removed, leaving the α-chain only.

Dataset split for validation purpose. Alleles were split into several datasets for different purposes. The *model quality evaluation set* was used to assess how closely the structural models reproduce experimentally determined crystal structures for the 40 alleles that have crystal structures available (Supplemental Table I). The *reference panel* was used to estimate the performance of present and previous methods, including 31 HLA-A, 57 HLA-B, and 22 HLA-C alleles that were classified into supertypes with high confidence in previous studies (Supplemental Table II). The HLA-A and HLA-B alleles were taken directly from the reference panel used in Sidney et al. (30), the supertype classifications of which were based on the experimentally established peptide binding motif. The HLA-C alleles were picked from the alleles of the consensus of the two supertype classification methods reported in Reche and Reinherz (33).

Structural modeling and coarse gaining. A total of 451 HLA structures, including 449 populated alleles plus 2 rare alleles that belong to the reference panel (B*08:02 and B*27:09), were modeled using ColabFold (39), an implementation of AlphaFold2 (47), that accelerates predictions by using fast homolog sequence searching with MMseqs2 (48, 49). The TM domains were trimmed and not modeled, as the TM domain is far from the peptide binding groove and is expected to have a negligible influence on the folding of the binding domain. Also, because the models were evaluated by overall quality, a poorly modeled TM domain may dominate scoring of model quality, concealing details in the peptide binding domain. Models were generated with the AlphaFold2_batch.ipynb (version 1.3) Google Colaboratory notebook. For each allele, five models were built with three rounds of recycles. The model with the highest predicted local distance difference test Cα score was then relaxed using the Rosetta FastRelax protocol (50) with fixed backbone to minimize steric clashes and optimize side chain positioning, because the constraint is not available in current ColabFold Amber implementation. We generated 20 relaxed replicas for each model and selected the one with the lowest total score calculated by the *ref_2015* Rosetta energy function (51). Finally, all models were trimmed to include only the peptide binding domain using Biopython (version 1.78) (52) and then superimposed onto the crystal structure of the most widely studied allele, HLA-A*02:01 (Protein Data Bank [PDB] ID: 1i4f) (53), with PyMOL by aligning the residues of the α helices and β sheets that form the binding groove in the α1 and α2 domains. The alignment is important for calculating structure distances (SDs) in the following step because the SD metric, which is defined below, is sensitive to the relative orientations of two structures.

Side-chain positioning is important in HLA–peptide interactions but is difficult to predict accurately (54). Coarse graining approximations reduce Q:13 the number of df and achieve balance between predictive power and computational cost, and thus have been successfully used in studying peptide–MHC interactions (55–57). Also, most protein chemical and physical similarities are analyzed at the residue level rather than the atomic level. Therefore, to simplify calculation and minimize errors introduced by modeling, the HLA models were coarse-grained using Python scripts. In this step, each residue was represented by the center of mass of its side chain (hydrogen atom for glycine), whereas backbone atoms (N, CA, C, and O in the PDB naming scheme) were not explicitly output.

Evaluating model quality. The quality of ColabFold models was measured by the root mean square deviation (RMSD) between models in the model quality evaluation set and crystal structures of the same allele. Some alleles have multiple crystal structures available. To represent the average of each of such alleles, the centroid structure was selected for comparison: first, the mean structure was generated in the way that the coordinates of any atom in the mean structure are the average coordinates of that atom in all crystal structures; next, the all-atom RMSDs between crystal structures and the mean structure were calculated using PyMOL; and finally, the structure with the lowest RMSD from the mean structure was selected as the centroid structure. The use of a centroid structure, rather than the mean structure itself, avoids comparisons with an unphysical structure obtained by averaging. To estimate natural structural variation, RMSDs between crystal structures and the centroid structure were also calculated.

Three kinds of RMSDs were used to describe structural similarity: all-atom, backbone (Cα), and coarse-grained. The all-atom and backbone RMSDs were calculated using PyMOL, and the coarse grained RMSD between two structures, *P1* and *P2*, was calculated with Eq. 1:

$$RMSD(P1,P2)=\sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}\delta_i^2}, \tag{1}$$

where δ_i is the distance between residue i of $P1$ and the equivalent Q:14 residue of $P2$, and N is the number of matching residues between $P1$ and $P2$.

Calculating SD

Definition of coarse-grained SD. In this study, we aim to cluster alleles with similar functions, that is, peptide binding specificity. However, that information is unavailable for most alleles due to the high demand of time and effort of peptide binding assays. As an alternative, because structure and function are closely related, the similarity in function might be able to be predicted from the similarity in structure. Specifically, the major assumption in this work is that alleles with similar peptide binding groove structures have similar peptide binding specificities. Structural similarity in this study is described by a distance metric, adapted from an atomic-detail distance metric defined and validated in a previous study (58) and modified in the present study for use with coarse-grained models. First, a kernel function is defined as

$$F(X) = \frac{1}{\cosh^k(\sigma \cdot X)}. \qquad (2)$$

Conforming to Hoffmann et al. (58), the kernel function in Eq. 2 transforms the distance X $(X \in R \,|\, X \geq 0)$ between two residues into a value in [0, 1], and closer residues have larger values, which means higher weight in determining structure similarity and vice versa. Parameters k and σ fine-tune the shape of the kernel function and determine the sensitivity to displacements between the two corresponding residues. For large σ and k values, distant residues will have small kernel function output, whereas small σ and k are more tolerant and distant residues will have higher values. To find optimal σ and k values, several shape/parameter combinations were tested (see *Comparing SD and PD*), and we found that $\sigma = 0.3, k = 1$ showed the highest correlation with peptide binding specificity distance (PD), indicated by the largest Pearson correlation coefficient (Supplemental Table III), and thus were applied in the current study.

The structural similarity, *K,* between two proteins, *P1* and *P2*, is defined as:

$$K(P1, P2) = \sum\nolimits_{i \in P1} \sum\nolimits_{j \in P2} \sqrt{w_i w_j} \cdot S_{ij} \cdot \frac{1}{\cosh^k(\sigma \cdot x_i - x_j)}. \qquad (3)$$

In this equation, x_i and x_j represent the Cartesian coordinates of coarse-grained residues i and j. The input of the kernel function is the Euclidean distance $x_i - x_j$. The residue similarity S_{ij} measures the physicochemical similarity between residues i and j, which is derived from $1 - G_{i,j}/215$, where $G_{i,j}$ is Grantham distance between the two amino acids (59) (Table I). The Grantham distance matrix is widely used in calculating sequence and structural similarities between HLA alleles (60–62). We tested the performance of several amino acid similarity matrices (63, 64), and with properly tuned shape parameters the transformed Grantham similarity matrix outperformed others (Supplemental Table III). The weight factor w_i controls the importance of residue i and is described in detail below.

The similarity, $K(P1, P2)$, is dependent on the number of residues that are defined to form the binding groove. Thus, the SD metric $SD(P1, P2)$, which measures the SD between P1 and P2 and fulfills the three axioms of metric space, that is, minimality, symmetry, and triangle inequality (65), is defined as:

$$SD(P1, P2) = \sqrt{K(P1, P1) + K(P2, P2) - 2K(P1, P2)}. \qquad (4)$$

Residue weight factor w. Some residues in the binding groove contribute more toward peptide binding affinity and selectivity than others; for example, the residues B and F in the binding pocket are more important than other residues. To represent realistic HLA-peptide binding, 21 key residues were selected and weighted according to a previous study (Table II) (66). In that work, single-residue substitutions of each position on HLA were performed, exhibiting changed peptide binding specificities compared with the wild type. Substitution of a position causing a larger difference suggests higher impact on binding specificity. The position that caused a minimum difference, residue 7, led to a 4.5% difference, so its weight was set to 1. The next lowest difference is in position 24, which resulted in a 5.2% difference and thus was assigned a weight of 5.2/4.5 ≈ 1.2. Weight factors of other residues were determined in the same way. Other residues have weight 0 and thus were ignored in calculating the SD.

Peptide binding specificity distance. The function of an allele as defined in the current study is represented by its peptide binding specificity. By this definition, the functional relationships between alleles are measured by peptide PDs, which were calculated using the method adapted from MHCcluster (31). First, the binding affinities of each allele to a set of 50,000 random peptides of length 8–14 (Table III) were calculated using the NetMHCpan 4.1 server (67). The random peptides were generated by Expasy RandSeq tool (68), and the ratio of different lengths was in accordance with natural peptide length preference (69). Next, the *PD*(*P1*,*P2*) between two alleles *P1* and *P2* was calculated by the correspondence of the top 10% strongest binders (including 50,000 × 10% = 5000 peptides) of each allele, calculated as:

$$PD(P1, P2) = 1 - \frac{n(b1 \cap b2)}{5000}, \qquad (5)$$

where *b1* and *b2* are the top 10% strongest binding peptides of alleles *P1* and *P2*, respectively, and $n(b1 \cap b2)$ is the number of peptides that are strong binders to both alleles. If both alleles have the same set of strong binders ($n(b1 \cap b2) = 5000$), the distance is 0, whereas for completely different sets of strong binders ($n(b1 \cap b2) = 0$) the distance is 1.

Comparing SD and PD. The locus-wise matrices of the *SD* and *PD* of the reference panel alleles were compared. For this, two pairwise distance matrices (SD and PD) for each of the three loci (HLA-A, HLA-B, and HLA-C) were calculated using Python scripts and were visualized as heatmaps by Matplotlib (version 3.4.3) (70) and seaborn (version 0.11.2) (71). The correlation between the two distances was further investigated via linear regression. Both SD and PD matrices of the same locus were normalized to the range in [0, 1] using scikit-learn (version 1.1.1) (72) The MinMaxScaler function and then the PD were linearly fitted on SD using the SciPy (version 1.8.0) (73) linregress function. The correlation between PD and SD was derived with the Pearson correlation coefficient, which ranges in [−1, 1], and a larger absolute value of this coefficient indicates stronger correlation.

Hierarchical clustering based on SD

SD clustering. In this study, we propose an SD clustering method that performs hierarchical clustering based on the pairwise SD. The method was implemented using the AgglomerativeClustering function in the scikit-learn package, with the complete-linkage method that has been shown to obtain a coherent and compact clustering result (4, 25, 33).

Table I. Residue similarity matrix adapted from Grantham distance matrix, derived as described in *Definition of coarse-grained SD*

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
0.49	0.33	0.66	0.73	0.54	0.42	0.74	0.34	0.28	0.33	0.48	0.59	0.68	0.79	0.44	0.70	0.63	0.37	0.18	Ser
1.00	0.53	0.52	0.67	0.48	0.55	0.42	0.55	0.55	0.64	0.16	0.87	0.80	0.60	0.88	0.55	0.75	0.58	0.53	Arg
	1.00	0.54	0.57	0.55	0.85	0.36	0.98	0.90	0.83	0.08	0.54	0.47	0.29	0.50	0.20	0.36	0.93	0.72	Leu
		1.00	0.82	0.87	0.68	0.80	0.56	0.47	0.49	0.21	0.64	0.65	0.58	0.52	0.50	0.57	0.60	0.32	Pro
			1.00	0.73	0.68	0.73	0.59	0.52	0.57	0.31	0.78	0.80	0.70	0.64	0.60	0.70	0.62	0.40	Thr
				1.00	0.70	0.72	0.56	0.47	0.48	0.09	0.60	0.58	0.48	0.51	0.41	0.50	0.61	0.31	Ala
					1.00	0.49	0.87	0.77	0.74	0.11	0.61	0.55	0.38	0.55	0.29	0.44	0.90	0.59	Val
						1.00	0.37	0.29	0.32	0.26	0.54	0.60	0.63	0.41	0.56	0.54	0.41	0.14	Gly
							1.00	0.90	0.85	0.08	0.56	0.49	0.31	0.53	0.22	0.38	0.95	0.72	Ile
								1.00	0.90	0.05	0.53	0.46	0.27	0.53	0.18	0.35	0.87	0.81	Phe
									1.00	0.10	0.61	0.54	0.33	0.60	0.26	0.43	0.83	0.83	Tyr
										1.00	0.19	0.28	0.35	0.06	0.28	0.21	0.09	0.00	Cys
											1.00	0.89	0.68	0.85	0.62	0.81	0.60	0.47	His
												1.00	0.79	0.75	0.72	0.87	0.53	0.40	Gln
													1.00	0.56	0.89	0.80	0.34	0.19	Asn
														1.00	0.53	0.74	0.56	0.49	Lys
															1.00	0.79	0.26	0.16	Asp
																1.00	0.41	0.29	Glu
																	1.00	0.69	Met

Table II. Contact residue positions in HLA class I proteins and corresponding weight factor

Position[a] (i)	Overlap of Binding Specificity[b] (%)	Difference in Binding Specificity[c] (%)	Weight (w_i)
63	55.4	44.6	9.9
67	65.7	34.3	7.6
116	73.9	26.1	5.8
9	75.6	24.4	5.4
97	78.7	21.3	4.7
152	79.4	20.6	4.6
167	82.8	17.2	3.8
156	83.2	16.8	3.7
74	83.9	16.1	3.6
70	85.6	14.4	3.2
80	86.3	13.7	3.0
171	84.0	13.0	2.9
45	87.3	12.7	2.8
66	88.0	12.0	2.7
77	88.0	12.0	2.7
76	89.4	10.6	2.4
114	89.7	10.3	2.3
99	90.4	9.6	2.1
163	93.1	6.9	1.5
95	93.1	6.9	1.5
59	93.5	6.5	1.4
158	93.8	6.2	1.4
69	94.5	5.5	1.2
24	94.8	5.2	1.2
7	95.5	4.5	1.0

Positions are listed in descending order of weight.

[a] There are no insertions or deletions in the binding domain of most existing HLA I alleles and all alleles considered in this study. Thus, residue positions are constant.

[b] Overlap of binding specificity between altered HLA with single point mutation and wild type.

[c] The difference was calculated as (1 − overlap)%.

Assessing clustering performance. Good clustering members within a cluster are highly similar whereas members of different clusters are highly different. The effectiveness of clustering was represented here by intracluster similarity (cohesion), measured by the sum of squared errors (SSEs) and silhouette coefficient (SC) using Python scripts with the subsequent procedure.

The calculation of the SSE requires identification of the cluster centers, which is usually the mean of cluster members. As a practical alternative, the centroids of clusters were used here instead, given the following equation:

$$SSE = \sum_{C} \sum_{i \in C} distance^2(i, centroid(C)), \tag{6}$$

in which the *distance* refers to the SD metric when calculating SD SSE, and the PD when calculating PD SSE. Given the same number of clusters, a smaller SSE value suggests better clustering because members of each cluster are more homogeneous. Usually, as the number of clusters increases, the SSE decreases monotonically. If all clusters are homogeneous, or the number of clusters equals the number of samples, the SSE reaches its minimal value of 0.

Table III. Random peptide set for measuring peptide binding specificity distance

Peptide Length	Natural Relative Abundance[a]	Recalculated Ratio (%)[b]	Number of Peptides	Random Sequence Length[c]
8	0.207	8.2	4,120	4,127
9	1	39.8	19,904	19,912
10	0.422	16.8	8,400	8,409
11	0.366	14.6	7,285	7,295
12	0.244	9.7	4,857	4,868
13	0.179	7.1	3,563	3,575
14	0.094	3.7	1,871	1,884
15	0.065	—	—	—

[a] Relative abundance compared with 9-mer as reported in Pierini and Lenz (60).

[b] The ratio was recalculated as NetMHCpan only accept residue length from 8 to 14.

[c] The input of NetMHCpan is a long protein sequence, and then the NetMHCpan server split the sequence into peptides of certain length using a sliding window algorithm.

The SC compares the intracluster variation to the distances to the neighboring clusters. For each allele i that belongs to cluster C, the average distance to all other alleles in the same cluster is

$$a(i) = \frac{1}{size(C) - 1} \sum_{i,j \in C,\ i/=j} distance(i,j). \tag{7}$$

If cluster C contains only one allele, then $a(i)$ is set to 0 to avoid a divide-by-zero error. Next, the average distance to the closest neighboring cluster for each allele i is defined as

$$b(i) = \min_{C/=D} \frac{1}{size(D)} \sum_{k \in D} distance(i,k), \tag{8}$$

where k is the member of neighboring cluster D. The SC for the clustering result regarding all alleles is calculated as

$$SC = \frac{1}{N} \sum_{i}^{N} \frac{b(i) - a(a)}{\max(a(i), b(i))}, \tag{9}$$

where N is number of alleles included in clustering. The range of the SC is [−1, 1], and a high SC indicates proper clustering.

Hierarchical clustering of the reference panel. Alleles in the reference panel were clustered locus-wise by the method described in *SD clustering*, as well as three other methods for comparison. A second method, "direct clustering," also uses complete linkage hierarchical clustering but is based on PD. Third, we reproduced the classifications from two previous studies: Sidney et al. (30) proposed six HLA-A and six HLA-B supertypes, and Doytchinova et al. (33) proposed three HLA-A, three HLA-B, and two HLA-C supertypes. Finally, the "random even split" method provides the baseline performance, randomly clustering alleles into any given number of clusters of equal size. The performance of each method was assessed by the PD SSE of its clustering result. Because the SSE is dependent on the number of clusters (N), methods including the structure clustering, direct clustering, and random even split were performed on a series of N clusters to show the trend.

Hierarchical clustering of 449 populated HLA alleles. All 449 populated HLA I alleles were hierarchically clustered across all alleles simultaneously for two purposes. First, possible interlocus functional overlapping is investigated. Also, such an approach ensures that supertypes are clustered with the same degree of functional similarity for all alleles rather than vary from locus to locus. The SD matrix of populated alleles was calculated, based on which clustering was performed as described in *SD clustering*. The optimal number of clusters was determined by analyzing the elbow plot and silhouette plot, which show the SSE and SC based on SD as a function of the number of clusters (N), respectively. With stepwise increasing of N, clustering was performed, and the SSE and SC were calculated for the resulting clusters. The optimal values of N are indicated by elbow points and SC peaks. The corresponding dendrogram was generated with the SciPy dendrogram function and visualized with the Interactive Tree of Life (iTOL) Web server (74).

Cluster stability estimated by bootstrapping. Clustering stability refers to the addition of new members without perturbing the clustering hierarchy. One major issue with hierarchical clustering is unsatisfying stability against independent resampling, which is a type of robustness measurement (75–77). The stability of the hierarchical clustering result was calculated in this study using a bootstrapping strategy. The HLA alleles were randomly sampled 100 times with replacement, and then the SD clustering was performed on the 100 bootstrapped samples, with the same setting as the original sample. The correspondence between bootstrapped cluster A and corresponding original cluster B was calculated using the Jaccard index (78), which is defined as

$$J(A,B) = \frac{A \cap B}{A \cup B}. \tag{10}$$

The stability of each cluster was then calculated as the mean Jaccard index of that cluster among 100 bootstrapped clustering results.

Associated content

Structures, scripts, and instructions for use are available at https://github.com/yshen25/HLA_clustering. The modeled all-atom structures of all 451 HLA I alleles are available in the GitHub repository.

Results

HLA I structural modeling

Structural models of 451 alleles were built using ColabFold, then relaxed using Rosetta FastRelax (Fig. 1). The models were of high confidence, as indicated by their average predicted local distance

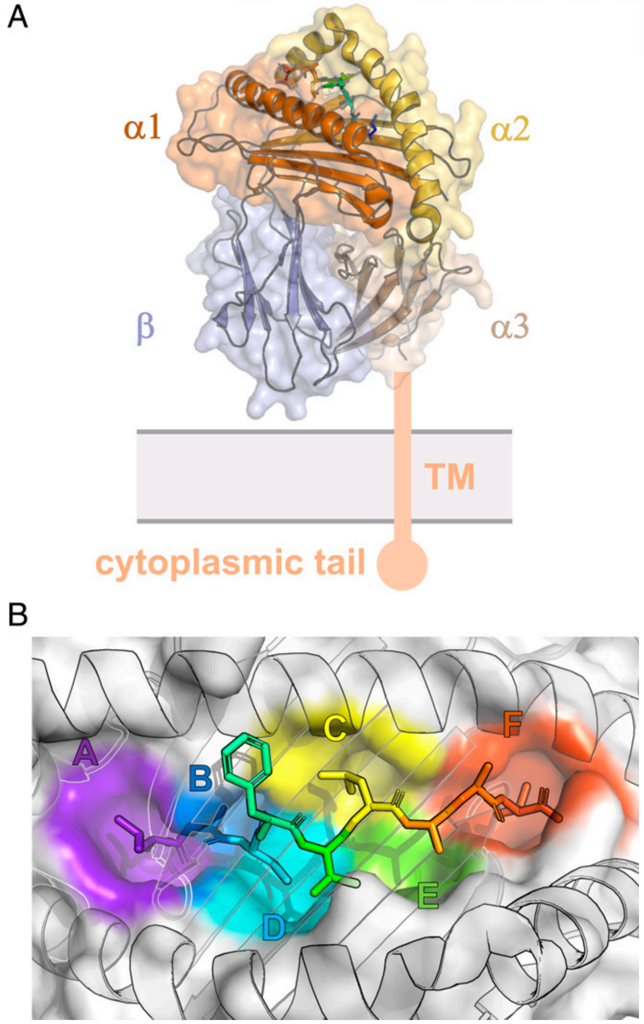

FIGURE 1. Structure of peptide–HLA complex (PDB ID: 3to2). (**A**) The whole complex demonstrating the domains and peptide binding groove. The transmembrane domain (TM), cytoplasmic tail, and lipid bilayer are shown schematically. (**B**) The top-down view showing six characteristic binding pockets (labeled A–F) along the binding groove. The peptide is shown in rainbow color with the N terminus in purple and the C terminus in orange. Pockets are colored the same with closest peptide residue.

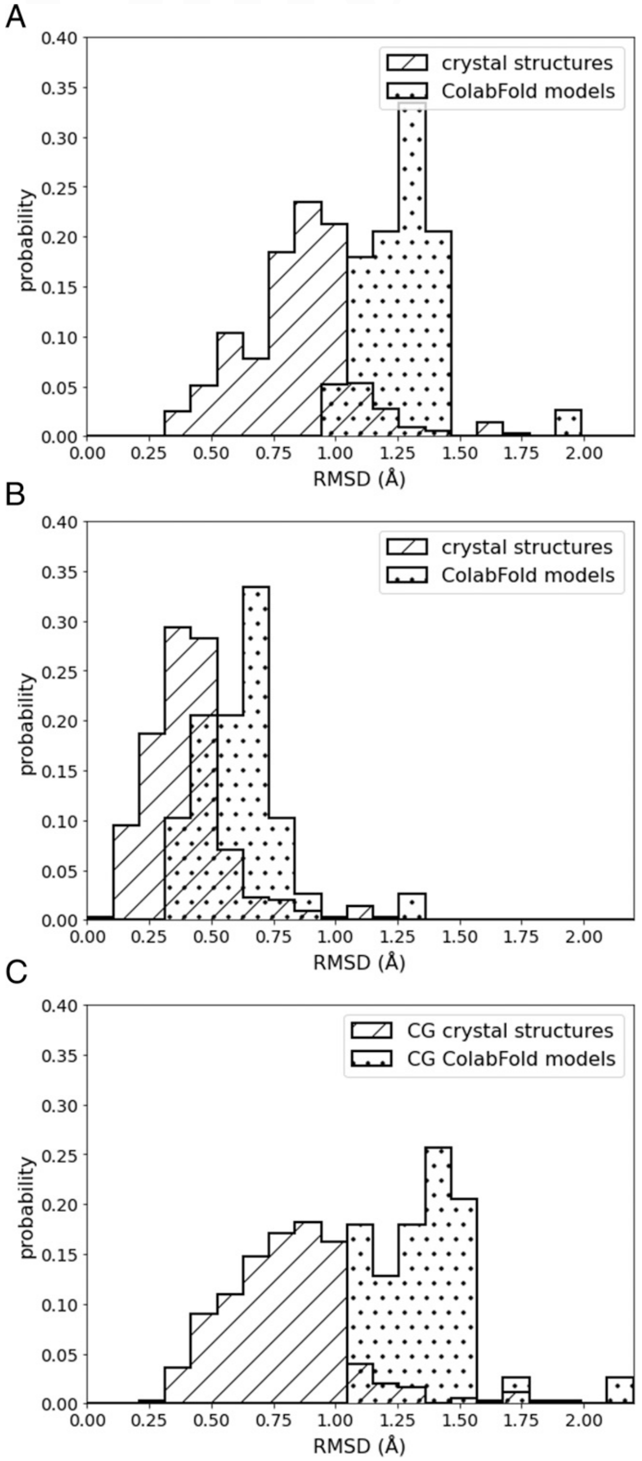

FIGURE 2. Distribution of all-atom, backbone, and coarse-grained RMSDs of crystal structures and ColabFold models compared with the centroid structure of multiple crystal structures (when available) for each allele. (**A–C**) All-atom (A), backbone (B), and coarse-grained (C) RMSDs.

difference test score of 96.2, and a minimum score of 94.0, with values >90 indicating high accuracy (47). The quality of the models was further tested by comparisons to the crystal structures of the same allele. The RMSDs, calculated with respect to the corresponding crystal structures or centroid structures, of the models in the model quality evaluation set were compared with the natural structural variations measured by RMSDs between crystal structures and centroid structures (Fig. 2). The all-atom RMSDs of the models (mean = 1.26 Å) are slightly larger than those of the crystal structures (mean = 0.85 Å). However, only one model (B*08:01, RMSD = 1.97 Å) exceeds the maximum RMSD of crystal structures (PDB ID: 4QRP, RMSD = 1.74 Å, compared with the centroid structure 4QRS) (Fig. 2A). The backbone RMSDs of the two groups are smaller than the all-atom RMSDs with the average 0.62 Å for models and 0.41 Å for the crystal structures (Fig. 2B), showing that the highest inaccuracy is in the side chain positioning. The distribution of coarse-grained RMSDs differs from the all-atom RMSDs (Fig. 2C), whereas the average (1.35 Å for models and 0.81 Å for crystal structures) is similar to the all-atom structures, showing that coarse graining has little influence on model quality. We also examined the model of B*08:01 that performed poorly in all three RMSD tests. In this study, the major differences between the model and crystal structure occur in the loop regions between the β strands, which are distant from the binding groove, and thus have only a minor impact on the clustering results, in line with the observation that the loop regions show flexibility in crystal structures as indicated, for example, by high B-factor values (40).

SD metric compared with PD

The SDs between alleles in the reference panel were compared with their PDs. Both distances were calculated locus-wise and then normalized to [0, 1], resulting in three pairs of intralocus distance matrices (Fig. 3). It is evident that the PD and SD heatmaps display similar patterns, suggesting that the two distances are in general agreement. This finding was further examined by linear regression between SD and PD. Among the three loci, HLA-B shows the highest coefficient ($R = 0.79$), followed by HLA-C ($R = 0.75$) and then HLA-A ($R = 0.73$), which confirms the strong correlation between the two distances. We also calculated the correlation coefficients between PD and the pseudo-sequence distance calculated by the BLOSUM62 substitution matrix (79), where the pseudo-sequence consists of the 34 contact residues as defined in NetMHCpan (67). The correlation coefficients for the BLOSUM62 substitution matrix are 0.69, 0.69, and 0.59 for HLA-A, HLA-B, and HLA-C, respectively, which are significantly inferior to the SDs (Supplemental Fig. 1). This finding supports the hypothesis that peptide binding specificity of the HLA I molecule is mainly determined by the structural landscape of the binding groove. Although given the dependence of function and binding in biology on 3D structure (80, 81), this result is not unexpected, and it is of particular significance in that a quantitative correlation between structure and peptide binding is found over a broad range of HLA types, emphasizing the usefulness of applying structural information in peptide/HLA affinity prediction over using sequence information alone.

Performance of reference panel clustering

The reference panel alleles were clustered locus-wise by four methods: 1) SD clustering, as described under *SD clustering*; 2) hierarchical clustering based on pairwise PD ("direct clustering"), which should provide the most accurate clustering among available binding data; 3) random even split representing the baseline performance; and 4) clustering results from two previous studies (30, 33). The performance, that is, the cluster cohesion, was calculated by the PD SSE, because the primary aim of supertype classification is to group alleles with similar peptide binding specificities (Fig. 4). A low PD SSE indicates higher intracluster similarity in peptide binding specificity, which is improved performance. In all three loci, the PD SSEs of the present clustering method are significantly lower than the random even split. Compared to the direct clustering curve, the SD clustering methods have very close SSE values, which indicates that the clustering based on SD is a reasonable approximation to clustering based on PD. The consistency between the two distances is also demonstrated by the SD

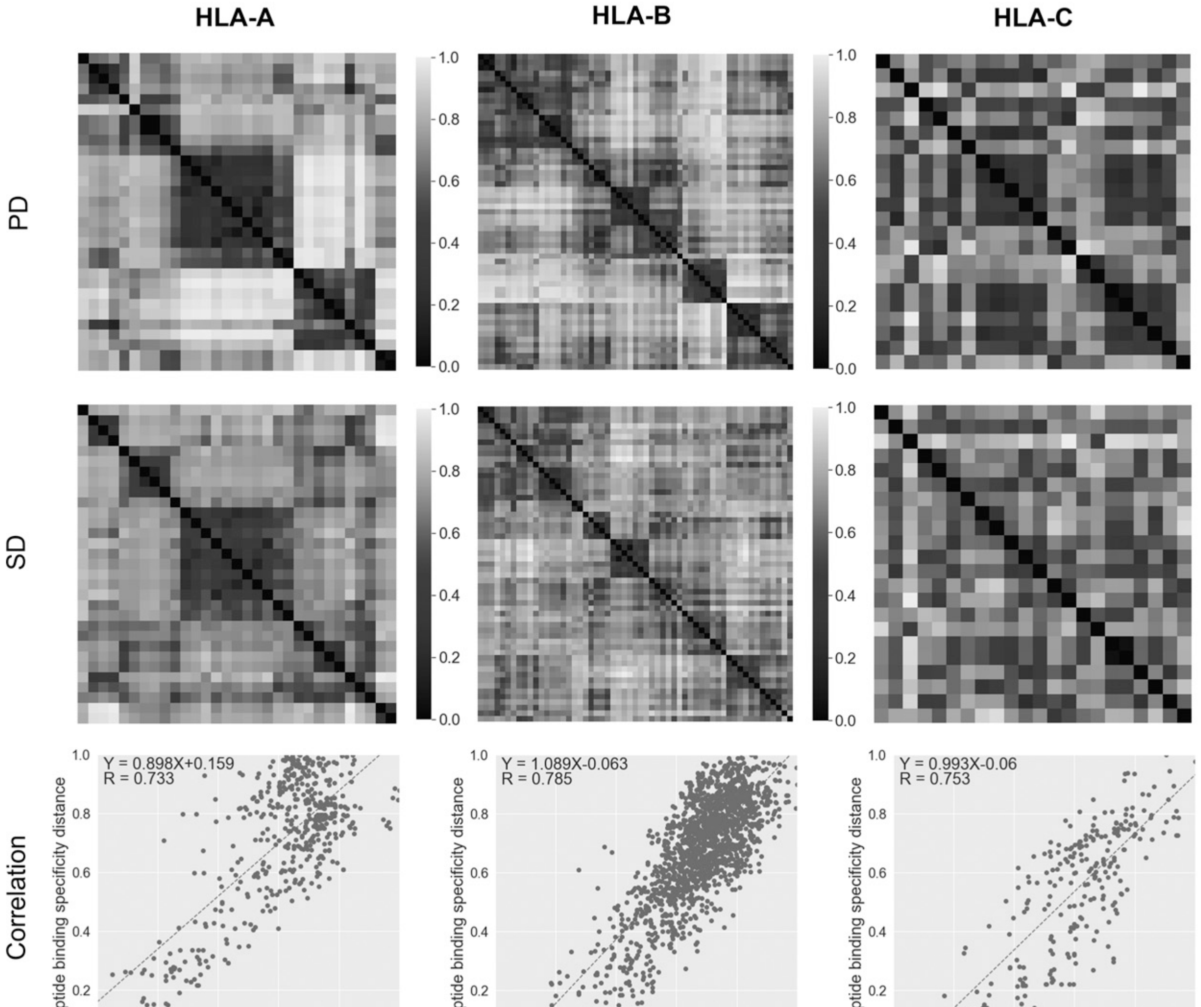

FIGURE 3. Comparison between peptide binding specificity distance (PD) predicted by NetMHCpan and structural distance (SD) defined in the present work. In the correlation plots, the fitted functions between two distances are shown with dotted line. Fitted function and Pearson correlation coefficient (R) are printed out.

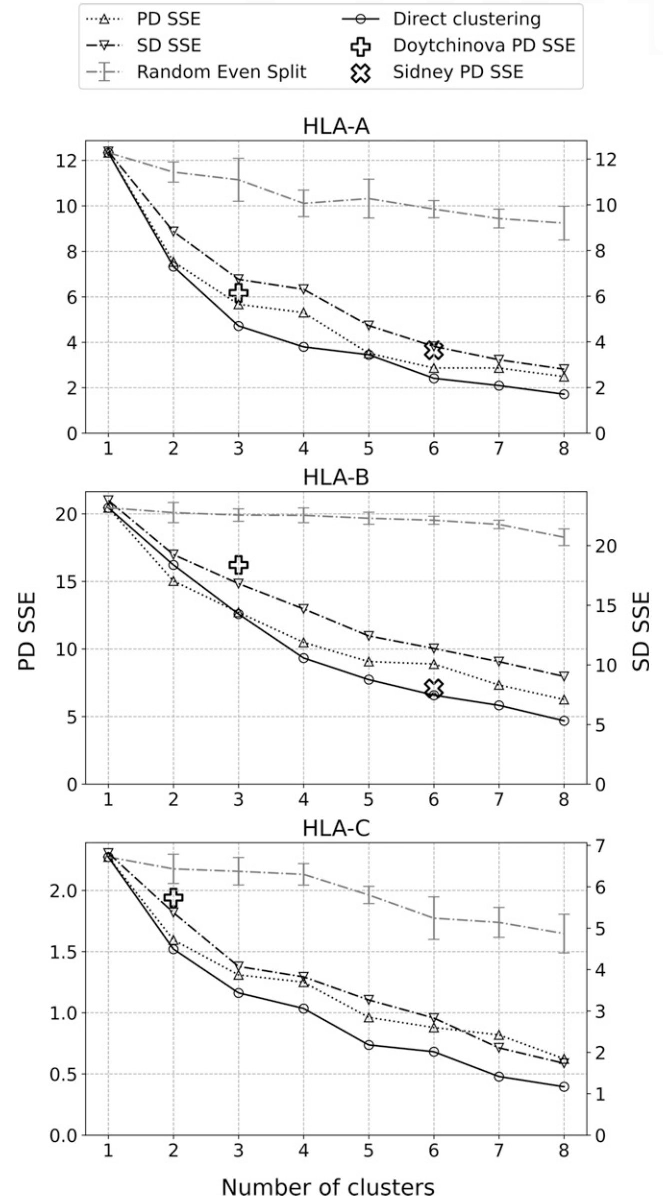

FIGURE 4. Elbow plot showing the sum of squared errors (SSEs) of HLA-A, HLA-B, and HLA-C reference panel alleles clustering. The *y*-axis on the left refers to SSE of the peptide binding specificity distance (PD SSE), and *y*-axis on the right refers to the SSE of the structural distance (SD SSE). The PD SSE of supertype classifications from Sidney et al. (25) and Doytchinova et al. (28) are shown as marked points.

SSE and the PD SSE curves, which show very similar shapes after scaling.

When compared with previous clustering methods, given the same number of clusters the present method outperforms findings in Doytchinova et al. (33) in all three loci, and it outperforms findings in Sidney et al. (30) in HLA-A but not in HLA-B. Importantly, note that the HLA-A and HLA-B alleles in the reference panel were derived directly from Sidney et al. (30), with experimentally validated peptide binding motifs. These results illustrate that the present classification method achieves accuracy comparable to experimentally determined supertypes, and better than previous structure-based methods. Furthermore, with the availability of high-quality models and the established correlation between SD and PD, it can be inferred that the present method should have sustainable clustering performance on alleles beyond the reference panel.

Supertype and subtype classification of populated HLA I alleles

Because the three classical HLA I loci, that is, A, B and C, are homologous, it is possible that alleles of different loci have overlapping peptide binding specificities. Some classification methods proposed mixed supertypes that include alleles from different loci (5, 29, 32). In the present study, we clustered 449 populated alleles (frequency >0.01) to investigate the possibility of overlapping peptide binding specificities and to ensure that the clustering of all HLA I alleles achieves the same level of functional similarity among all clusters.

The optimal number of clusters was determined by the elbow plot and silhouette plot method (Fig. 5). The most significant elbow point and SC peak appears at $n = 5$. However, dividing all HLA I alleles into only five supertypes may hide useful details and thus is not selected. Two optimized numbers of clusters, $n = 12$ and $n = 20$, are suggested by both the elbow points and SC peaks. Accordingly, the 449 populated HLA alleles were clustered into 12 supertypes and 20 subtypes, representing two levels of resolution (Fig. 6, Table IV). In this way, the supertypes describe the overall functional similarities, whereas the subtypes provide more details at enhanced resolution, which provides flexibility in application.

Subtypes are named after the most abundant allotype group in the cluster, following the naming convention, whereas supertypes were named after the included subtypes. The sizes of the supertypes and subtypes are imbalanced: each supertype contains one to three subtypes and 9–72 alleles, whereas each subtype contains 4–44 alleles, suggesting that distribution of the sampled 449 alleles is uneven in peptide binding specificity space.

Although the supertypes and subtypes were clustered based on structural similarity, they generally agree with allotype groups (the first set of digits of allele nomenclature) that are based on sequence similarity, because alleles that belong to the same group are usually clustered together. This agreement indicates that the allotype groups are reasonable approximations to supertypes and subtypes, as has been done previously (82–84). As for allotype groups that were separated in multiple clusters, single point mutations in the key residues and subsequent change of side chain positioning explain their split. For example, allele B*15:09 was clustered in subtype B14 rather than B15 where most of the B*15 allele group is situated. The full protein sequence comparison shows that B*15:09 has 8 mismatches with B*15:01, which belongs to B15, and 10 mismatches with B*14:01, which belongs to B14. However, among the 21 key residues defined in *Residue weight factor w*, B*15:09 has seven mismatches with B*15:01 but four mismatches with B*14:01. In addition, the

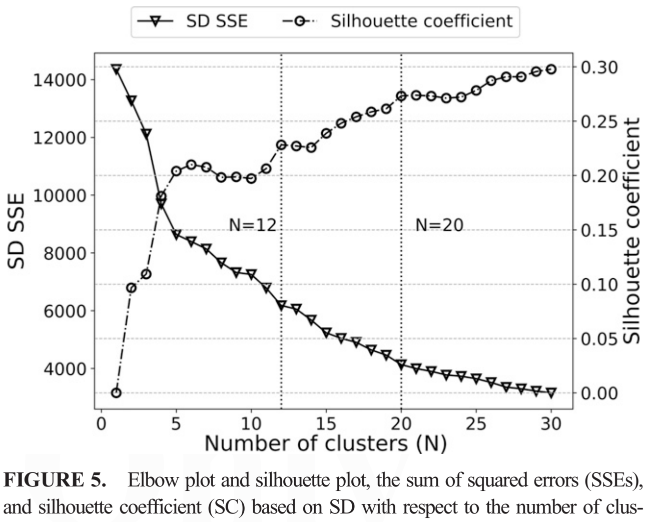

FIGURE 5. Elbow plot and silhouette plot, the sum of squared errors (SSEs), and silhouette coefficient (SC) based on SD with respect to the number of clusters. Two elbow points and silhouette peaks are shown as vertical dotted lines.

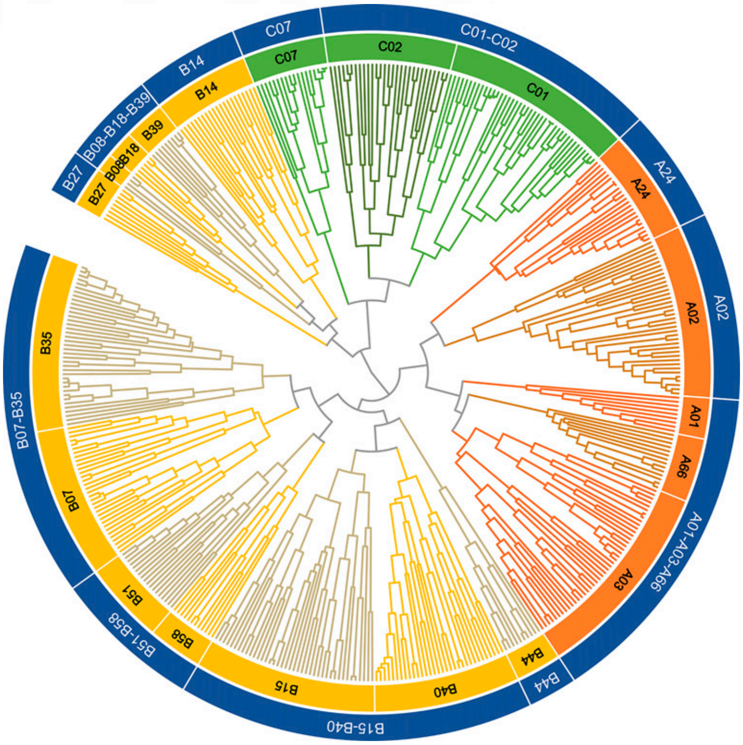

FIGURE 6. Dendrogram of 449 populated HLA alleles by hierarchical clustering. Clades of different subtypes are shown in different colors. Allele names were hidden for clarity. The detailed SD heatmap and dendrogram are available in the GitHub repository.

major structural difference between the three alleles is located near the F pocket: the orientation of residue 97 in B*15:01 is different from B*15:01 but similar to B*14:01 (Fig. 7).

At both the supertype and subtype levels, functional overlap between different loci is not significant, as most supertypes and subtypes are locus specific; the only exception is subtype C02 that includes two HLA-B alleles, B*46:01 and B*56:03 (Table IV). The allele B*46:01 is validated to have a close functional relationship with HLA-C (85), which provides circumstantial evidence that supports the present classification.

Stability of present supertype and subtype classification

The stability of supertypes and subtypes, that is, invariance, is an important concern, as more alleles will be included and clustered in future studies. To investigate whether the present supertype and subtype classification is reproducible given different sets of alleles, the stability was evaluated with a bootstrapping strategy. The stability of each supertype and subtype was calculated by its average Jaccard index among 100 repetitions (Table V). In previous studies, the stability of supertypes has rarely been reported, as we found only one study that reported the bootstrap values of 12 proposed HLA-A and HLA-B supertypes, the average of which is 0.54 in the range of [0, 1] (22). The average stability for our classification approach is 0.61 for supertypes and 0.75 for subtypes, respectively. In comparison, then, the present classification shows better stability than previous methods and is expected to be more robust. In addition, the stability represents the confidence of each cluster, as clusters of low stability may arise from random aggregation and be likely to collapse upon minor changes in the data. For supertypes of low stability, in particular B08-B18-B39, the structural and functional similarity may be falsely represented, and these supertypes should thus be used with caution. The subtypes show higher stability than supertypes, meaning that the functional similarity at the subtype level is more reliable than the supertype level. We also simulated a scenario in which new alleles are included: by adding the two already modeled rare alleles, B*08:02 and B*27:09, the pairwise distance matrix calculation and hierarchical clustering of the 451 alleles were performed the same way as for the 449 populated alleles. The two alleles, B*08:02 and B*27:09, were clustered into subtype B08 and B27, respectively. The subtypes remained the same, whereas supertypes showed a different pattern, as subtype B08 detached from B08-B18-B39 and combined with C07. This demonstrated that most subtypes are stable while some supertypes with low stability are less satisfying.

Discussion

HLA genes are extremely polymorphic, resulting in numerous alleles with diverse peptide binding specificities, thus allowing the human adaptive immune system to respond to a very wide range of Ags. However, this polymorphism also poses a significant difficulty in studies including, but not limited to, organ and cell transplantation, disease association studies, and peptide vaccine development. To reduce the conceptual complexity of the HLA system, the concept of supertypes was introduced to cluster alleles that have similar peptide binding specificities. Although, in the past, a number of experiments have been carried out for supertype classification, these cover only a small portion of binding peptide sequence space. For example, when considering only 8-mer, 9-mer, and 10-mer peptides consisting of the 20 canonical amino acids, the binding peptide sequence has $20^8 + 20^9 + 20^{10}$ ($\sim 1.1 \times 10^{13}$) possibilities, whereas HLA-A*02:01, one of the most studied alleles, has binding assay data for only 71,115 unique peptides in the Immune Epitope Database (86). Although a large fraction of those peptides will be poor binders or have no biological significance, there is still a massive gap, and thus supertypes can hardly be determined entirely by experiment.

An alternative to experimental methods is to use predicted peptide–HLA affinities, for example, using deep learning models trained by experimental affinities, and this is perhaps the most direct in silico approach. However, the coverage and accuracy of peptide–HLA

Table IV. Supertype and subtype assignment for 449 populated HLA I alleles

A01-A03-A66											
A01		A03							A66		
A*01:0	A*36:0	A*03:0	A*11:0	A*29:0	A*30:0	A*31:0	A*32:0	A*74:0	A*25:0	A*26:1	
A*01:0	A*80:0	A*03:0	A*11:0	A*29:1	A*30:0	A*31:0	A*32:1	A*74:0	A*26:0	A*26:1	
A*01:0		A*03:0	A*11:0	A*29:1	A*30:1	A*31:0	A*33:0		A*26:0	A*34:0	
A*01:0		A*03:2	A*11:1	A*29:2	A*30:1	A*31:1	A*33:0		A*26:0	A*43:0	
A*01:1		A*11:0	A*11:1	A*29:5	A*31:0	A*31:2	A*33:1		A*26:0	A*66:0	
A*01:2		A*11:0	A*29:0	A*30:0	A*31:0	A*32:0	A*34:0		A*26:1	A*66:0	
A*01:3		A*11:0	A*29:0	A*30:0	A*31:0	A*32:0	A*74:0		A*26:1	A*66:0	
A02											
A02											
A*02:0	A*02:0	A*02:0	A*02:1	A*02:1	A*02:2	A*02:4	A*02:5	A*68:0	A*68:3		
A*02:0	A*02:0	A*02:1	A*02:1	A*02:1	A*02:2	A*02:4	A*02:6	A*68:0	A*69:0		
A*02:0	A*02:0	A*02:1	A*02:1	A*02:2	A*02:3	A*02:4	A*68:0	A*68:0			
A*02:0	A*02:0	A*02:1	A*02:1	A*02:2	A*02:3	A*02:5	A*68:0	A*68:1			
A24											
A24											
A*23:0	A*23:1	A*24:0	A*24:0	A*24:0	A*24:1	A*24:1	A*24:1	A*24:2	A*24:2	A*24:2	A*24:5
A*23:0	A*24:0	A*24:0	A*24:0	A*24:0	A*24:1	A*24:1	A*24:2	A*24:2	A*24:2	A*24:4	
B07-B35											
B07					B35						
B*07:0	B*07:3	B*42:0	B*55:0	B*67:0	B*15:0	B*35:0	B*35:1	B*35:2	B*35:3	B*40:0	
B*07:0	B*07:7	B*42:1	B*55:1	B*78:0	B*15:1	B*35:0	B*35:1	B*35:2	B*35:4	B*53:0	
B*07:0	B*07:9	B*54:0	B*56:0	B*81:0	B*15:2	B*35:0	B*35:1	B*35:2	B*35:4	B*53:0	
B*07:0	B*07:9	B*54:1	B*56:0	B*82:0	B*18:0	B*35:0	B*35:1	B*35:2	B*35:6	B*53:0	
B*07:0	B*39:1	B*55:0	B*56:0	B*82:0	B*35:0	B*35:0	B*35:1	B*35:2	B*35:6		
B*07:0	B*39:2	B*55:0	B*56:0		B*35:0	B*35:1	B*35:1	B*35:2	B*35:7		
B*07:1	B*42:0	B*55:0	B*56:4		B*35:0	B*35:1	B*35:1	B*35:3	B*40:0		
B51-B58											
B51							B58				
B*51:0	B*51:0	B*51:0	B*51:1	B*51:1	B*51:7		B*15:1	B*57:0	B*57:0	B*58:0	
B*51:0	B*51:0	B*51:0	B*51:1	B*51:3	B*52:0		B*15:1	B*57:0	B*57:0	B*58:0	
B*51:0	B*51:0	B*51:1	B*51:1	B*51:6	B*59:0		B*15:6	B*57:0	B*57:2	B*58:0	
B08-B18-B39											
B08			B18				B39				
B*08:0	B*08:0		B*18:0	B*18:0	B*18:0		B*39:0	B*39:0	B*39:0	B*39:2	B*73:0
B*08:0	B*08:0		B*18:0	B*18:0	B*18:0		B*39:0	B*39:0	B*39:1	B*39:5	
B14											
B14											
B*14:0	B*14:0	B*14:0	B*14:1	B*15:1	B*15:2	B*15:3	B*38:0	B*38:0	B*39:0	B*78:0	
B*14:0	B*14:0	B*14:0	B*15:0	B*15:1	B*15:2	B*15:9	B*38:0	B*39:0	B*39:1		
B15-B40											
B15						B40					
B*13:0	B*15:0	B*15:2	B*15:3	B*40:0	B*47:0	B*15:3	B*40:0	B*40:1	B*41:0	B*49:0	
B*13:0	B*15:0	B*15:2	B*15:3	B*40:0	B*48:0	B*15:5	B*40:1	B*40:2	B*41:2	B*50:0	
B*13:0	B*15:0	B*15:2	B*15:3	B*40:0	B*48:0	B*37:0	B*40:1	B*40:3	B*44:0	B*50:0	
B*13:3	B*15:0	B*15:2	B*15:4	B*40:0	B*48:0	B*39:0	B*40:1	B*40:4	B*44:1	B*50:0	
B*15:0	B*15:1	B*15:3	B*18:0	B*40:2	B*48:0	B*39:0	B*40:1	B*40:4	B*44:1		
B*15:0	B*15:1	B*15:3	B*35:2	B*40:3		B*40:0	B*40:1	B*41:0	B*45:0		
B*15:0	B*15:1	B*15:3	B*35:2	B*47:0		B*40:0	B*40:1	B*41:0	B*48:0		
B27											
B27											
B*27:0	B*27:0	B*27:0	B*27:0	B*27:0	B*27:0	B*27:0	B*27:0	B*27:1			
B44											
B44											
B*44:0	B*44:0	B*44:0	B*44:0	B*44:0	B*44:0	B*44:0	B*44:1	B*44:2	B*44:2		
C01-C02											
C01							C02				
C*01:0	C*03:0	C*03:1	C*04:1	C*08:0	C*15:0	C*15:1	B*46:0	C*02:1	C*06:0	C*12:0	C*15:0
C*01:0	C*03:0	C*04:0	C*04:1	C*08:0	C*15:0	C*17:0	B*56:0	C*03:0	C*06:0	C*12:0	C*15:0
C*01:0	C*03:0	C*04:0	C*04:4	C*08:0	C*15:0	C*17:0	C*01:0	C*03:1	C*07:2	C*14:0	C*16:0
C*01:4	C*03:0	C*04:0	C*05:0	C*08:0	C*15:0	C*18:0	C*02:0	C*03:1	C*12:0	C*14:0	C*16:0
C*01:5	C*03:0	C*04:0	C*05:0	C*08:1	C*15:0	C*18:0	C*02:0	C*06:0	C*12:0	C*14:0	C*16:0
C*03:0	C*03:1	C*04:0	C*08:0	C*08:7	C*15:1		C*02:0	C*06:0	C*12:0	C*14:0	
C07											
C07											
C*07:0	C*07:0	C*07:0	C*07:0	C*07:1	C*07:1	C*07:1	C*07:2	C*07:2			
C*07:0	C*07:0	C*07:0	C*07:0	C*07:1	C*07:1	C*07:1	C*07:2	C*07:3			

affinity prediction methods are limited, with the leading issue being the unsatisfying availability of training data, as mentioned above. Another issue comes from the peptide binding assays themselves. The most widely used measurement of peptide−HLA affinity is the IC_{50}, which is defined as the concentration of a query peptide that blocks 50% of standard peptide binding, but different standard

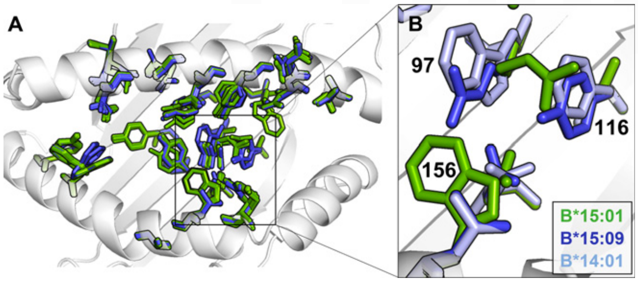

FIGURE 7. Comparison of structures of alleles in allotype group HLA-B*15 that were clustered in subtype B14 and B15. (**A**) Top-down view of the peptide binding groove. Alleles that clustered in subtype B14 are colored in blue; B15 alleles are in green. (**B**) Zoom-in near the F pocket, showing the difference of B*15:09, B*15:01, and B*14:01 in residue 97, 116, and 156. Compared to B*15:09 and B*14:01, B*15:01 has Trp in position 156, which is bulk in size and pushed Arg^{97} toward Ser^{116}, resulting in the difference in the shape of the F pocket. Hydrogen atoms and residues from other alleles in (B) were omitted for clarity.

peptides have been used for different HLA alleles (87). As a result, the binding assay data of different alleles are in different bases and thus not, in principle, comparable. However, they have been combined in the training sets of pan-specific predictors, which may lead to inaccuracy.

Another alternative to time-consuming experiments is to cluster HLA alleles based on sequence or structural similarity, which recognizes that the peptide binding specificity is determined by the spatial and chemical properties of the peptide binding groove. Sequence/structure-based methods have advantages compared with affinity prediction-based methods. First, the sequences and structures of HLAs are available with less effort than binding assays; second, the methods may be applicable to all HLA alleles and not just those with available binding affinity data, permitting reliable clustering performance on understudied alleles.

Both sequence-based and structure-based approaches have been successfully applied in general protein binding pocket comparison (88–93). In the present study, we have demonstrated that both the sequence and SDs we use for HLA I are highly correlated with PDs. Furthermore, as structure-based methods should, in principle, incorporate more direct functional detail than simple sequences, in this study, indeed our structure-based method outperforms sequence-based methods in that SD has higher correlation coefficient with PD than sequence distance.

One limitation of structure-based methods, including the SD clustering, is the imperfect quality of HLA crystal structures and predicted models (91). As illustrated above, natural structural variations occur in crystal structures of the same HLA allele, which can be caused by peptide-induced conformational change (94–96), chaperones (e.g., tapasin) (97, 98), or crystal packing (99). To minimize the impact of imperfect HLA models, relaxation and coarse graining were applied in the present study. Coarse graining decreases the number of df, thus reducing the impact of possible inaccuracies in side chain orientations and improves its robustness. However, coarse graining may also reduce clustering accuracy through the loss of information that would be present in a fully atomistic structure. Clearly, the development of all-atom methods may further improve the present approach. Furthermore, static HLA structures cannot reveal certain factors that may influence peptide binding specificity, such as the conformational flexibility of a peptide–HLA complex (98, 100–103). Also note that the difference in peptide binding specificity cannot fully explain certain functional differences. For example, B*27:05 and B*27:09 have very similar sequences (single mutation) and structures (104). Thus, both alleles are included in subtype B27 in the current study, and this is validated by the experimental result that they share an ~80% repertoire (105). However, B*27:05 is strongly associated with ankylosing spondylitis whereas B*27:09 is not (106), and this cannot be explained by the present method but, rather, requires dynamical considerations (103).

Table V. Supertype and subtype stability measured by average Jaccard index in bootstrapping

Supertype	Average Jaccard Index	Subtype	Average Jaccard Index
A01-A03-A66	0.65	A01	0.72
		A03	0.70
		A66	0.68
A02	0.58	A02	0.83
A24	0.66	A24	0.89
B07-B35	0.60	B07	0.70
		B35	0.70
B51-B58	0.74	B51	0.91
		B58	0.94
B08-B18-B39	0.25	B08	0.66
		B18	0.66
		B39	0.54
B14	0.68	B14	0.74
B15-B40	0.62	B15	0.54
		B40	0.65
B27	0.68	B27	0.96
B44	0.38	B44	0.64
C01-C02	0.87	C01	0.87
		C02	0.76
C07	0.66	C07	0.93

The reasons for the performance differences between supertype classification methods compared in the current study are complex. Apart from the limited data quality available to previous studies, there are differences in the definition of supertypes. All methods agree that supertypes include functionally similar alleles, although they focus on different aspects, including the selectivity of anchor residues of binding peptides (peptide motif) (21, 22, 30), the interaction profile of the binding groove (25, 33, 107), and, as adopted in the current study, peptide binding specificity (31). Therefore, these supertype classifications should not be used interchangeably.

In conclusion, we have quantified the correlation between structural similarity of HLA binding groove and peptide binding specificity using a newly defined SD metric, based on which we propose a new HLA classification scheme. Our results show that binding specificity is mainly determined by the structural landscape of the binding groove. Although this is not unexpected, the present results demonstrate it at scale and thus elevate the importance of structural considerations.

Relative to previous classifications, our approach achieved advances in four aspects. First, the present supertypes and subtypes better represent similarity in peptide binding specificity, as illustrated by the improved clustering performance (cohesion). Also, the flexibility of classification is improved, in that clustering at different resolutions, supertype and subtype, is incorporated. Third, the method is demonstrated to have better stability, meaning that the classification was performed with higher confidence, and users can add user-defined alleles without perturbing the existing classification structure. Finally, the method is broadly applicable across HLA alleles.

Improvements could be expected when applying the clustering result in developing supertype/subtype-specific affinity prediction tools and supertype/subtype–disease association studies. With advances in the understanding of peptide–HLA interactions and the further development of structural modeling approaches, structure-based methods are destined to improve further.

Disclosures

The authors have no financial conflicts of interest.

References

1. Klein, J., and A. Sato. 2000. The HLA system. *N. Engl. J. Med.* 343: 702–709.
2. Hewitt, E. W. 2003. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* 110: 163–169.
3. Bird, L. 2004. Advantages to being different. *Nat. Rev. Immunol.* 4: 577.
4. Lee, K. H., Y. C. Chang, T. F. Chen, H. F. Juan, H. K. Tsai, and C. Y. Chen. 2021. Connecting MHC-I-binding motifs with HLA alleles via deep learning. *Commun. Biol.* 4: 1194.
5. Di Marco, M., H. Schuster, L. Backert, M. Ghosh, H. G. Rammensee, and S. Stevanović. 2017. Unveiling the peptide motifs of HLA-C and HLA-G from naturally presented peptides and generation of binding prediction matrices. *J. Immunol.* 199: 2639–2651.
6. O'Donnell, T. J., A. Rubinsteyn, M. Bonsack, A. B. Riemer, U. Laserson, and J. Hammerbacher. 2018. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7: 129–132.e4.
7. Bui, H. H., J. Sidney, K. Dinh, S. Southwood, M. J. Newman, and A. Sette. 2006. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics* 7: 153.
8. Tian, W., F. Zhu, J. Cai, L. Li, H. Jin, and W. Wang. 2020. Multiple low-frequency and rare HLA-B allelic variants are associated with reduced risk in 1,105 nasopharyngeal carcinoma patients in Hunan province, southern China. *Int. J. Cancer* 147: 1397–1404.
9. Trachtenberg, E., B. Korber, C. Sollars, T. B. Kepler, P. T. Hraber, E. Hayes, R. Funkhouser, M. Fugate, J. Theiler, Y. S. Hsu, et al. 2003. Advantage of rare HLA supertype in HIV disease progression. *Nat. Med.* 9: 928–935.
10. Das Ghosh, D., I. Mukhopadhyay, A. Bhattacharya, R. Roy Chowdhury, N. R. Mandal, S. Roy, and S. Sengupta. 2017. Impact of genetic variations and transcriptional alterations of HLA class I genes on cervical cancer pathogenesis. *Int. J. Cancer* 140: 2498–2508.
11. Wennink, R. A. W., J. H. de Boer, S. Hiddingh, A. J. W. Haasnoot, V. Kalinina Ayuso, T. de Hoop, J. van Setten, E. Spierings, and J. J. W. Kuiper. 2021. Next-generation HLA sequence analysis uncovers shared risk alleles between clinically distinct forms of childhood uveitis. *Invest. Ophthalmol. Vis. Sci.* 62: 19.
12. Jin, P., and E. Wang. 2003. Polymorphism in clinical immunology—from HLA typing to immunogenetic profiling. *J. Transl. Med.* 1: 8.
13. Kishore, A., and M. Petrek. 2018. Next-generation sequencing based HLA typing: deciphering immunogenetic aspects of sarcoidosis. *Front. Genet.* 9: 503.
14. Sidney, J., M.-F. del Guercio, S. Southwood, V. H. Engelhard, E. Appella, H.-G. Rammensee, K. Falk, O. Rötzschke, M. Takiguchi, R. T. Kubo, et al. 1995. Several HLA alleles share overlapping peptide specificities. *J. Immunol.* 154: 247–259.
15. del Guercio, M.-F., J. Sidney, G. Hermanson, C. Perez, H. M. Grey, R. T. Kubo, and A. Sette. 1995. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. *J. Immunol.* 154: 685–693.
16. Fruci, D., P. Rovero, G. Falasca, A. Chersi, R. Sorrentino, R. Butler, N. Tanigaki, and R. Tosi. 1993. Anchor residue motifs of HLA class-I-binding peptides analyzed by the direct binding of synthetic peptides to HLA class I α chains. *Hum. Immunol.* 38: 187–192.
17. Sidney, J., M.-F. del Guercio, S. Southwood, G. Hermanson, A. Maewal, E. Appella, and A. Sette. 1997. The HLA-A*0207 peptide binding repertoire is limited to a subset of the A*0201 repertoire. *Hum. Immunol.* 58: 12–20.
18. Sidney, J., H. M. Grey, S. Southwood, E. Celis, P. A. Wentworth, M.-F. del Guercio, R. T. Kubo, R. W. Chesnut, and A. Sette. 1996. Definition of an HLA-A3-like supermotif demonstrates the overlapping peptide-binding repertoires of common HLA molecules. *Hum. Immunol.* 45: 79–93.
19. Sidney, J., S. Southwood, M.-F. del Guercio, H. M. Grey, R. W. Chesnut, R. T. Kubo, and A. Sette. 1996. Specificity and degeneracy in peptide binding to HLA-B7-like class I molecules. *J. Immunol.* 157: 3480–3490.
20. Sidney, J., H. M. Grey, R. T. Kubo, and A. Sette. 1996. Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. *Immunol. Today* 17: 261–266.
21. Sette, A., and J. Sidney. 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 50: 201–212.
22. Lund, O., M. Nielsen, C. Kesmir, A. G. Petersen, C. Lundegaard, P. Worning, C. Sylvester-Hvid, K. Lamberth, G. Røder, S. Justesen, et al. 2004. Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 55: 797–810.
23. Kobayashi, H., J. Lu, and E. Celis. 2001. Identification of helper T-cell epitopes that encompass or lie proximal to cytotoxic T-cell epitopes in the gp100 melanoma tumor antigen. *Cancer Res.* 61: 7577–7584.
24. Panigada, M., T. Sturniolo, G. Besozzi, M. G. Boccieri, F. Sinigaglia, G. G. Grassi, and F. Grassi. 2002. Identification of a promiscuous T-cell epitope in *Mycobacterium tuberculosis* Mce proteins. *Infect. Immun.* 70: 79–85.
25. Doytchinova, I. A., and D. R. Flower. 2005. In silico identification of supertypes for class II MHCs. *J. Immunol.* 174: 7085–7095.
26. Cano, P., B. Fan, and S. Stass. 1998. A geometric study of the amino acid sequence of class I HLA molecules. *Immunogenetics* 48: 324–334.
27. McKenzie, L. M., J. Pecon-Slattery, M. Carrington, and S. J. O'Brien. 1999. Taxonomic hierarchy of HLA class I allele sequences. *Genes Immun.* 1: 120–129.
28. Chelvanayagam, G. 1996. A roadmap for HLA-A, HLA-B, and HLA-C peptide binding specificities. *Immunogenetics* 45: 15–26.
29. Zhang, C., A. Anderson, and C. DeLisi. 1998. Structural principles that govern the peptide-binding motifs of class I MHC molecules. *J. Mol. Biol.* 281: 929–947.
30. Sidney, J., B. Peters, N. Frahm, C. Brander, and A. Sette. 2008. HLA class I supertypes: a revised and updated classification. *BMC Immunol.* 9: 1–15.
31. Thomsen, M., C. Lundegaard, S. Buus, O. Lund, and M. Nielsen. 2013. MHCcluster, a method for functional clustering of MHC molecules. *Immunogenetics* 65: 655–665.
32. Reche, P. A., and E. L. Reinherz. 2007. Definition of MHC supertypes through clustering of MHC peptide-binding repertoires. *Methods Mol. Biol.* 409: 163–173.
33. Doytchinova, I. A., P. Guan, and D. R. Flower. 2004. Identifiying human MHC supertypes using bioinformatic methods. *J. Immunol.* 172: 4314–4323.
34. Tong, J. C., T. W. Tan, and S. Ranganathan. 2007. In silico grouping of peptide/HLA class I complexes using structural interaction characteristics. *Bioinformatics* 23: 177–183.
35. Shao, X. M., R. Bhattacharya, J. Huang, I. K. A. Sivakumar, C. Tokheim, L. Zheng, D. Hirsch, B. Kaminow, A. Omdahl, M. Bonsack, et al. 2020. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunol. Res.* 8: 396–408.
36. Bonsack, M., S. Hoppe, J. Winter, D. Tichy, C. Zeller, M. D. Küpper, E. C. Schitter, R. Blatnik, and A. B. Riemer. 2019. Performance evaluation of MHC class-I binding prediction tools based on an experimentally validated MHC-peptide binding data set. *Cancer Immunol. Res.* 7: 719–736.
37. N'Diaye, A., J. K. Haile, A. T. Cory, F. R. Clarke, J. M. Clarke, R. E. Knox, and C. J. Pozniak. 2017. Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. [Published erratum appears in 2017 *PLoS One* 12: e0187178.] *PLoS One* 12: e0170941. **Q:19**
38. Abed, A., and F. Belzile. 2019. Comparing single-SNP, multi-SNP, and haplotype-based approaches in association studies for major traits in barley. *Plant Genome* 12: 190036.
39. Mirdita, M., K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger. 2022. ColabFold: making protein folding accessible to all. *Nat. Methods* 19: 679–682.
40. Wieczorek, M., E. T. Abualrous, J. Sticht, M. Álvaro-Benito, S. Stolzenberg, F. Noé, and C. Freund. 2017. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. *Front. Immunol.* 8: 292.
41. Nguyen, A. T., C. Szeto, and S. Gras. 2021. The pockets guide to HLA class I molecules. *Biochem. Soc. Trans.* 49: 2319–2331.
42. Marsh, S. G., E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, M. Fernández-Viña, D. E. Geraghty, R. Holdsworth, C. K. Hurley, et al. 2010. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* 75: 291–455.
43. Gonzalez-Galarza, F. F., A. McCabe, E. J. M. D. Santos, J. Jones, L. Takeshita, N. D. Ortega-Rivera, G. M. D. Cid-Pavon, K. Ramsbottom, G. Ghattaoraya, A. Alfirevic, et al. 2020. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* 48(D1): D783–D788.
44. Robinson, J., J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, and S. G. Marsh. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 43(D1): D423–D431.
45. Ehrenmann, F., Q. Kaas, and M. P. Lefranc. 2010. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res.* 38(Suppl. 1): D301–D307.
46. Schrödinger, LLC. 2015. The PyMOL molecular graphics system, version 2.5. Available at: https://pymol.org/2/.
47. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.
48. Steinegger, M., and J. Söding. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35: 1026–1028.
49. Mirdita, M., M. Steinegger, and J. Söding. 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35: 2856–2858.
50. Conway, P., M. D. Tyka, F. DiMaio, D. E. Konerding, and D. Baker. 2014. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.* 23: 47–55.
51. Alford, R. F., A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, et al. 2017. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 13: 3031–3048.
52. Cock, P. J., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. de Hoon. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423.
53. Hillig, R. C., P. G. Coulie, V. Stroobant, W. Saenger, A. Ziegler, and M. Hülsmeyer. 2001. High-resolution structure of HLA-A*0201 in complex with a tumour-specific antigenic peptide encoded by the MAGE-A4 gene. *J. Mol. Biol.* 310: 1167–1176.
54. Kingsford, C. L., B. Chazelle, and M. Singh. 2005. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* 21: 1028–1036.
55. Tian, F., R. Tan, T. Guo, P. Zhou, and L. Yang. 2013. Fast and reliable prediction of domain-peptide binding affinity using coarse-grained structure models. *Biosystems* 113: 40–49.
56. Knapp, B., S. Demharter, C. M. Deane, and P. Minary. 2016. Exploring peptide/MHC detachment processes using hierarchical natural move Monte Carlo. *Bioinformatics* 32: 181–186.
57. Huang, M., W. Huang, F. Wen, and R. G. Larson. 2017. Efficient estimation of binding free energies between peptides and an MHC class II molecule using coarse-grained molecular dynamics simulations with a weighted histogram analysis method. *J. Comput. Chem.* 38: 2007–2019.
58. Hoffmann, B., M. Zaslavskiy, J.-P. Vert, and V. Stoven. 2010. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics* 11: 99.
59. Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185: 862–864.

60. Pierini, F., and T. L. Lenz. 2018. Divergent allele advantage at human MHC genes: signatures of past and ongoing selection. *Mol. Biol. Evol.* 35: 2145–2158.
61. Arora, J., F. Pierini, P. J. McLaren, M. Carrington, J. Fellay, and T. L. Lenz. 2020. HLA heterozygote advantage against HIV-1 is driven by quantitative and qualitative differences in HLA allele-specific peptide presentation. *Mol. Biol. Evol.* 37: 639–650.
62. Schetelig, J., F. Heidenreich, H. Baldauf, S. Trost, B. Falk, C. Hoßbach, R. Real, A. Roers, D. Lindemann, A. Dalpke, et al. 2021. Individual *HLA-A, -B, -C*, and *-DRB1* genotypes are no major factors which determine COVID-19 severity. *Front. Immunol.* 12: 698193.
63. Kim, Y., J. Sidney, C. Pinilla, A. Sette, and B. Peters. 2009. Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior. *BMC Bioinformatics* 10: 394.
64. Dosztányi, Z., and A. E. Torda. 2001. Amino acid similarity matrices based on force fields. *Bioinformatics* 17: 686–699.
65. Sacks, P. 2017. Metric spaces. In *Techniques of Functional Analysis for Differential and Integral Equations.* Academic, New York, p. 35–50.
66. van Deutekom, H. W., and C. Keşmir. 2015. Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most? *Immunogenetics* 67: 425–436.
67. Reynisson, B., B. Alvarez, S. Paul, B. Peters, and M. Nielsen. 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 48(W1): W449–W454.
68. Duvaud, S., C. Gabella, F. Lisacek, H. Stockinger, V. Ioannidis, and C. Durinx. 2021. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res.* 49(W1): W216–W227.
69. Trolle, T., C. P. McMurtrey, J. Sidney, W. Bardet, S. C. Osborn, T. Kaever, A. Sette, W. H. Hildebrand, M. Nielsen, and B. Peters. 2016. The length distribution of class I–restricted T cell epitopes is determined by both peptide supply and MHC allele–specific binding preference. *J. Immunol.* 196: 1480–1487.
70. Hunter, J. D. 2007. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9: 90–95.
71. Waskom, M. L. 2021. Seaborn: statistical data visualization. *J. Open Source Softw.* 6: 3021.
72. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg. 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12: 2825–2830.
73. Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al.; SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. [Published erratum appears in 2020 *Nat. Methods* 17: 352.] *Nat. Methods* 17: 261–272.
74. Letunic, I., and P. Bork. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49(W1): W293–W296.
75. Smith, S. P., and R. Dubes. 1980. Stability of a hierarchical clustering. *Pattern Recognit.* 12: 177–187.
76. Carlsson, G. E., and F. Mémoli. 2010. Characterization, stability and convergence of hierarchical clustering methods. *J. Mach. Learn. Res.* 11: 1425–1470.
77. Saunders, A., D. Ashlock, and S. Houghten. 2018. *Hierarchical clustering and tree stability*. In *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, May 30–June 2*. IEEE, St. Louis, MO.
78. Halkidi, M., Y. Batistakis, and M. Vazirgiannis. 2002. Cluster validity methods: part I. *SIGMOD Rec.* 31: 40–45.
79. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915–10919.
80. Murthy, V. L., and L. J. Stern. 1997. The class II MHC protein HLA-DR1 in complex with an endogenous peptide: implications for the structural basis of the specificity of peptide binding. *Structure* 5: 1385–1396.
81. Wucherpfennig, K. W., B. Yu, K. Bhol, D. S. Monos, E. Argyris, R. W. Karr, A. R. Ahmed, and J. L. Strominger. 1995. Structural basis for major histocompatibility complex (MHC)-linked susceptibility to autoimmunity: charged residues of a single MHC binding pocket confer selective presentation of self-peptides in pemphigus vulgaris. *Proc. Natl. Acad. Sci. USA* 92: 11935–11939.
82. Mathieu, A., F. Paladini, A. Vacca, A. Cauli, M. T. Fiorillo, and R. Sorrentino. 2009. The interplay between the geographic distribution of HLA-B27 alleles and their role in infectious and autoimmune diseases: a unifying hypothesis. *Autoimmun. Rev.* 8: 420–425.
83. Elahi, S., W. L. Dinges, N. Lejarcegui, K. J. Laing, A. C. Collier, D. M. Koelle, M. J. McElrath, and H. Horton. 2011. Protective HIV-specific $CD8^+$ T cells evade Treg cell suppression. [Published erratum appears in 2011 *Nat. Med.* 17: 1153.] *Nat. Med.* 17: 989–995.
84. Nielsen, C. M., J. Vekemans, M. Lievens, K. E. Kester, J. A. Regules, and C. F. Ockenhouse. 2018. RTS,S malaria vaccine efficacy and immunogenicity during *Plasmodium falciparum* challenge is associated with HLA genotype. *Vaccine* 36: 1637–1642.
85. Barber, L. D., L. Percival, N. M. Valiante, L. Chen, C. Lee, J. E. Gumperz, J. H. Phillips, L. L. Lanier, J. C. Bigge, R. B. Parekh, and P. Parham. 1996. The inter-locus recombinant HLA-B*4601 has high selectivity in peptide binding and functions characteristic of HLA-C. *J. Exp. Med.* 184: 735–740.
86. Vita, R., S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette, and B. Peters. 2019. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* 47(D1): D339–D343.
87. Sette, A., J. Sidney, M.-F. del Guercio, S. Southwood, J. Ruppert, C. Dahlberg, H. M. Grey, and R. T. Kubo. 1994. Peptide binding to the most frequent HLA-A class I alleles measured by quantitative molecular binding assays. *Mol. Immunol.* 31: 813–822.
88. Shulman-Peleg, A., R. Nussinov, and H.J. Wolfson. 2005. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Res.* 33(Web Server issue): W337–W341.
89. Gao, M., and J. Skolnick. 2013. APoc: large-scale identification of similar protein pockets. *Bioinformatics* 29: 597–604.
90. Yu, D. J., J. Hu, J. Yang, H. B. Shen, J. Tang, and J. Y. Yang. 2013. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 10: 994–1008.
91. Maheshwari, S., and M. Brylinski. 2015. Predicting protein interface residues using easily accessible on-line resources. *Brief. Bioinform.* 16: 1025–1034.
92. Lee, H. S., and W. Im. 2016. G-LoSA: an efficient computational tool for local structure-centric biological studies and drug design. *Protein Sci.* 25: 865–876.
93. Simonovsky, M., and J. Meyers. 2020. DeeplyTough: learning structural comparison of protein binding sites. *J. Chem. Inf. Model.* 60: 2356–2366.
94. Léger, C., T. Di Meo, M. Aumont-Nicaise, C. Velours, D. Durand, I. Li de la Sierra-Gallay, H. van Tilbeurgh, N. Hildebrandt, M. Desmadril, A. Urvoas, et al. 2019. Ligand-induced conformational switch in an artificial bidomain protein scaffold. *Sci. Rep.* 9: 1178.
95. Zarutskie, J. A., A. K. Sato, M. M. Rushe, I. C. Chan, A. Lomakin, G. B. Benedek, and L. J. Stern. 1999. A conformational change in the human major histocompatibility complex protein HLA-DR1 induced by peptide binding. *Biochemistry* 38: 5878–5887.
96. Kumar, P., A. Vahedi-Faridi, W. Saenger, A. Ziegler, and B. Uchanska-Ziegler. 2009. Conformational changes within the HLA-A1:MAGE-A1 complex induced by binding of a recombinant antibody fragment with TCR-like specificity. *Protein Sci.* 18: 37–49.
97. Sadegh-Nasseri, S., M. Chen, K. Narayan, and M. Bouvier. 2008. The convergent roles of tapasin and HLA-DM in antigen presentation. *Trends Immunol.* 29: 141–147.
98. Sieker, F., S. Springer, and M. Zacharias. 2007. Comparative molecular dynamics analysis of tapasin-dependent and -independent MHC class I alleles. *Protein Sci.* 16: 299–308.
99. Eyal, E., S. Gerzon, V. Potapov, M. Edelman, and V. Sobolev. 2005. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J. Mol. Biol.* 351: 431–442.
100. Jiang, J., D. K. Taylor, E. J. Kim, L. F. Boyd, J. Ahmad, M. G. Mage, H. V. Truong, C. H. Woodward, N. G. Sgourakis, P. Cresswell, et al. 2022. Structural mechanism of tapasin-mediated MHC-I peptide loading in antigen presentation. *Nat. Commun.* 13: 5470.
101. Cheng, X., Y. Mei, X. Ji, Q. Xue, and D. Chen. 2016. Molecular mechanism of the susceptibility difference between HLA-B*27:02/04/05 and HLA-B*27:06/09 to ankylosing spondylitis: substitution analysis, MD simulation, QSAR modelling, and in vitro assay. *SAR QSAR Environ. Res.* 27: 409–425.
102. Aranha, M. P., Y. S. M. Jewel, R. A. Beckman, L. M. Weiner, J. C. Mitchell, J. M. Parks, and J. C. Smith. 2020. Combining three-dimensional modeling with artificial intelligence to increase specificity and precision in peptide-MHC binding predictions. *J. Immunol.* 205: 1962–1977.
103. Pöhlmann, T., R. A. Böckmann, H. Grubmüller, B. Uchanska-Ziegler, A. Ziegler, and U. Alexiev. 2004. Differential peptide dynamics is linked to major histocompatibility complex polymorphism. *J. Biol. Chem.* 279: 28197–28201.
104. Del Porto, P., M. D'Amato, M. T. Fiorillo, L. Tuosto, E. Piccolella, and R. Sorrentino. 1994. Identification of a novel HLA-B27 subtype by restriction analysis of a cytotoxic gamma delta T cell clone. *J. Immunol.* 153: 3093–3100.
105. Ramos, M., A. Paradela, M. Vazquez, A. Marina, J. Vazquez, and J. A. Lopez de Castro. 2002. Differential association of HLA-B*2705 and B*2709 to ankylosing spondylitis correlates with limited peptide subsets but not with altered cell surface stability. *J. Biol. Chem.* 277: 28749–28756.
106. Chen, B., J. Li, C. He, D. Li, W. Tong, Y. Zou, and W. Xu. 2017. Role of HLA-B27 in the pathogenesis of ankylosing spondylitis (Review). *Mol. Med. Rep.* 15: 1943–1951.
107. Tong, J. C., T. W. Tan, and S. Ranganathan. 2007. In silico grouping of peptide/HLA class I complexes using structural interaction characteristics. *Bioinformatics* 23: 177–183.

Key Points

- Structural similarity is highly correlated with peptide binding specificity.
- A comprehensive HLA supertype classification method is presented.
- Improved breadth, accuracy, flexibility, and stability relative to previous methods are shown.