Briefings in Bioinformatics, 2022, **23(3)**, 1–13

https://doi.org/10.1093/bib/bbac076

Problem Solving Protocol

HLA3D: an integrated structure-based computational toolkit for immunotherapy

Xingyu Li[†], Xue Lin[†], Xueyin Mei, Pin Chen, Anna Liu, Weicheng Liang, Shan Chang and Jian Li

Corresponding author: Jian Li, Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, 210096, China.
Tel: +86-13052881142; E-mail: jianli2014@seu.edu.cn

[†]These authors contributed equally to this work.

Abstract

The human major histocompatibility complex (MHC), also known as human leukocyte antigen (HLA), plays an important role in the adaptive immune system by presenting non-self-peptides to T cell receptors. The MHC region has been shown to be associated with a variety of diseases, including autoimmune diseases, organ transplantation and tumours. However, structural analytic tools of HLA are still sparse compared to the number of identified HLA alleles, which hinders the disclosure of its pathogenic mechanism. To provide an integrative analysis of HLA, we first collected 1296 amino acid sequences, 256 protein data bank structures, 120 000 frequency data of HLA alleles in different populations, 73 000 publications and 39 000 disease-associated single nucleotide polymorphism sites, as well as 212 modelled HLA heterodimer structures. Then, we put forward two new strategies for building up a toolkit for transplantation and tumour immunotherapy, designing risk alignment pipeline and antigenic peptide prediction pipeline by integrating different resources and bioinformatic tools. By integrating 100 000 calculated HLA conformation difference and online tools, risk alignment pipeline provides users with the functions of structural alignment, sequence alignment, residue visualization and risk report generation of mismatched HLA molecules. For tumour antigen prediction, we first predicted 370 000 immunogenic peptides based on the affinity between peptides and MHC to generate the neoantigen catalogue for 11 common tumours. We then designed an antigenic peptide prediction pipeline to provide the functions of mutation prediction, peptide prediction, immunogenicity assessment and docking simulation. We also present a case study of hepatitis B virus mutations associated with liver cancer that demonstrates the high legitimacy of our antigenic peptide prediction process. HLA3D, including different HLA analytic tools and the prediction pipelines, is available at http://www.hla3d.cn/.

Keywords: HLA structure, mutation, organ transplantation, neoantigen prediction

Introduction

The human major histocompatibility complex (MHC), also known as human leukocyte antigen (HLA), is indispensable in the adaptive immune system by delivering non-self-peptides for capture by T cells. The MHC region has been shown to be associated with a variety of diseases and therapies, including autoimmune diseases [1], tumour [2] and organ transplantation [3]. Some pathogenic MHC associations have been uncovered by genome-wide association study (GWAS), such as psoriasis, myasthenia gravis and ankylosing spondylitis. However, compared with the number of pathogenic variants of MHC, little progress has been made in the study of its pathogenesis and immune mechanisms [4]. In addition to the high degree of polymorphism in this region, its complexity is also reflected in the genetic heterogeneity among different populations. Explore the change of MHC–peptide binding pattern could help to reveal the pathogenesis of MHC-related diseases. As for the methods for predicting the binding affinity between HLA-I molecules and peptides, there are some excellent tools available in recent years [5, 6], but there is still a lack of a comprehensive platform, which can directly analyse the interaction between HLA molecules and peptides based

Xingyu Li is a master student at the Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, China.
Xue Lin is an assistant professor at the Department of Bioinformatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China. Her research focuses on bioinformatics, data mining and machine learning.
Xueyin Mei is a master student at the Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, China.
Pin Chen is a master student at the Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, China.
Anna Liu is a master student at the Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, China.
Weicheng Liang is a master student at the Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, China.
Shan Chang is a professor at Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, China. His research interests include bioinformatics, artificial intelligence and molecular simulation.
Jian Li is a professor at the Key Laboratory of DGHD, MOE, School of Life Science and Technology, Southeast University, Nanjing, China. His research interests lie in bioinformatics, genomics and big data computing.

Received: December 20, 2021. **Revised:** February 12, 2022. **Accepted:** February 14, 2022

on their structures. This may be due to the structural data of HLA that is still sparse compared to the number of identified HLA alleles, and most published HLA structural databases focus on some prevalent HLA alleles [7–10]. In addition, previous homologous modelling of HLA molecules mostly used the basic local alignment search tool (BLAST) software to find structures with high sequence similarity as templates. However, we found that the sequence similarity between the top structures calculated by BLAST was very high, so it was difficult to determine which template was most suitable for modelling. As we know, the high sequence polymorphism of HLA is mainly consistent with its non-overlapping antigen binding groove, where individual amino acid substitutions may even alter the immunogenicity of the entire molecule. Therefore, the selection of template should consider not only the similarity of HLA sequences but also the similarity of their functional grooves.

In order to provide a comprehensive analysis of HLA molecular mechanism for different populations, we established HLA3D, an integrated toolkit including different types and dimensions of data around HLA and several useful tools, and we also collected tumour-associated mutations of significantly mutated genes (SMGs), HLA, and hepatitis B virus (HBV). For HLA, at the sequence level, HLA3D contains 1296 amino acid sequences and nucleotide sequences of MHC alleles, which included all the common and well-documented MHC class I alleles in the American, European and Chinese populations. At the structural level, HLA3D includes 256 HLA structures collected from protein data bank (PDB) and 212 predicted heterodimers obtained through homologous modelling and protein docking. Differently, apart from the norm of sequence similarity and structure resolution, the selection of the modelling template was also based on the knowledge of HLA antigen serological characteristic [11], the classification of supertype [12] and the prediction results of MHC peptide affinity by MHCcluster [13]. Moreover, HLA3D, combining the data of 120 000 frequencies, 73 000 literature and 39 000 single nucleotide polymorphism (SNP) sites from public datasets, provides the users with comprehensive HLA information query. For mutation, we collected 1743 tumour-related mutations by mining the literature, including 213 tumour-associated HLA mutations, 989 hotspot mutations of SMGs and 402 hepatitis B virus mutations associated with chronic hepatitis B or liver cancer in the Chinese population.

In order to promote the application of HLA structure data in the field of immunotherapy, we established a risk alignment pipeline for organ transplantation and an antigenic peptides prediction pipeline for the design of tumour vaccine. On the one hand, we quantified the structural differences of all HLA molecules in HLA3D, obtained more than 100 000 root mean square deviation (RMSD) records, and manually collected 26 amino acid mismatch sites associated with acute graft versus host disease (aGVHD) to construct the risk alignment pipeline. This platform provides the users with the functions of structure alignment, sequence alignment, residue visualization and risk report generation of mismatched HLA donors. By integrating the knowledge of immunogenicity of sequence and structure differences of HLA, we provide a reference for the priority of irrelevant HLA mismatched donors during hematopoietic stem cell transplantation. Also, we predicted more than 370 000 high-affinity mutated peptides and generated a catalogue of candidate neoantigens for common tumours. In addition, we took HBV mutations as an example to build a complete process of antigenic peptide prediction in HLA3D. Here, we took into account the key characteristics that affect the immunogenicity of the peptide and designed the antigenic peptide prediction pipeline by incorporating open source and our software. It provides users with the functions of mutation prediction, peptide prediction, immunogenicity assessment and docking simulation, which aims to narrow down the immunogenic peptides used in peptide-based vaccine design.

HLA3D toolkit does not only provide a convenient, user-friendly interface for users to search, browse, predict and download information about HLA genes but also provides users with useful pipelines to accomplish personalized prediction. We believe that the HLA3D toolkit will contribute to the further exploration of structure-based epitope prediction and the pathogenesis of MHC-related diseases and promote the application of HLA molecules in the field of immunotherapy. The HLA3D toolkit is publicly available at http://www.hla3d.cn/.

Methods and materials

The process of constructing HLA3D is depicted in Figure 1. The data of HLA3D were obtained from public datasets, literature and our laboratory. The detailed procedure is explained in the following sections.

Collection and collation of HLA basic information

In HLA3D, the sequence, frequency and literature information of HLA class I genes were obtained from public databases. We collected all common and well-documented HLA class I genes from The American Society for Histocompatibility and Immunogenetics (ASHI) CWD 2.0.0 catalogue [14], The European Federation for Immunogenetics (EFI) catalogue [15] and Chinese HLA CWD catalogue [16], and a total of 1296 HLA alleles were collected after the removal of alleles that did not encode proteins. Amino acid sequences and nucleotide sequences of 1296 HLA class I genes reported in the above three HLA-CWD catalogues were manually collected from IMGT/HLA (https://www.ebi.ac.uk/ipd/imgt/hla/). In all 120 000 pieces of HLA frequency data around the world were derived from the allele frequency net database (AFND) (http://www.allelefrequencies.net/default.asp). The detailed information of a total of 73 000 literature, covering hot research topics on

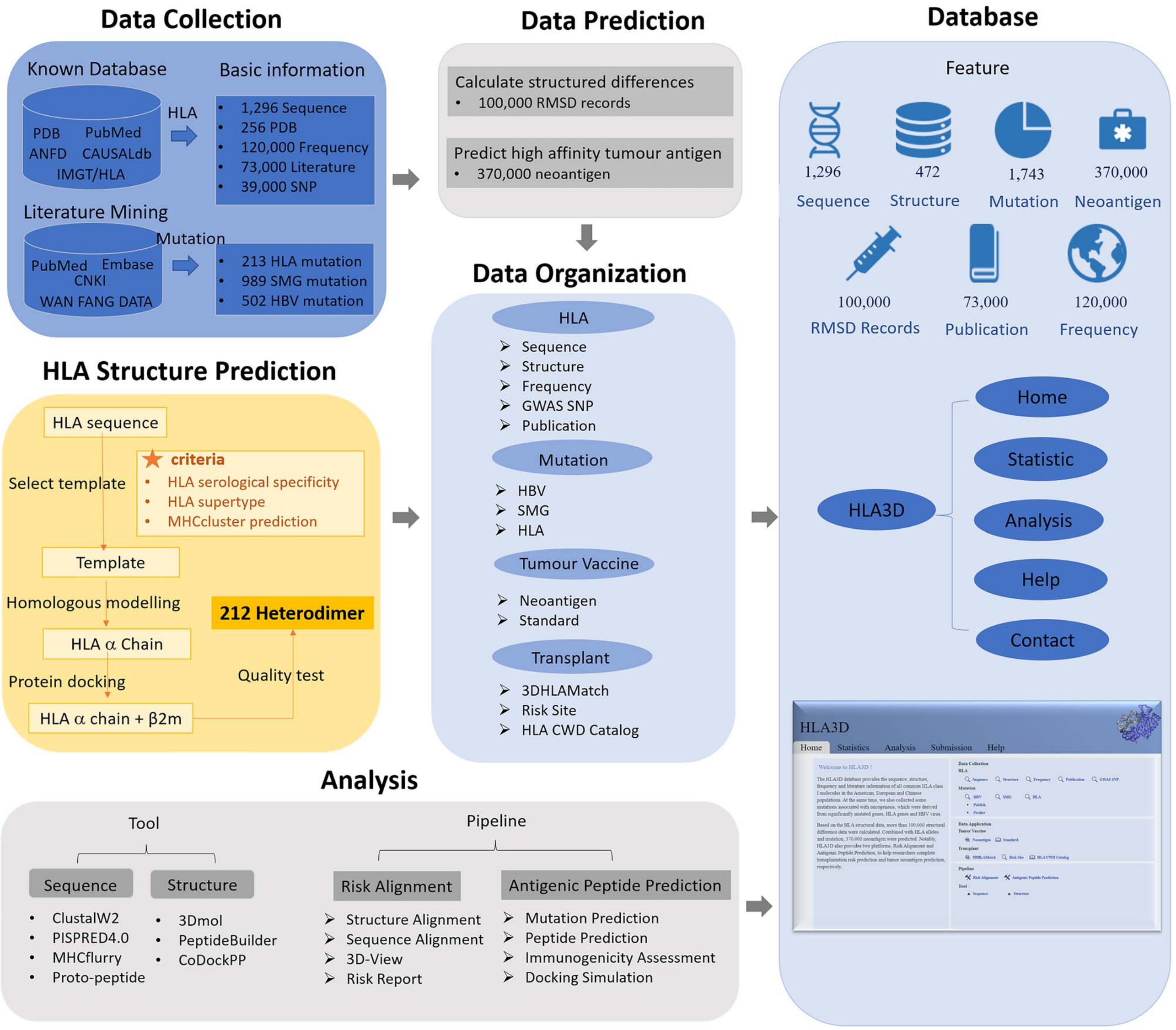


Figure 1. Pipeline for construction of the HLA3D toolkit.

immunotherapy, HLA matching and so on, was collected from PubMed (https://pubmed.ncbi.nlm.nih.gov/). And we also collected 39 000 HLA SNP sites associated with the diseases from CAUSALdb (http://mulinlab.tmu.edu.cn/causaldb).

Collection and construction of HLA structure data

The construction process of the HLA structure is shown in Figure 2. The structure data were retrieved from the PDB database or developed by homology modelling and protein docking. For HLA alleles with structures, we used the mmseq2 method to query in PDB according to the sequences of HLA molecules and collected all the matching structures [17], and a total of 256 structures were collected. For HLA molecules without structure information, 3D structures of 212 common HLA I genes were constructed by homology modelling, protein structure alignment and necessary quality tests. As shown in Figure 2, the complete HLA class I alpha chain amino acid sequences were collected from IPD-IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla/). Then, we selected the most suitable template for modelling according to the sequence similarity and functional similarity of HLA molecules. The selection of templates following four criteria: (i) select the structure with high sequence similarity and high resolution in the same serological group based on antigenicity of HLA [11]; (ii) if no template is available in the serological group, choose the equal in the same supertype group based on the peptide binding specificity [12]; (iii) use MHCcluster for HLA functional clustering [13], and pick out the structure with the closest peptide binding specificity (Figure 3) and (iv) for structures that do not meet the above criteria, the most suitable template was elected according to the sequence similarity and resolution. After the modelling work was accomplished, 212 predicted structures were submitted to SAVES (https://saves.mbi.ucla.edu/) for reliability test. Ramachandran plots of modelled structures were done and validated in PROCHECK [18]. Stereo-chemical excellence and overall quality was tested by Verify 3D

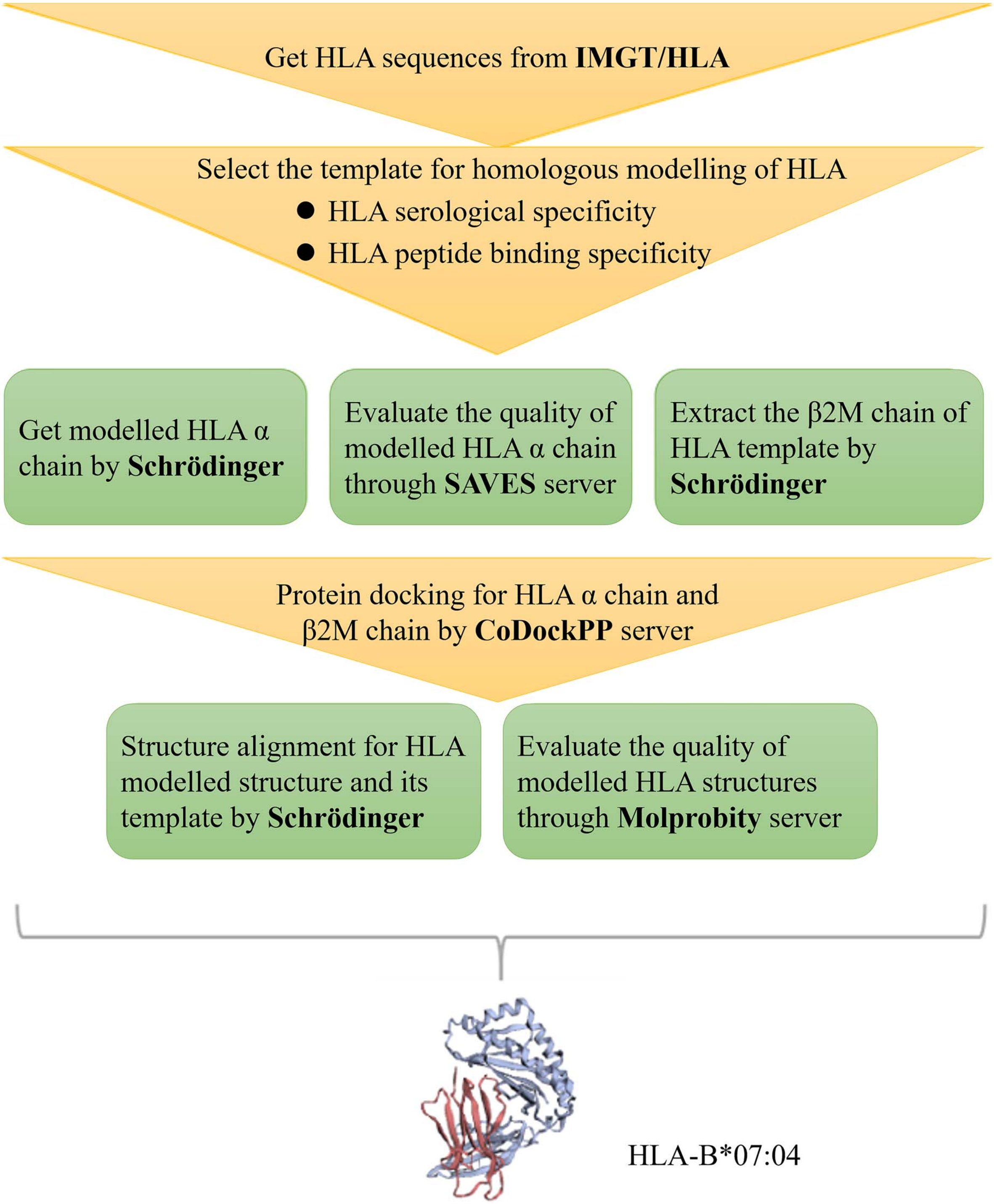

Get HLA sequences from **IMGT/HLA**
Select the template for homologous modelling of HLA
● HLA serological specificity
● HLA peptide binding specificity
Get modelled HLA α chain by **Schrödinger**
Evaluate the quality of modelled HLA α chain through **SAVES** server
Extract the β2M chain of HLA template by **Schrödinger**
Protein docking for HLA α chain and β2M chain by **CoDockPP** server
Structure alignment for HLA modelled structure and its template by **Schrödinger**
Evaluate the quality of modelled HLA structures through **Molprobity** server
HLA-B*07:04

Figure 2. The process of HLA structure construction. The complete HLA I class alpha chain amino acid sequences were collected from IPD-IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla/). The structure with high sequence similarity, high structural resolution and belonging to the same serological group or supertype group were used as template. We used the Advanced protein modelling function of Schrodinger (2020–24 release) to construct the 3D structure of the alpha chain of HLA molecule. Then, these predicted structures were submitted to SAVES (https://saves.mbi.ucla.edu/) for reliability test. Then, refined HLA class I alpha chains and the beta chains from templates were docked by CoDockPP software. Considering ligand RMSD and docking scores together, the best conformation was preserved. Finally, the heterodimers were uploaded to Molprobity (http://molprobity.biochem.duke.edu/) for quality test. The modelling information and quality parameters of each model structure are all recorded in HLA3D.

and ERRAT analysis [19, 20]. Then, the refined HLA class I alpha chain model were docked with the beta chain in the template by CoDockPP [21]. Considering ligand RMSD and docking scores together, the best conformation was preserved. Finally, the heterodimers were uploaded to Molprobity for quality test [22].

Acquisition and curation of mutation datasets

Mutation data in HLA3D were manually collected from the recent literature. We first screened the literature that met our standards in public databases, and then annotated the key information from it. Now, mutations in HLA3D are made up of three types of data, including

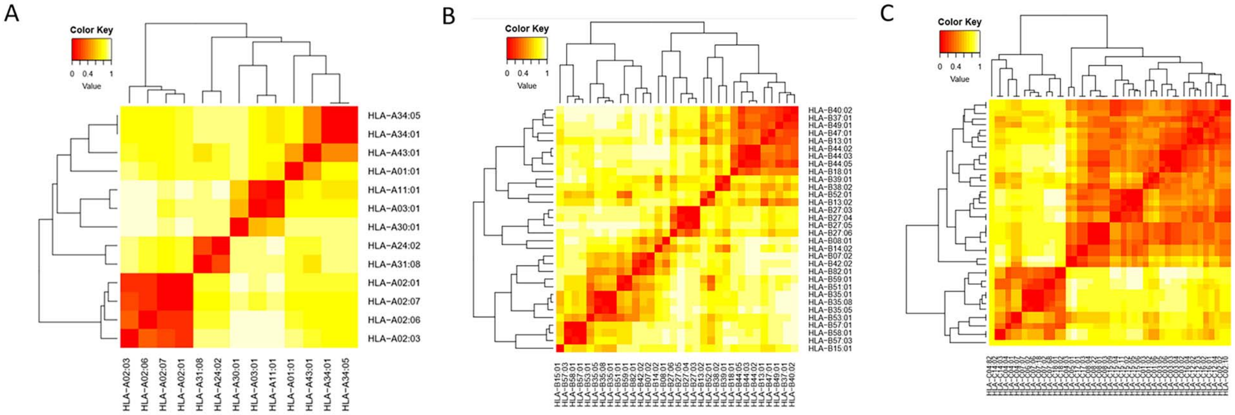


Figure 3. The clustering results of unclassifiable HLA-A(A), HLA-B(B) and HLA-C(C) genes.

213 tumour-associated HLA mutations, 989 hotspot mutations of significant mutated genes and 402 HBV mutations associated with chronic hepatitis B or liver cancer in the Chinese population. To ensure the accuracy of collected data about HBV, we manually collected a total of 1267 publications from Embase (https://www.embase.com), PubMed (https://pubmed.ncbi.nlm.nih.gov/) and other public databases. First, we filtered 259 publications based on four criteria: (i) removal of studies with duplicated results; (ii) removal of studies unrelated to HBV mutations; (iii) removal of studies involving HCV infection and (iv) removal of the review. Second, we reviewed the full text of these publications in detail and extracted the general information of HBV mutations related in liver cancer, and we obtained HBV mutations. Then, these mutations were annotated according to mutation, region, gene, gene type, disease and publication, which can be used as a reference panel to study the interaction mechanism between HBV virus and MHC immune system in the Chinese population. In addition, we collected 985 hotspot mutations based on a systematic analysis of 3281 tumours from 12 cancer types, covering 127 SMG involved in different signal pathways and enzyme processes [23]. Since HLA mutations may cause immune escape in tumours, 213 possible HLA functional mutations, including loss of function events (nonsense, frame-insertion loss), were identified from a WES (whole exome sequencing) analysis of 7930 cancer patients [24].

Acquisition of HLA differential immunogenicity data

We provided a reference for the preferable selection of recipient-donor pairs based on HLA structure and sequence differences. On the one hand, we calculated the differential immunogenicity of HLA structures in HLA3D. The distinction in peptide-binding groove between all structures was quantified with the protein structure alignment facilities of Prime (Schrödinger 2020–24 release). It uses double dynamic programming to align secondary structure elements and an iterative rigid body superposition that minimizes the RMSD of Cα atoms. In particular, amino acid residues (AAR) used for the calculation did not include residues that are not involved in antigen presentation or T-cell receptor (TCR) recognition, such as random curls. The region of AAR used for RMSD calculations were aa3-13, aa20-38, aa45-85, aa92-103, aa109-127 and aa132-178 [25]. A total of over 100 000 RMSD records were obtained and organized in HLA3D. On the other hand, at the sequence level, we manually collected 26 amino acid mismatch sites reported in the literature that are related to immune rejection after transplantation [2, 26–31]. The active sites of amino acid in HLA antigen-binding pocket, which interacted with peptide or TCR, were annotated, as well [32]. Moreover, the common and well-documented (CWD) status of the HLA gene in European, American and Chinese populations [14–16] was incorporated to provide the information of the distribution and frequency of candidate donors.

Acquisition of high affinity neoantigen data

We predicted the binding affinity between mutated peptides generated from 989 hotspot mutations in 11 common tumours and 250 common HLA class I alleles in the HLA3D toolkit, and more than 370 000 predicted immunogenic peptides were obtained and organized. Firstly, we collected the amino acid sequences of the 127 SMGs in the Uniprot database (https://www.uniprot.org/). Then, we used Java language to write a script for batch generation of overlapping peptides, and by setting the original residue, position, new residue and length, we could get the overlapping peptides containing mutations. Secondly, we used NetMHCpan4.0 [33] to predict the binding affinity of 8–11 amino acid peptides to HLA molecules, and only strong binders (SBs) or weak binders (WBs) were retained according to the rank. Considering that the presentation and recognition of antigenic peptides is a complex process, we collected some key parameters to facilitate the screening of antigenic peptides. The key parameters controlling tumour immunogenicity in the tumour epitope immunogenicity model proposed by the tumour neoantigen selection alliance include MHC binding affinity, tumour abundance, MHC binding stability, agretopicity and foreignness [34].

Results

Toolkit contents

The HLA3D toolkit includes 1296 amino acid sequences of HLA alleles (Figure 4A), 256 PDB structures and 212 modulated structures (Figure 4B), 120 000 frequency data, 73 000 references, 39 000 SNP sites and 1743 mutations associated with tumorigenesis. In addition, we calculated the structural differences in HLA antigen binding grooves and obtained more than 100 000 RMSD records. RMSD data of HLA-A molecules are concentrated in the range of 0.4–0.7. Values lower than 0.4 are considered relatively 'low' in the overall range, and values higher than 0.8 are relatively considered 'high'. The structural difference data of HLA-B molecules are distributed evenly, and the structural difference data of HLA-C class molecules are mostly concentrated in the two ranges of 0–0.2 and 0.5–0.8 (Figure 4C). We also took hotspot mutations in common tumours as an example and predicted more than 370 000 neoantigens with high affinity. We classified and counted the predicted antigenic peptides according to different tumour types (Figure 4D) and sorted the top 20 genes and HLA alleles with the highest number of predicted antigens in different tumours (Figure 4E and F).

Website interface

The main components of the HLA3D web server are interrelated tables, datasets, pipelines and tools for immunotherapy based on HLA structural data. Figure 5 shows the user interface of HLA3D website and describes the functions of main page.

Searching and browsing

The data collection module provides users with concise pages to search, download and visualize information about HLA class I molecules and tumour-associated mutations (Figure 5A). The HLA section provides five types of HLA data: (i) Sequence: users can search by HLA allele to get the amino acid sequence and nucleotide sequence; (ii) Structure: users can search by HLA allele to get the structural information of HLA (Figure 5B). For the structures derived from PDB, we annotated the PDB ID, resolution, description and literature of each. For structures through modelling, we annotated the templates for their alpha chains, the beta chains and the quality information of each HLA molecules (Figure 5C). We also integrated the 3Dmol tool to provide the visualization of the protein structures and peptide epitopes. You can download the PDB structure files from the download tab; (iii) Frequency: the frequency information of HLA around the world is provided, including population, sample size, allele frequency and other information; (iv) Publication: HLA-related research literature, including title, author, keywords and other information was provided and (v) GWAS SNP: HLA SNP information related to GWAS was provided according to 14 traits such as anatomy, organism and disease. While the Mutation section provides users with mutations in the case of SMGs, HLA and HBV viruses. For HLA and significant mutated genes in common tumours, interested mutations can be retrieved by searching the gene or the cancer type. As for HBV, two web pages are designed: (i) Publish: HBV mutations related to liver cancer or chronic hepatitis B in the Chinese population are provided here. Interested mutations can be retrieved by searching the gene and gene type; (ii) Predict: the analysis process of HBV hotspot mutations associated with liver cancer in American and Chinese populations is shown.

The data application module provides data that we have calculated or simulated using bioinformatic tools based on HLA structure or sequence data. For the tumour vaccine part, we have designed two web pages: neoantigen and standard (Figure 5A). Neoantigen provides the information of predicted antigen peptides, which can be identified by searching by the keywords of cancer, gene, mutation, HLA allele and peptide. Users can also view the information of the mutation state wild type or mutant type (WT or MUT), position and bind level (SB or WB) of antigenic peptides. To help users to verify the antigenicity of antigenic peptides, we also provide an immune epitope database and analysis resource (IEDB) browser (https://www.iedb.org/). The standard subpage provides detailed descriptions of the key parameters controlling tumour immunogenicity: MHC binding affinity, tumour abundance, MHC binding stability, agretopicity and foreignness. For the transplant part, we have designed three pages: 3DHLAMatch, risk site and HLA CWD catalogue (Figure 5A). The 3DHLAMatch page provides data on the structural differences of all HLA structure antigen-binding pocket in HLA3D. Users can search for RMSD scores of common HLA molecular structure antigen-binding pockets by HLA allele, aligned allele or gene pair. Risk site page demonstrates the amino acid mismatch sites associated with acute rejection after transplantation. HLA CWD catalogue page provides the prevalence and sample information of HLA in America, Europe and China.

Functionality

We integrated different types of data and tools and established two pipelines to promote the application of HLA molecules in the biomedical field (Figure 5D). Details of the usage of tools can be found in the tool manual in HLA3D (listed in the Supplementary Data).

For stem cell therapy and organ transplantation, we have built the risk alignment pipeline to help users assess the transplant risk of mismatched HLA donors. We integrated the structural difference data of HLA molecules and the collected HLA sequence mismatch sites as data resources. Then, we integrated ClusterW2 and 3Dmol [35]

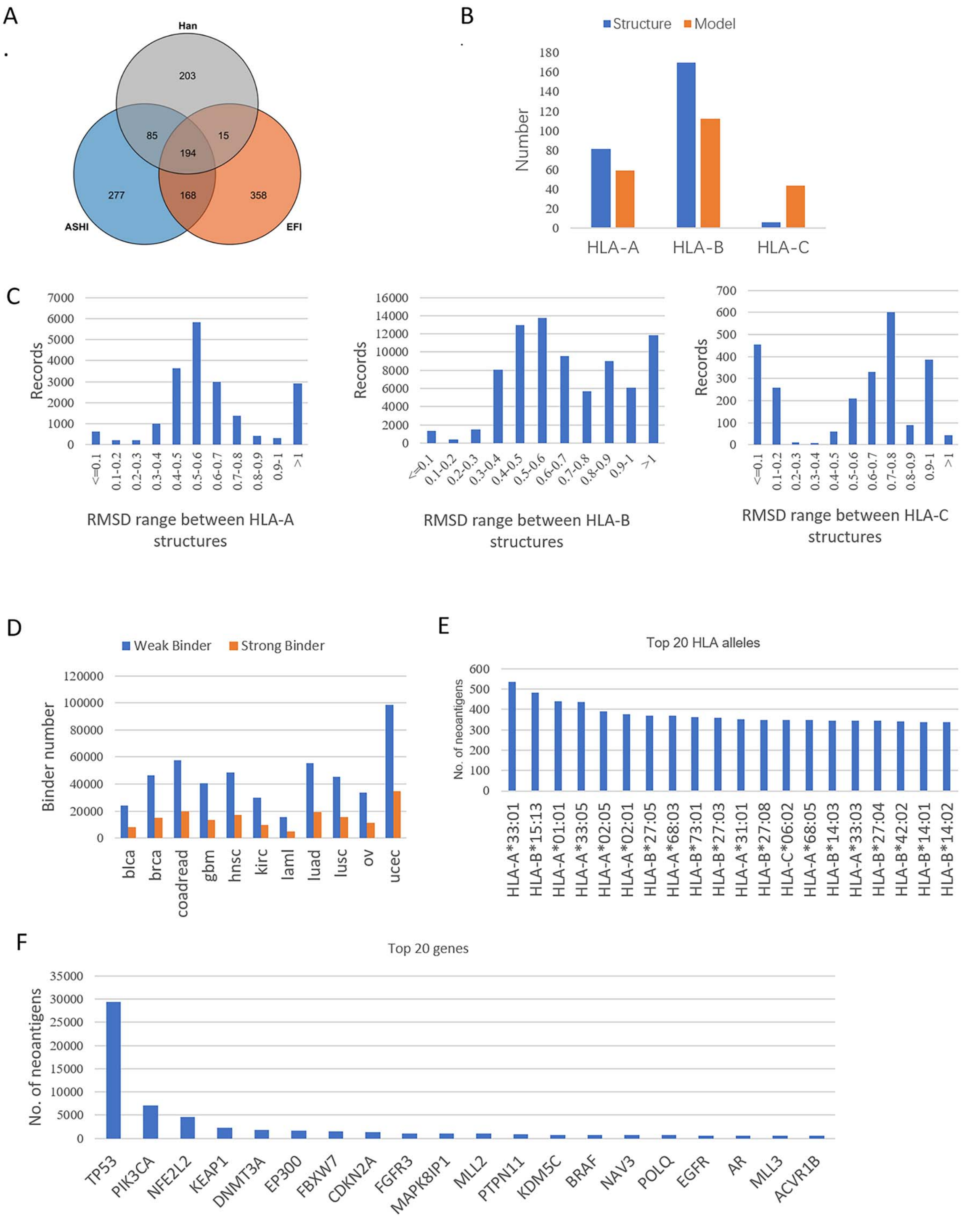
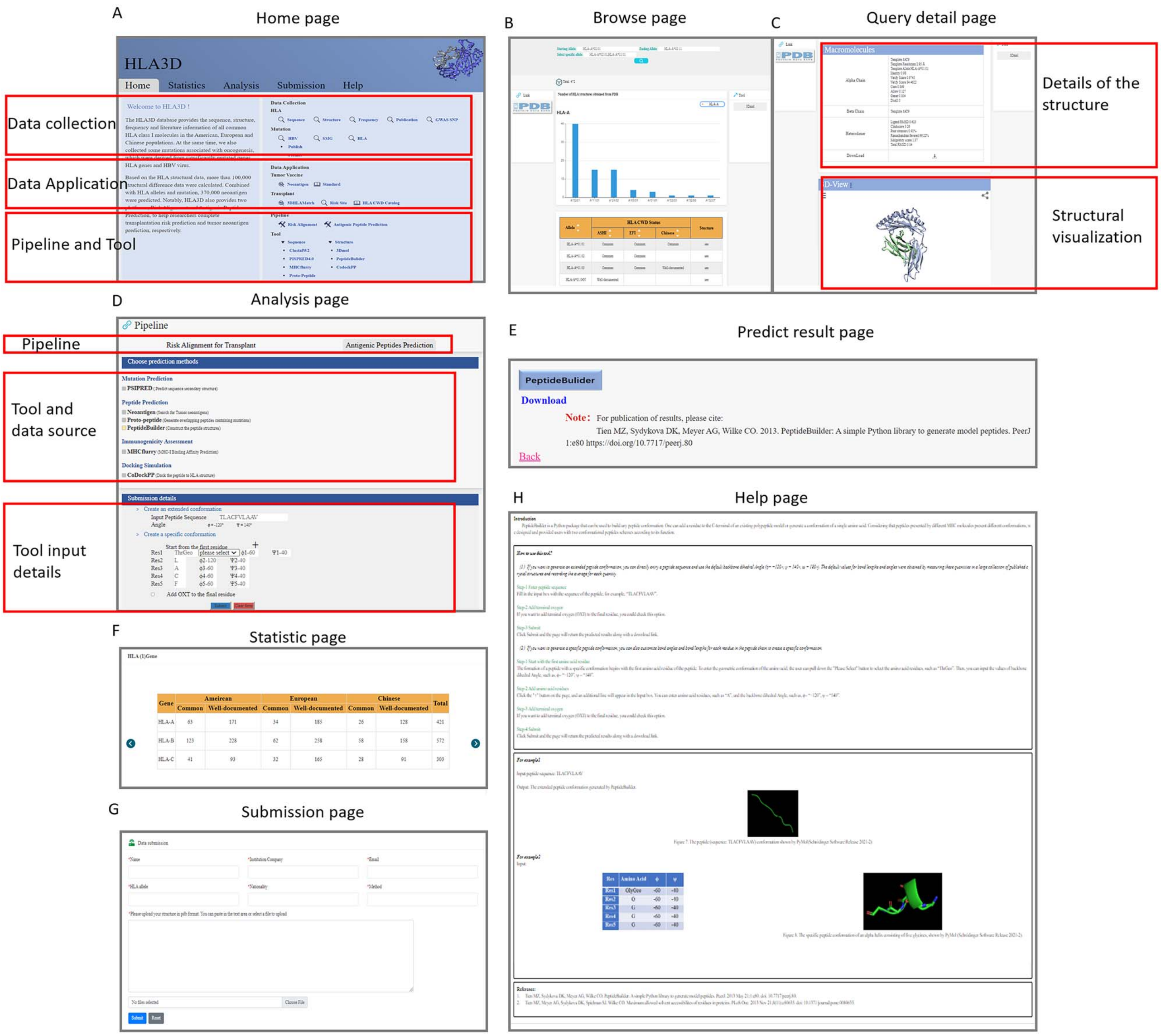


Figure 4. Statistical analysis of the data in HLA3D toolkit. (A) The collection of HLA class I alleles from ASHI CWD2.0.0 Catalogue (ASHI), EFI CWD catalogue (EFI) and Chinese CWD catalogue (Han). (B) The collection of HLA structures in HLA3D. Blue bars represent crystal structure from PDB. Orange bars represent structure obtained through homologous modelling and protein docking. (C) The range of RMSD records of HLA-A, HLA-B and HLA-C structures in HLA3D. (D) The predicted neoantigens in common tumours. Blue bars represent WBs and orange bars represent SBs. The definition of WB and SB is based on the Rank calculated in NetMHCpan4.0. The top 20 HLA alleles (E) and genes (F) with the most neoantigens in lung squamous cell carcinoma.

tool to provide users with the function of sequence alignment and 3D visualization. Finally, a risk assessment report would be generated for users.

For tumour vaccine, we established the antigenic peptide prediction pipeline to help users to predict and screen immunogenic peptides. We integrated the PSRPRED4.0 [36] to provide the users with sequence secondary analysis to narrow the range of candidate mutations. Then, we provide the users with the function of peptide prediction. Users can not only query on the Neoantigen page to access the information of predicted antigenic peptides with high affinity in common tumours but also obtain overlapping peptides containing the candidate mutations using our own Proto-peptide tool.



Welcome to HLA3D !

The HLA3D database provides the sequence, structure, frequency and literature information of all common HLA class I molecules in the American, European and Chinese populations. At the same time, we also collected some mutations associated with oncogenesis, which were derived from significantly mutated genes, HLA genes and HBV virus.

Based on the HLA structural data, more than 100,000 structural difference data were calculated. Combined with HLA alleles and mutation, 370,000 neoantigen were predicted. Notably, HLA3D also provides two platforms, Risk Alignment and Antigenic Peptide Prediction, to help researchers complete transplantation risk prediction and tumor neoantigen prediction, respectively.

Input Peptide Sequence TLACFVLAAV
Angle ϕ = -120° Ψ = 140°

Start from the first residue

Res1	ThrGeo	please select	ϕ1-60	Ψ1-40
Res2	L		ϕ2-120	Ψ2-40
Res3	A		ϕ3-60	Ψ3-40
Res4	C		ϕ4-60	Ψ4-40
Res5	F		ϕ5-60	Ψ5-40

□ Add OXT to the final residue

Note : For publication of results, please cite:
Tien MZ, Sydykova DK, Meyer AG, Wilke CO. 2013. PeptideBuilder: A simple Python library to generate model peptides. PeerJ 1:e80 https://doi.org/10.7717/peerj.80

Gene	Ameircan		European		Chinese		Total
	Common	Well-documented	Common	Well-documented	Common	Well-documented	
HLA-A	63	171	34	185	26	128	421
HLA-B	123	228	62	258	58	158	572
HLA-C	41	93	32	165	28	91	303

*Please upload your structure in pdb format. You can paste in the text area or select a file to upload.

No files selected

Figure 5. The interface of HLA3D toolkit. (A) Home page: users can search different types of data on HLA molecules and tumour-associated mutations here. (B) Browse page: users can browse specific entries via search on different page. The query condition of HLA structure page is shown here. Users can search for HLA structure by entering the name of HLA allele. (C) Query detail page: shown here is the response for searching for HLA structure, including annotation information and structural visualization results. (D) Analysis page: two constructed Pipelines are presented here, namely Risk Alignment and Antigenic Peptides Prediction. Users can flexibly choose tools to realize different predictions. (E) Predict result page: users can download the predicted results here. (F) Statistic page: the statistical details of the data in the HLA3D toolkit are shown here. (G) Submission page: users can upload structural data of HLA molecules in this page. (H) Help page: this page contains a description of the HLA3D toolkit, including toolkit background, data sources, tool sources, data feature and detailed description of the pipelines and tools usage.

We also integrated the MHCflurry [37] package and designed convenient interface to provide the function of immunogenicity assessment based on peptide-HLA affinity. Finally, the function of docking simulation is provided for users to realize the molecular docking of peptide and HLA. We integrated the PeptideBuilder package [38] and designed the input page to provide users with two ways to generate peptide conformation (Figure 5D). Users can input a peptide sequence, generate an extended structure with default values for the backbone dihedral angles ($\phi = -120\circ$, $\psi = 140\circ$, $\omega = 180\circ$), and can also use their custom bond angles and bond lengths for every residue to create a specific conformation (Figure 5E). Moreover, we integrated our own CoDockPP software [21] and modified the site constraints of the software according to the binding characteristics of the peptide and HLA to provide the function of peptide–HLA docking. Users can upload the peptide and HLA structure through CoDockPP and predict the initial docking conformation for subsequent dynamic simulation and so on. Thus, users can dock the peptide to the HLA in Ambiguous type without any site constraints, and multiple docking between peptide and HLA molecule can be performed by inputting several sites (total sites < 8) on HLA molecule and peptide as well.

Example use case: predicting the differential immunogenicity of HLA mismatched donors by risk alignment pipeline

The risk alignment prediction process is shown in the Figure 6A. It provides users with the following functions: (i) structure alignment, which provides information about the conformational differences of HLA antigen binding groove; (ii) sequence alignment, which aims to provide users with the function of HLA sequence alignment; (iii) 3D-view, which aims to provide users with the function of structural visualization of HLA sequence mismatch sites and (iv) risk report, which aims to provide users with a transplant risk assess according to the function of mismatch sites. Here, we give an example of a transplant with a single allele level mismatch in HLA-B: B*35:01/B*35:08, to illustrate how to use the Risk Alignment pipeline to evaluate the aGVHD severity of mismatched donors.

Structure alignment

For example, typing 'HLA-B *35:01' on HLA Allele and 'HLA-B *35:08' on aligned allele. Since there are multiple PDB structures in these two HLA genes, the search results show that the antigen binding pocket difference between the two HLA genes is 0.174–0.313.

Sequence alignment

Amino acid sequences of 'HLA-B*35:01' and 'HLA-B*35*08' were uploaded for sequence alignment through ClustalW2, and the mismatched amino acid was highlighted in yellow. According to the output results, the 180th amino acid of HLA-B*35:01 and HLA-B*35*08 is inconsistent.

3D view

Upload the structure of HLA-B*35:01 (PDB ID: 4PRB) and locate the 180th amino acid on the 3D structure by 3Dmol [35]. The results showed that the site is located in the alpha helix of HLA and may be related to antigenic peptide or TCR recognition [39].

Risk report

The risk report generated the structural annotation information of amino acid mismatch sites of HLA-B*35:01 and HLA-B*35:08. This mismatch site is located in the α-helix of HLA and participates in the composition of HLA antigen binding pockets D and E and interacts with peptide ligands. It is a previously reported risk site [40].

Example use case: predicting the immunogenic peptides by antigenic peptide prediction pipeline

The Antigenic Peptide Prediction process is shown in the Figure 6B. It provides users with the following functions: (i) mutation prediction, which aims to help users narrow the range of candidate mutations; (ii) peptide prediction, which aims to help users obtain the sequence of the mutated peptides; (iii) immunogenicity assessment, which aims to help users obtain the potential immunogenic peptides and (iv) docking simulation, which aims to help users get the initial conformation of peptide–HLA docking. Users can flexibly choose different tools to meet their own research needs.

Mutation analysis

PSIPPRED4.0 provides users with the secondary structure prediction of the sequence [36]. Users can also refer to the HBV Predict page for HBV virus hotspot mutation analysis process, including sequence conservatism analysis, hydrophobicity analysis and transmembrane analysis and select other online tools to identify the hotspot mutations [41].

Peptide prediction

There are two ways for users to obtain mutated peptide sequences of interest. One is that, on the neoantigen page, it provides a catalogue of 11 common tumour neoantigens predicted with high affinity, and users can search for tumour antigen peptides of interest. We also provide users with histograms of the top 20 genes and HLA alleles to predict the number of antigenic peptides in each tumour (Figure 5E and F). And the other way is that the users can use the Proto-peptide tool to obtain the overlapping peptides containing mutations by setting up the parameters. It can be used as input for tools, such as NetMHCpan, to predict the affinity of 8–11 peptides to MHC molecules [33]. In addition, users can choose PeptideBuilder to generate the default extended peptide conformation or customize a specific peptide conformation [38].

Immunogenicity assessment

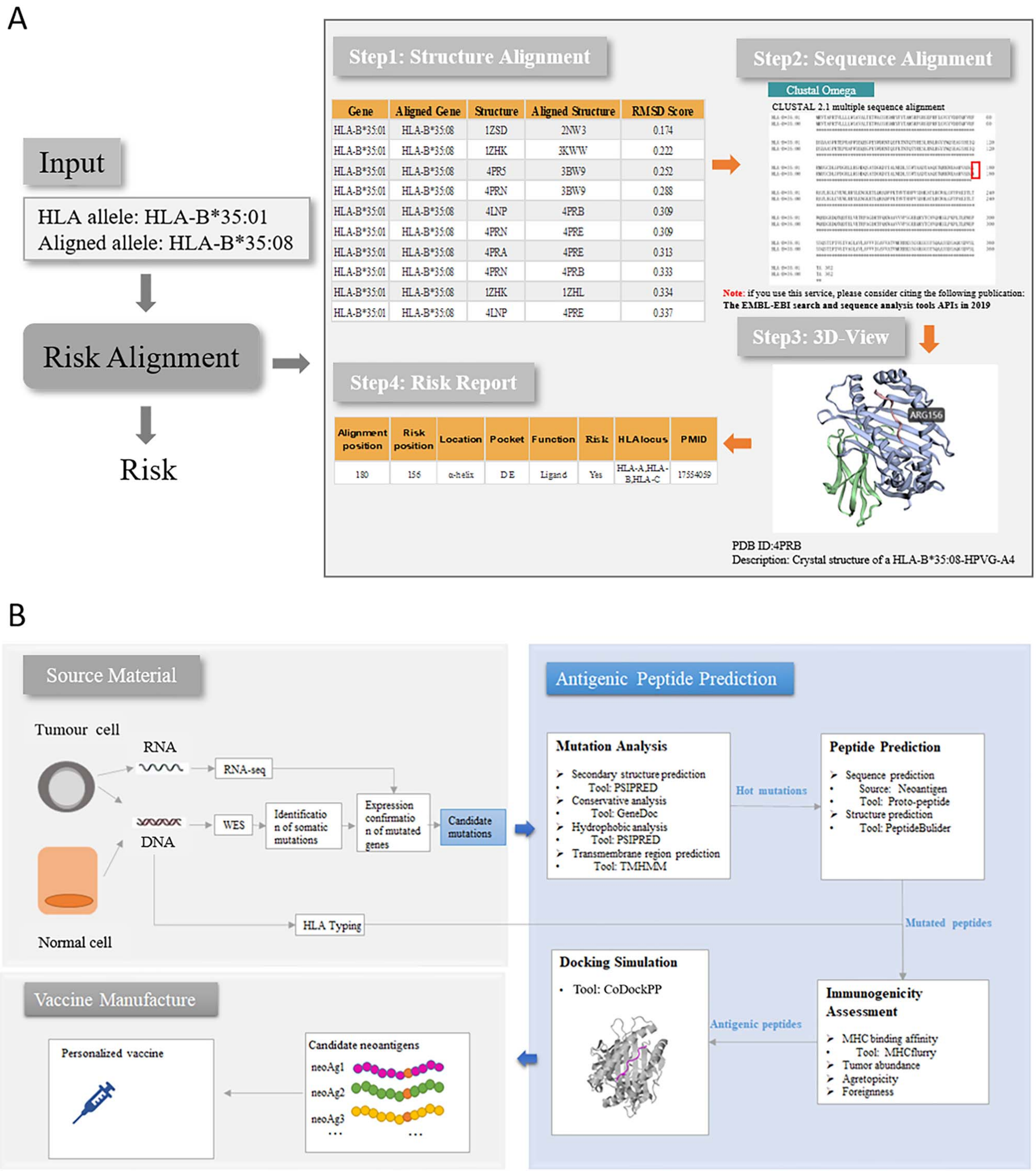
Here, we designed a user-friendly interface based on the key features of MHCFlurry [37]. It provides users with two methods to predict the binding affinity of mutated peptides. Users can choose 'MHCFlurry predict' method to predict the binding affinity of individual peptides to MHC molecules, or select 'MHCFlurry predict scan' method to scan protein sequences for epitopes. A more detailed tutorial can be found in the tools manual in HLA3D (listed in the Supplementary Data).

Docking simulation

According to the binding characteristics of HLA and peptide, we provide users with the tools to generate peptide conformation and to complete docking simulation. First, users can use PeptideBuilder [38] to generate any peptide structure based on the peptide sequence and the dihedral Angle of the backbone. Then, users can upload the structure of peptides and HLA to CoDockPP to obtain the initial docking conformation by defining the constraint sites on the peptide and HLA molecules [21].

Discussion

HLA is characterized by high polymorphism, high disease correlation and genetic heterogeneity, and its prevalence frequency varies greatly among different



Gene	Aligned Gene	Structure	Aligned Structure	RMSD Score
HLA-B*35:01	HLA-B*35:08	1ZSD	2NW3	0.174
HLA-B*35:01	HLA-B*35:08	1ZHK	3KWW	0.222
HLA-B*35:01	HLA-B*35:08	4PR5	3BW9	0.252
HLA-B*35:01	HLA-B*35:08	4PRN	3BW9	0.288
HLA-B*35:01	HLA-B*35:08	4LNP	4PRB	0.309
HLA-B*35:01	HLA-B*35:08	4PRN	4PRE	0.309
HLA-B*35:01	HLA-B*35:08	4PRA	4PRE	0.313
HLA-B*35:01	HLA-B*35:08	4PRN	4PRB	0.333
HLA-B*35:01	HLA-B*35:08	1ZHK	1ZHL	0.334
HLA-B*35:01	HLA-B*35:08	4LNP	4PRE	0.337

Note: if you use this service, please consider citing the following publication:
The EMBL-EBI search and sequence analysis tools APIs in 2019

Alignment position	Risk position	Location	Pocket	Function	Risk	HLAlocus	PMID
180	156	α-helix	DE	Ligand	Yes	HLA-A,HLA-B,HLA-C	17554059

PDB ID:4PRB
Description: Crystal structure of a HLA-B*35:08-HPVG-A4

Identification of somatic mutations

Expression confirmation of mutated genes

Candidate mutations

Mutation Analysis
- Secondary structure prediction
- Tool: PSIPRED
- Conservative analysis
- Tool: GeneDoc
- Hydrophobic analysis
- Tool: PSIPRED
- Transmembrane region prediction
- Tool: TMHMM

Peptide Prediction
- Sequence prediction
- Source: Neoantigen
- Tool: Proto-peptide
- Structure prediction
- Tool: PeptideBulider

Docking Simulation
- Tool: CoDockPP

Immunogenicity Assessment
- MHC binding affinity
- Tool: MHCflurry
- Tumor abundance
- Agretopicity
- Foreignness

Figure 6. Pipeline design for immunotherapy in the HLA3D toolkit. (A) The workflow of Risk Alignment pipeline. The prediction process of this pipeline consists of four steps, and the prediction results of mismatched donor HLA-B*35:01 and HLA-B*35:08 are show. (B) The interpretation of tumour vaccine design process. For tumour antigens, the first step is to get candidate mutations from source material, which contains the sequencing data of the tumour and normal cell. The HLA3D toolkit provides users with three types of candidate mutation data: HLA, SMG and HBV virus. Users can choose by themselves or upload their own mutation data for prediction. Then, these mutations could be analysed in the Antigenic Peptide Prediction pipeline. The grey line represents the information flow. This platform provides users with four types of prediction: mutation analysis, peptide prediction, immunogenicity assessment and docking simulation. Users can flexibly choose different tools or source to meet their research needs.

populations [42]. These sequence characteristics emphasize the importance of studying disease-related MHC genes in population-specific reference panels. Thanks to the establishment of sequence reference panels in recent years [4], many susceptible MHC genes were discovered in different populations, but little was understood about their pathogenesis. Exploring the change of MHC–peptide binding pattern is the key to revealing the pathogenesis of MHC-related diseases. However, the number of HLA structures available is still small compared with the number of identified HLA alleles, and the published HLA structural databases focused more on HLA alleles that are prevalent in the population.

Here, we developed a comprehensive toolkit of HLA, HLA3D, to provide different resources and useful pipelines and tools for different populations. The current version of HLA3D includes sequences and structures of all common MHC class I alleles in the American, European and Chinese populations. The structural data of the HLA include not only the PDB structures available in public databases but also those obtained by homology modelling and molecular docking. Some excellent HLA modelling tasks have been accomplished in the past, but the problem of how to choose the best template to retain the antigenicity and immunogenicity of the HLA structure has not been resolved. Here, we consider both functional similarity and sequence similarity between the template and the model. Structures with high sequence similarity, high structural resolution and belonging to the same serological group or supertype group were used as templates [11, 12]. Notably, the results of the three classification methods we used were basically the same. In most cases, the results of supertype classification follow the serological grouping. The clustering results of MHCcluster are also the same as the classification of supertype definition. Serological classification focuses on the antigenicity of HLA molecules, while supertype and MHCcluster both focus on the immunogenicity of HLA molecules, which are two basic characteristics of antigens. The modelling information and quality parameters of each model structure are recorded in HLA3D to ensure the accuracy of the structure.

The HLA3D web server not only provides an integrative resource for users who are interested in HLA but also aims to promote structure-based immunotherapy. We have designed the 'Risk Assessment' pipeline for organ transplantation and the 'Antigenic Peptide Prediction' pipeline for the design of tumour vaccine.

During the process of hematopoietic stem cell transplantation (HSCT), HLA compatibility between donor and recipient is closely related to the severity of acute graft-versus-host disease (aGVHD) [43]. However, some primary structure differences of HLA proteins do not lead to clinical allograft rejection [26]. This is because certain amino acids are not directly involved in the contact between MHC and antigenic peptides and TCR. For example, it is reported that the mismatches C*03:03/C*03:04 are the most common (68.7%) in grafts with a single-allele-level mismatch in HLA-C [30]. Virtually, structure-based predictions are superior to sequence-based ones because the binding feature of MHC–peptide hinges on the conformation, hydrophobicity and charge distribution of the groove. Here, we combined the knowledge of HLA structural differences and sequence differences to design a risk alignment pipeline to help users comprehensively evaluate the differential immunogenicity of HLA mismatched donors. Notably, the RMSD score is a quantitative indicator of HLA conformational differences, which provides a quick query method for assessing the severity of aGVHD before transplantation. A donor with a higher RMSD value is considered to have a higher risk of postoperative aGVHD and is not recommended as a transplant donor. Moreover, additional considerations regarding the position and function of HLA sequence mismatch sites in the 3D structure make the prediction more accurate. We believe that, in the future, HLA matching data in HLA3D toolkit could complement other immunogenicity prediction methods to provide more convenient and precise predictions for patients in urgent need of transplantation.

HLA molecules also play a crucial role in tumour immunotherapy. Tumour neoantigen is an abnormal protein that is not expressed in normal cells but is expressed in tumour cells and can activate the immune system. By obtaining immunogenic tumour antigens, personalized tumour vaccines can be prepared. However, the experimental method for detecting immunogenic peptides is very time-consuming. In fact, not all mutated antigens are immunogenic, only those antigens that bind to the MHC and are stably presented on the cell surface meet the requirements [3]. Therefore, we used bioinformatic approaches to predict the antigenic peptides of common tumour, thereby narrowing the range of candidate neoantigens and promoting subsequent experimental verification. In addition, the success of neoantigen identification lies not only in the identification of candidate antigenic peptides, but also in the evaluation of immunogenicity of antigenic peptides. At present, there are many tools available to predict immunogenicity of peptides using a single indicator, such as binding affinity or binding stability, which may lead to high false positive rates. We previously screened candidate mutations of HBV virus through conservative analysis, hydrophobicity analysis and transmembrane analysis. Then, molecular docking and molecular dynamics simulation were used to study the changes of affinity induced by these mutations. The final results explained the role of 11 HBV mutations in immune escape of liver cancer, including V351A and V354A, which are S region mutations of HBV subtype A [41]. We synthesized the mutant peptide and HLA-A*02:01 protein *in vitro* and measured the affinity using biofilm interference technology. The results showed that the mutant peptide had no affinity with HLA-A*02:01 protein, while the WT had affinity (Supplementary Figure 1, see Supplementary Data available online at https://academic.oup.com/bib). This is consistent with the results of our molecular dynamics simulation, indicating that our antigenic peptides analysis process is accurate. Here, in the case of HBV virus mutation, we integrated this prediction process into HLA3D toolkit and designed an antigenic peptide prediction pipeline for users to realize *de novo* prediction of immunogenic peptides. We have integrated different tools to design four functions for users, including mutation analysis, peptide prediction, immunogenicity assessment and docking simulation. Users can flexibly combine tools according to actual needs to complete the prediction and analysis of immunogenic peptides.

Currently, the structures and pipelines in HLA3D are all about the modelling and docking of MHC class I molecules and peptides. However, in the cellular immune response, the activation of TCR is essential. In the future, we will further understand the modelling and dynamic changes of TCRs and develop MHC–peptide-TCR docking benchmarks to promote the application of HLA molecular structure data in the field of immunotherapy.

Key Points

- To provide comprehensive analysis of HLA for different populations, we developed HLA3D, a comprehensive toolkit that collected 1296 sequences, 256 PDB structures, 212 modelled structures, 120 000 frequency data, 73 000 associated literature, 39 000 disease-associated SNP of HLA and 1604 oncogenic mutations.
- Based on common HLA molecules in HLA3D, we qualified the HLA structure differences and obtained 100 000 RMSD records. In addition, we predicted 370 000 antigenic peptides with high affinity in common tumours, which helps to narrow the range of candidate neoantigens and promote subsequent experimental verification.
- By integrating the knowledge of differential immunogenicity in HLA sequences and structures, HLA3D established a risk alignment pipeline, providing users with the functions of structure alignment, sequence alignment, 3D-View and risk report, to help users predict the severity of aGVHD of mismatch HLA donors before transplantation.
- In view of the key characteristics that affect the immunogenicity of mutated peptides, HLA3D established an antigenic peptide prediction pipeline to provide users with a series of applications, such as mutation prediction, peptide prediction, immunogenicity assessment and docking simulation, to help users complete the prediction and analysis of immunogenic peptides.

Abbreviations

HLA, human leukocyte antigen; MHC, major histocompatibility complex; TCR, T cell receptor; PDB, Protein Data Bank; IEDB, immune epitope database and analysis resource; AFND, allele frequency net database; HLA CWD catalogue, The catalogue of common and well-documented (CWD) alleles of HLA; ASHI, The American Society for Histocompatibility and Immunogenetics; EFI, The European Federation for Immunogenetics; BLAST, basic local alignment and search tool; SMG, significantly mutated gene; GWAS, genome-wide association study; SNP, single nucleotide polymorphism; HBV, hepatitis B virus; RMSD, root mean square deviation; aGVHD, acute graft versus host disease; HSCT, hematopoietic stem cell transplantation; WT, wild type; Mut, mutant type; SB, strong binder; WB, weak binder.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

National Natural Science Foundation of China (31871322, 31900473); Fundamental Research Funds for the Central Universities (2242017K3DN23, 2242017 K41041). The funding bodies were not involved in the design of the study, collection, analysis or interpretation of data.

References

1. Matzaraki V, Kumar V, Wijmenga C, *et al.* The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol* 2017;**18**(1):76.
2. Capietto AH, Jhunjhunwala S, Delamarre L. Characterizing neoantigens for personalized cancer immunotherapy. *Curr Opin Immunol* 2017;**46**:58–65.
3. Pidala J, Wang T, Haagenson M, *et al.* Amino acid substitution at peptide-binding pockets of HLA class I molecules increases risk of severe acute GVHD and mortality. *Blood* 2013;**122**(22):3651–8.
4. Zhou F, Cao H, Zuo X, *et al.* Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat Genet* 2016;**48**(7):740–6.
5. Mei S, Li F, Xiang D, *et al.* Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Brief Bioinform* 2021;**22**:bbaa415. https://doi.org/10.1093/bib/bbaa415.
6. Mei S, Li F, Leier A, *et al.* A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2020;**21**:1119–35.
7. Menezes Teles e Oliveira D, Melo Santos de Serpa Brandão R, Claudio Demes da Mata Sousa L, *et al.* pHLA3D: an online database of predicted three-dimensional structures of HLA molecules. *Hum Immunol* 2019;**80**(10):834–41.
8. Schönbach C, Koh JL, Sheng X, *et al.* FIMM: a database of functional molecular immunology. *Nucleic Acids Res* 2000;**28**(1):222–4.
9. Schönbach C, Koh JL, Flower DR, *et al.* FIMM, a database of functional molecular immunology: update 2002. *Nucleic Acids Res* 2002;**30**(1):226–9.
10. Sinigaglia M, Antunes DA, Rigo MM, *et al.* CrossTope: a curated repository of 3D structures of immunogenic peptide: MHC complexes. *Database (Oxford)* 2013;**2013**:bat002.
11. Nunes E, Heslop H, Fernandez-Vina M, *et al.* Definitions of histocompatibility typing terms. *Blood* 2011;**118**(23):e180–3.
12. Sidney J, Peters B, Frahm N, *et al.* HLA class I supertypes: a revised and updated classification. *BMC Immunol* 2008;**9**(1):1.
13. Thomsen M, Lundegaard C, Buus S, *et al.* MHCcluster, a method for functional clustering of MHC molecules. *Immunogenetics* 2013;**65**(9):655–65.
14. Mack SJ, Cano P, Hollenbach JA, *et al.* Common and well-documented HLA alleles: 2012 update to the CWD catalogue. *Tissue Antigens* 2013;**81**(4):194–203.

15. Sanchez-Mazas A, Nunes JM, Middleton D, *et al.* Common and well-documented HLA alleles over all of Europe and within European sub-regions: a catalogue from the European Federation for Immunogenetics. *HLA* 2017;**89**(2):104–13.
16. He Y, Li J, Mao W, *et al.* HLA common and well-documented alleles in China. *HLA* 2018;**92**(4):199–205.
17. Burley SK, Bhikadiya C, Bi C, *et al.* RCSB protein data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acid Res* 2021;**49**(D1): D437–51.
18. Pontius J, Richelle J, Wodak SJ. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol* 1996;**264**(1):121–36.
19. von Grotthuss M, Pas J, Wyrwicz L, *et al.* Application of 3D-jury, GRDB, and Verify3D in fold recognition. *Proteins* 2003;**53**(Suppl 6): 418–23.
20. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci* 1993;**2**(9):1511–9.
21. Kong R, Wang F, Zhang J, *et al.* CoDockPP: a multistage approach for global and site-specific protein-protein docking. *J Chem Inf Model* 2019;**59**(8):3556–64.
22. Williams CJ, Headd JJ, Moriarty NW, *et al.* MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* 2018;**27**(1):293–315.
23. Kandoth C, McLellan MD, Vandin F, *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* 2013;**502**(7471):333–9.
24. Shukla SA, Rooney MS, Rajasagi M, *et al.* Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* 2015;**33**(11):1152–8.
25. Han H, Yuan F, Sun Y, *et al.* Three-dimensional structure discrepancy between HLA alleles for effective prediction of aGVHD severity and optimal selection of recipient-donor pairs: a proof-of-concept study. *Oncotarget* 2015;**6**(37):40337–59.
26. Heemskerk MB, Roelen DL, Dankers MK, *et al.* Allogeneic MHC class I molecules with numerous sequence differences do not elicit a CTL response. *Hum Immunol* 2005;**66**(9):969–76.
27. Kawase T, Morishima Y, Matsuo K, *et al.* High-risk HLA allele mismatch combinations responsible for severe acute graft-versus-host disease and implication for its molecular mechanism. *Blood* 2007;**110**(7):2235–41.
28. Kawase T, Matsuo K, Kashiwase K, *et al.* HLA mismatch combinations associated with decreased risk of relapse: implications for the molecular mechanism. *Blood* 2009;**113**(12):2851–8.
29. Marino SR, Lin S, Maiers M, *et al.* Identification by random forest method of HLA class I amino acid substitutions associated with lower survival at day 100 in unrelated donor hematopoietic cell transplantation. *Bone Marrow Transplant* 2012;**47**(2):217–26.
30. Fernandez-Viña MA, Wang T, Lee SJ, *et al.* Identification of a permissible HLA mismatch in hematopoietic stem cell transplantation. *Blood* 2014;**123**(8):1270–8.
31. Bacigalupo A. A closer look at permissive HLA mismatch. *Blood* 2013;**122**(22):3555–6.
32. Bjorkman PJ, Saper MA, Samraoui B, *et al.* The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 1987;**329**(6139):512–8.
33. Jurtz V, Paul S, Andreatta M, *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;**199**(9): 3360–8.
34. Wells DK, van Buuren MM, Dang KK, *et al.* Key parameters of tumour epitope immunogenicity revealed through a consortium approach improve Neoantigen prediction. *Cell* 2020;**183**(3):818–834.e13.
35. Rego N, Koes D. 3Dmol.Js: molecular visualization with WebGL. *Bioinformatics* 2015;**31**(8):1322–4.
36. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**(2): 195–202.
37. O'Donnell TJ, Rubinsteyn A, Bonsack M, *et al.* MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst* 2018;**7**(1):129–132.e4.
38. Tien MZ, Sydykova DK, Meyer AG, *et al.* PeptideBuilder: a simple python library to generate model peptides. *PeerJ* 2013;**1**:e80.
39. Bjorkman PJ, Saper MA, Samraoui B, *et al.* Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 1987;**329**(6139):506–12.
40. Tynan FE, Elhassen D, Purcell AW, *et al.* The immunogenicity of a viral cytotoxic T cell epitope is controlled by its MHC-bound conformation. *J Exp Med* 2005;**202**(9):1249–60.
41. Gu S, Lv L, Lin X, *et al.* Using structural analysis to explore the role of hepatitis B virus mutations in immune escape from liver cancer in Chinese, European and American populations. *J Biomol Struct Dyn* 2020;**40**:1586–96. https://doi.org/10.1080/07391102.2020.1830852.
42. Meyer D, C. Aguiar VR, Bitarello BD, *et al.* A genomic perspective on HLA evolution. *Immunogenetics* 2018;**70**(1):5–27.
43. Claas FH, Dankers MK, Oudshoorn M, *et al.* Differential immunogenicity of HLA mismatches in clinical transplantation. *Transpl Immunol* 2005;**14**(3–4):187–91.