Briefings in Bioinformatics, 2022, **23**, 1–15

https://doi.org/10.1093/bib/bbab496

Case Study

Immuno-informatics analysis predicts B and T cell consensus epitopes for designing peptide vaccine against SARS-CoV-2 with 99.82% global population coverage

Priyank Shukla, Preeti Pandey, Bodhayan Prasad, Tony Robinson, Rituraj Purohit, Leon G. D'Cruz, Murtaza M. Tambuwala, Ankur Mutreja, Jim Harkin, Taranjit Singh Rai, Elaine K. Murray, David S. Gibson and Anthony J. Bjourson

Corresponding author. Priyank Shukla, Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, C-TRIC Building, Altnagelvin Area Hospital, Glenshane Road, Derry/Londonderry, BT47 6SB, UK. Tel.: +442871675690; E-mail: p.shukla@ulster.ac.uk

Abstract

The current global pandemic due to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has taken a substantial number of lives across the world. Although few vaccines have been rolled-out, a number of vaccine candidates are still under clinical trials at various pharmaceutical companies and laboratories around the world. Considering the intrinsic nature of viruses in mutating and evolving over time, persistent efforts are needed to develop better vaccine candidates. In this study, various immuno-informatics tools and bioinformatics databases were deployed to derive consensus B-cell and T-cell epitope sequences of SARS-CoV-2 spike glycoprotein. This approach has identified four potential epitopes which have the capability to initiate both antibody and cell-mediated immune responses, are non-allergenic and do not trigger autoimmunity. These peptide sequences were also evaluated to show 99.82% of global population coverage based on the genotypic frequencies of HLA binding alleles for both MHC class-I and class-II and are unique for SARS-CoV-2 isolated from human as a host species. Epitope number 2 alone had a global population coverage of 98.2%. Therefore, we further validated binding and interaction of its constituent T-cell epitopes with their corresponding HLA proteins using molecular docking and molecular dynamics simulation experiments, followed by binding free energy calculations with molecular mechanics Poisson–Boltzmann surface area, essential dynamics analysis and free energy landscape analysis. The immuno-informatics pipeline described and the candidate epitopes discovered herein could have significant impact upon efforts to develop globally effective SARS-CoV-2 vaccines.

Keywords: immuno-informatics, bio-informatics, vaccine, peptide, SARS-CoV-2

Introduction

The rise of the current pandemic of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) started in December 2019 with the reports of severe pneumonia cases of unknown aetiology from the city of Wuhan by the Chinese Center for Disease Control (China CDC) [1]. Coronaviruses (CoV) are enveloped positive-stranded RNA-viruses [2] belonging to a family of zoonotic viruses

Priyank Shukla is a Lecturer in Stratified Medicine (Bioinformatics) at the Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, UK.

Preeti Pandey is a Postdoctoral Researcher at the Department of Cell and Molecular Pharmacology and Experimental Therapeutics, College of Medicine, Medical University of South Carolina, Charleston, USA.

Bodhayan Prasad is a PhD Researcher at the Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, UK.

Tony Robinson is a PhD Researcher at the School of Computing, Engineering and Intelligent Systems, Ulster University, Magee Campus, UK.

Rituraj Purohit is a Principal Scientist at the Structural Bioinformatics Lab, Division of Biotechnology, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, Himachal Pradesh, India.

Leon G D'Cruz is a Research Associate at the Respiratory Medicine Department and Clinical Trials Unit, Queen Alexandra Hospital, Portsmouth, UK.

Murtaza M Tambuwala is a Lecturer in Pharmacy at the School of Pharmacy and Pharmaceutical Sciences, Ulster University, Coleraine Campus, UK.

Ankur Mutreja is a Group Leader at the Department of Medicine, University of Cambridge, Cambridge, UK.

Jim Harkin is a Professor at the School of Computing, Engineering and Intelligent Systems, Ulster University, Magee Campus, UK.

Taranjit Singh Rai is a Lecturer in Cellular Aging at the Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, UK.

Elaine K Murray is a Lecturer in Stratified Medicine (Mental Health) at the Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, UK.

David S Gibson is a Senior Lecturer in Inflammatory Disease at the Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, UK.

Anthony J Bjourson is a Professor of Genomics at the Northern Ireland Centre for Stratified Medicine, Biomedical Sciences Research Institute, Ulster University, UK.

Received: July 11, 2021. **Revised:** October 3, 2021. **Accepted:** October 30, 2021

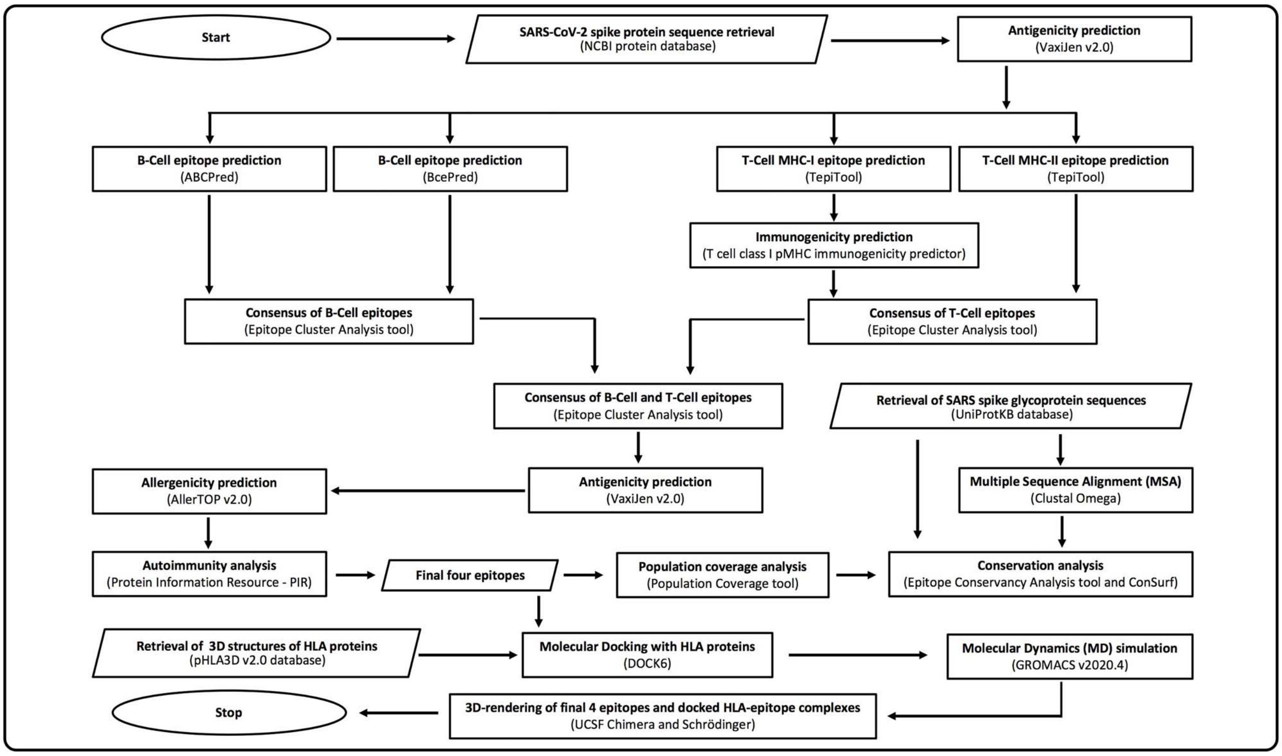
Start
SARS-CoV-2 spike protein sequence retrieval (NCBI protein database)
Antigenicity prediction (VaxiJen v2.0)
B-Cell epitope prediction (ABCPred)
B-Cell epitope prediction (BcePred)
T-Cell MHC-I epitope prediction (TepiTool)
T-Cell MHC-II epitope prediction (TepiTool)
Immunogenicity prediction (T cell class I pMHC immunogenicity predictor)
Consensus of B-Cell epitopes (Epitope Cluster Analysis tool)
Consensus of T-Cell epitopes (Epitope Cluster Analysis tool)
Consensus of B-Cell and T-Cell epitopes (Epitope Cluster Analysis tool)
Retrieval of SARS spike glycoprotein sequences (UniProtKB database)
Allergenicity prediction (AllerTOP v2.0)
Antigenicity prediction (VaxiJen v2.0)
Multiple Sequence Alignment (MSA) (Clustal Omega)
Autoimmunity analysis (Protein Information Resource - PIR)
Final four epitopes
Population coverage analysis (Population Coverage tool)
Conservation analysis (Epitope Conservancy Analysis tool and ConSurf)
Retrieval of 3D structures of HLA proteins (pHLA3D v2.0 database)
Molecular Docking with HLA proteins (DOCK6)
Molecular Dynamics (MD) simulation (GROMACS v2020.4)
Stop
3D-rendering of final 4 epitopes and docked HLA-epitope complexes (UCSF Chimera and Schrödinger)

Figure 1. Flowchart of immuno-informatics analysis pipeline *viz.* Peptide-based Vaccine Prediction Pipeline (PVPredPip). Oval shapes represent start/stop of the pipeline. Parallelogram boxes represent input/output. Rectangular boxes represent processing steps. Servers/tools/software and databases used are mentioned in parenthesis.

[3] which infect a variety of mammals including bats and humans [4]. They are implicated in previous outbreaks such as, Middle East Respiratory Syndrome (MERS-CoV) [5], Severe Acute Respiratory Syndrome (SARS-CoV) [4] and most recently SARS-CoV-2 [6], which has created an urgent need to develop diagnostics, therapeutics and vaccines against SARS-CoV-2. Although a few vaccines have been rolled-out in some major developed and developing countries, a number of vaccine candidates are still under clinical trials at various laboratories across the world in non-profit, public, academic, multinational pharmaceutical companies and other industries [7]. Considering the intrinsic nature of viruses in mutating and evolving over time (https://www.cdc.gov/coronavirus/2019-ncov/transmission/variant.html), persistent and timely efforts are needed to develop better vaccine candidates [8].

Recent efforts towards utilizing epitope prediction for designing a peptide-based vaccine against SARS-CoV-2 are either focused on only T-cell epitopes [9] or on HLA allele frequencies amongst specific populations such as, Japan [9] or China [10]. Some have focused on virus E protein [3], whereas others have investigated homology between SARS-CoV and SARS-CoV-2 to derive both common and unique epitopes [1].

In this study, we have deployed various immuno-informatics tools and bioinformatics databases to predict B-cell and T-cell consensus epitopes as peptide-based vaccine candidates for SARS-CoV-2, which show maximum population coverage across all continents and thus can be effective globally. Recognition of the antigenic epitope was carried out strategically in such a way that the selected epitopes are capable of generating both antibody and cell-mediated immune responses. The designed epitopes are also predicted to be non-allergenic and show no autoimmune response in humans. We further went on computationally validating the binding of constituent T-cell epitopes of the best consensus epitope with their corresponding HLA proteins using molecular docking and molecular dynamics (MD) simulation experiments, followed by binding free energy calculations with molecular mechanics Poisson–Boltzmann surface area (MM-PBSA), essential dynamics analysis and free energy landscape (FEL) analysis.

Materials and methods

A summary flowchart of the immuno-informatics analysis pipeline *viz.* Peptide-based Vaccine Prediction Pipeline (PVPredPip) is presented in Figure 1 and each step of the pipeline is described in detail in Supplementary File 1 (see Supplementary Data available online at http://bib.oxfordjournals.org/) with all the tools, software, servers, databases and specific parameters used for each of them during the analysis. All the in-house scripts used in this pipeline have been deposited in the following public repository https://github.com/ShuklaLab/PVPredPip.

Results

We have developed a robust immuno-informatics pipeline PVPredPip (Figure 1) for B-cell and T-cell consensus epitope prediction for SARS-CoV-2 by meticulously choosing and deploying various tools and bioinformatics databases. The results from each step of this pipeline are presented in the following sub-sections.

Antigenicity prediction of spike glycoprotein of SARS-CoV-2

The amino acid sequence of the spike glycoprotein of SARS-CoV-2, which was retrieved from NCBI Protein database (GenBank ID: QIC53213.1, NCBI Reference Sequence: YP_009724390.1) was 1273 amino acid residues long and was predicted to be antigenic by VaxiJen 2.0 [11]. The predicted antigenic nature of the retrieved spike glycoprotein sequence of SARS-CoV-2 endorses it to be a potential candidate to find B-cell and T-cell epitopes for designing a peptide-based vaccine.

Identification of B-cell epitopes

The B-cell epitopes predicted by the two servers ABCPred [12] and BcePred [13] produced variable results. Hence, a consensus of their results was developed with IEDB's Epitope Cluster Analysis tool [14], which yielded a total of twenty-four candidate B-cell epitope sequences (Supplementary Table 1, see Supplementary Data available online at http://bib.oxfordjournals.org/). These epitopes ranged in size from 14 to 35 residues and were distributed throughout all the domains of the spike glycoprotein, suggesting that further downstream analysis was required to refine potential B-cell epitope candidates.

Identification of T-cell epitopes

We chose a panel of the 27 most common HLA MHC-I binding A and B alleles and the 26 most common HLA MHC-II binding (DP, DQ and DR; A and B) alleles to ensure that predicted epitopes cover the majority of the global population. We also fixed the MHC-I epitopes length to 9-mer, as this length is suggested to be most preferable for binding majority of the ligands presented by HLA alleles [15]. In the case of MHC-II epitopes, length was fixed to 15-mer as recommended by the TepiTool. We further went on first refining the list of MHC-I epitopes by IEDB's Class-I Immunogenicity predictor [16] which groups residues based on their physico-chemical properties and uses them as a feature for immunogenicity prediction. Following this, we built the consensus of MHC-I and MHC-II overlapping epitope sequences with IEDB's Epitope Cluster Analysis tool [14], which yielded a total of 51 candidate T-cell epitope sequences (Supplementary Table 2, see Supplementary Data available online at http://bib.oxfordjournals.org/). These epitopes ranged in size from 15 to 27 residues and were distributed throughout all the domains of the spike glycoprotein, suggesting that further downstream analysis was required to refine the list of potential T-cell epitope candidates.

Final B-cell and T-cell consensus epitopes

Using the IEDB's Epitope Cluster Analysis tool [14], we identified 11 clusters, which were based on overlapping sequences of B-cell and T-cell epitopes. The rationale behind building these consensus sequences was to ensure that the selected epitopes are capable of generating both humoral and cytotoxic immune response. We went on to further rigorously refine these 11 epitopes by conducting a battery of tests: (i) antigenicity test with VaxiJen 2.0 [11], (ii) allergenicity test with AllerTOP 2.0 [17] and (iii) auto-immunity test with Protein Information Resource's peptide search service [18], which finally yielded four B-cell and T-cell consensus epitopes (Table 1). These epitopes range in size from 18 to 39 residues and are present in the S1 region of the SARS-CoV-2 spike glycoprotein (Figure 2). More specifically, epitope number 1 is present in the C-terminal domain 2 (CTD2) and epitopes numbered 2–4 are present in the N-terminal domain (NTD) (Figure 2).

Population coverage of final B-cell and T-cell consensus epitopes

Global population coverage for the set of final four epitopes was computed to be 99.82% by IEDB's Population Coverage tool (Table 2). This computation is based on genotypic frequencies of MHC-I and MHC-II HLA binding alleles of each epitope presented in Table 1. When analysed individually, epitope number 2 in particular showed high coverage for all the continent-area-specific populations including overall world population. Epitope number 1 on the other hand had not only the least overall world population coverage of only 19.83% but also very low coverage for each continent-area-specific populations. All four epitopes have less coverage ranging between 0.00 and 59.15% in the South African population (Table 2), which has also been previously demonstrated for other vaccines [20], and more recently for SARS-CoV-2 vaccines [21–24]. Epitope number 3 showed a high degree of variability in its population coverage ranging between 17.10 and 77.25% across all the continents.

Conservation of final B-cell and T-cell consensus epitopes

With this analysis, we sought to determine if the final four epitopes are unique for SARS-CoV-2 or can also be used against SARS-CoV. We also wanted to confirm if they are conserved across all SARS spike glycoprotein sequences which have been isolated from different host species (*viz.* human, bats, civet and bovine). A high degree of protein sequence similarity has been found between SARS-CoV-2 and SARS-CoV, but not between SARS-CoV-2 and MERS [1]. Therefore, we excluded MERS spike glycoprotein sequences from this analysis. We performed Multiple Sequence Alignment (MSA) of 144 non-redundant SARS spike glycoprotein sequences with EMBL-EBI's Clustal Omega server. These sequences were annotated in UniProt database as either belonging to SARS-CoV or SARS-CoV-2 and originated from either human, bat, civet

Table 1. Final four B-cell and T-cell consensus epitope sequences. Their start and end residue positions on the Spike Glycoprotein sequence of SARS Coronavirus-2 (GenBank ID: QIC53213.1, NCBI Reference Sequence: YP_009724390.1), length and the corresponding HLA binding alleles of T-cell epitopes are presented here. For more details of constituent B-cell and T-cell epitopes and the specific HLA binding alleles of each constituent T-cell epitopes, please see Supplementary Table 3, see Supplementary Data available online at http://bib.oxfordjournals.org/

Epitope No.	B-cell and T-Cell consensus epitope sequence	Residue positions	Length	HLA Class-I binding alleles of constituent T-cell epitopes	HLA Class-II binding alleles of constituent T-cell epitopes
1	PVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEH	621–655	35	HLA-A*02:06	HLA-DRB1*07:01
2	LHRSYLTPGDSSSGWTAGAAAYYVGYLQPR	244–273	30	HLA-A*01:01, HLA-A*26:01, HLA-A*30:02, HLA-A*68:02	HLA-DPA1*01:03, HLA-DPB1*02:01, HLA-DQA1*01:01, HLA-DQA1*01:02, HLA-DQA1*04:01, HLA-DQA1*05:01, HLA-DQB1*03:01, HLA-DQB1*04:02, HLA-DQB1*05:01, HLA-DQB1*06:02, HLA-DRB1*09:01
3	VLPFNDGVYFASTEKSNI	83–100	18	HLA-A*03:01, HLA-A*11:01	HLA-DPA1*02:01, HLA-DPB1*05:01, HLA-DRB1*04:01
4	WMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLR	152–190	39	HLA-A*01:01, HLA-A*30:02, HLA-B*15:01, HLA-B*35:01	HLA-DPA1*01:03, HLA-DPB1*02:01, HLA-DRB1*04:05, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*15:01, HLA-DRB3*02:02

or bovine species. The MSA results were further analysed by ConSurf which graded most of the residues as variable regions, i.e. least conserved (Table 3). Interestingly, epitope number 2, which showed maximum population coverage (Table 2), showed least conservation with IEDB's Population Coverage tool, such that 79% of 144 non-redundant sequences of SARS spike glycoprotein showed less than 30% identify with the epitope sequence (Table 3). Epitope number 4, which was second-best in terms of population coverage (Table 2), also showed poor conservation, where 94% of 144 non-redundant sequences of SARS spike glycoprotein showed less than 40% identity with the epitope sequence (Table 3). Epitopes numbered 1 and 3 were better conserved with 88–98% of 144 non-redundant sequences of SARS spike glycoprotein showing 70% identity with the two epitope sequences (Table 3), but their population coverage was less than the other two epitopes (Table 2). These results infer that the four epitopes, particularly epitopes 2 and 4 with maximum human population coverage, are unique for SARS-CoV-2 found in humans. These results also confirm that even though SARS-CoV-2 and SARS-CoV protein sequences are similar, the amino acid sequence of spike glycoprotein differs substantially when analysing different strains based on host origin, which warrants host-origin-specific vaccine development.

Molecular docking of potential T-cell epitopes

For a multi-epitope vaccine to induce protective immunity, it should satisfy at least three criteria: (i) the peptides must match with the epitope naturally presented to the immune cells during infection, (ii) elicit an adequate immune response and (iii) must have an optimal population coverage. In our earlier analysis, we found epitope number 2 to satisfy all these three criteria. But, since the HLA molecules are extremely polymorphic in nature (more than 600 allelic forms, encoding diverse amino acid sequence), with the sequence diversity mostly concentrated in the peptide binding region (antigen binding groove between the two helices of MHC molecules), the binding affinity of the epitope towards different HLA molecules may differ. Therefore, to understand the binding affinity of the predicted epitope towards different HLA molecules and subsequent interaction between them, we performed molecular docking studies. The docking scores of five constituent T-cell epitopes numbered 2.2.1–2.2.5 of consensus epitope number 2 (Supplementary Table 3, see Supplementary Data available online at http://bib.oxfordjournals.org/) towards different HLA molecules are listed in Table 4. The docking results indicate that these five constituent T-cell epitopes possess good binding affinity towards different HLA molecules which is in agreement with our earlier results. The binding affinity for the MHC-I molecules lies in the range of −97.67 to −84.02 kcal/mol, while for MHC-II molecules, it is in the range of −136.36 to −77.86 kcal/mol. In both cases, the major contributions are coming from the *van der* Waals interactions as hydrophobic residues dominate in the binding pocket as well as in epitopes (Table 4). The best binding affinity for MHC-I molecules was observed for epitope number 2.2.2 against HLA-A*30:02 (−97.67 kcal/mol) and epitope number 2.2.3 against HLA-A*01:01 (−94.93 kcal/mol) and HLA-A*26:01 (−90.97 kcal/mol). And, for the

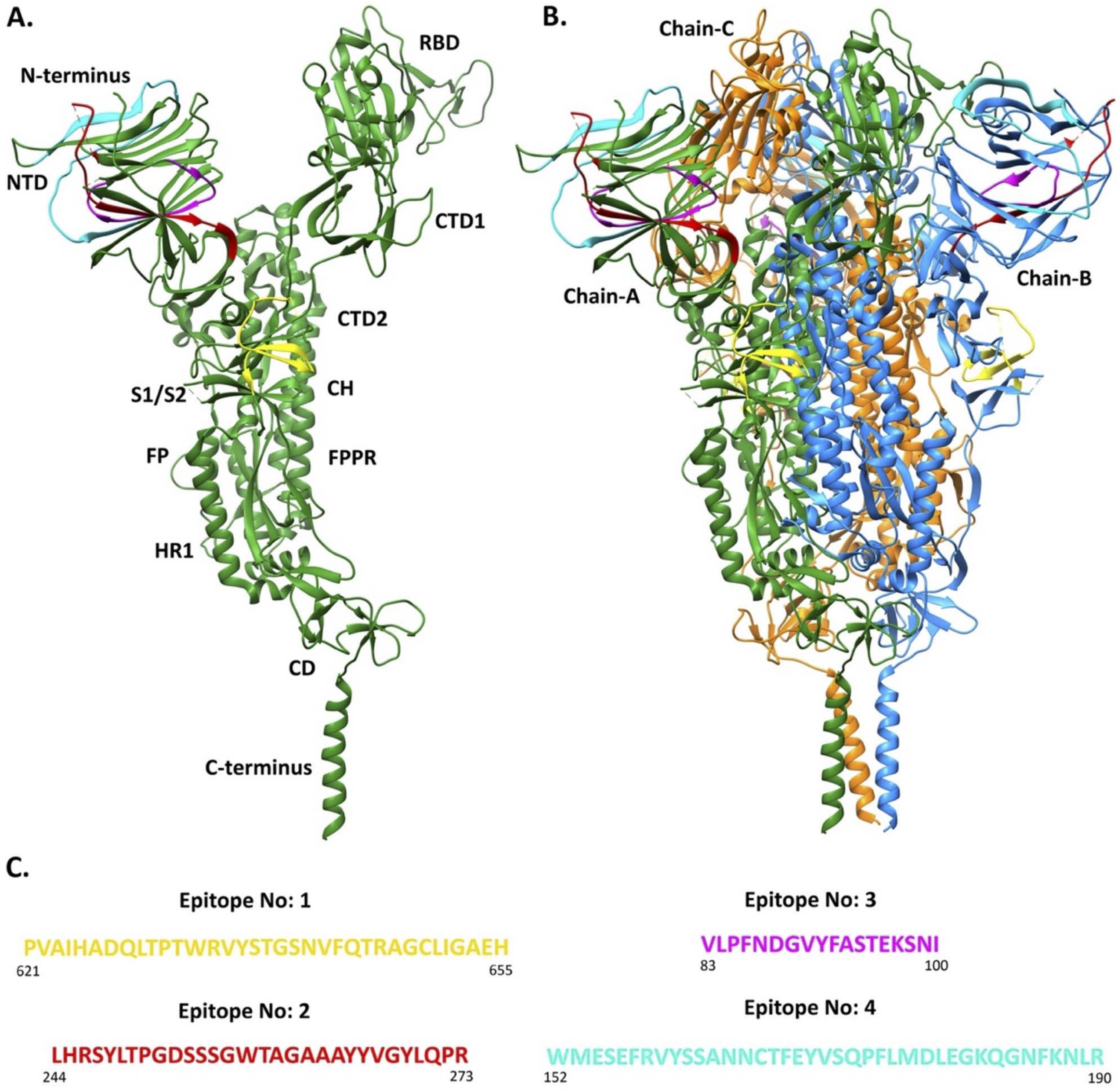

Figure 2. Visualization of final four epitopes in SARS-CoV-2 spike protein (PDB ID: 6XR8). Regions of final four B-cell and T-cell consensus epitope sequences are highlighted in both (**A**) monomer and (**B**) trimer of the spike protein, wherein chain-A is highlighted in green, chain-B in blue and chain-C in orange. Epitope 1, highlighted in yellow, is present in C-terminal domain 2 (CTD2) and epitopes 2–4, highlighted in red, magenta and cyan, respectively, are in N-terminal domain (NTD). (**C**) Sequences of final four B-cell and T-cell consensus epitope. RBD = receptor binding domain, CTD1 = C-terminal domain 1, S1/S2 = S1/S2 cleavage site, CH = central helix region, CD = connector domain, HR1 = heptad repeat 1, FP = fusion peptide and FPPR = fusion peptide proximal region, are indicated in Figure 2A. SARS-CoV-2 spike protein's domain information has been derived from Figure 1A of Cai *et al*. study [2]. 3D-rendering was performed using UCSF Chimera [19].

MHC-II molecules, epitope numbers 2.2.4 and 2.2.1 showed a promising binding affinity towards HLA-DPA1*01:03/HLA-DPB1*02:01, HLA-DQA1*05:01/HLA-DQB1*03:01 and HLA-DQA1*04:01/HLA-DQB1*04:02, respectively. It is also to be noted here that the two epitopes, i.e. 2.2.2 and 2.2.3, have eight residues in common and essentially differ by a single residue window shift leading to a difference of one residue flanking on each side (Supplementary Table 3, see Supplementary Data available online at http://bib.oxfordjournals.org/), but they possess a significantly different affinity towards HLA-A*30:02 (dock score differs by ∼10 kcal/mol). Quite similar is the case for epitope numbers 2.2.1 and 2.2.4 against HLA-DQA1*05:01/HLA-DQB1*03:01 (MHC-II). These results clearly indicate that few residues flanking the common motif in a novel designed peptide can significantly influence the overall binding preference of a peptide. Interestingly, we also observed that the same epitope can have a very different binding affinity towards different HLA molecules. For example, epitope number 2.2.4 possesses a very good binding affinity towards HLA-DPA1*01:03/HLA-DPB1*02:01, while the affinity differs by ∼52 kcal/mol for HLA-DQA1*01:01/HLA-DQB1*05:01 (Table 4). This is expected as the binding pockets of the HLA molecules are highly diverse with respect to the amino acid sequence, and therefore, the antigen binding

Table 2. Population coverage analysis of final four B-cell and T-cell consensus epitope sequences. Values in each cell were computed by IEDB's Population Coverage Tool and they represent the percentage of population covered by the epitope based on the genotypic frequencies of HLA binding alleles (presented in Table 1). NA = data not available

Epitope No.	World	East Asia	North-East Asia	South Asia	South-East Asia	South-West Asia	Europe	East Africa	West Africa	Central Africa	North Africa	South Africa	West Indies	North America	Central America	South America	Oceania
1	19.83	21.66	14.22	33.97	12.29	12.60	23.96	9.22	9.19	13.23	25.63	0.00	18.09	20.78	11.22	9.76	3.54
2	98.22	91.24	97.44	98.18	92.55	93.43	99.79	99.94	99.86	99.46	95.67	42.83	98.18	99.98	91.42	98.87	98.82
3	65.45	67.63	77.25	60.62	72.03	24.33	55.89	42.45	75.74	47.08	17.10	18.54	25.83	46.01	NA	48.25	81.74
4	88.53	81.45	86.38	92.55	47.83	52.60	98.72	76.45	76.06	79.90	65.94	32.48	43.90	99.71	33.14	83.73	87.62
Set of 4 epitopes	99.82	99.33	99.77	99.89	98.83	96.39	99.99	99.98	99.99	99.86	98.17	59.15	99.23	100.00	94.90	99.77	99.91

Table 3. Conservation analysis of final four B-cell and T-cell consensus epitope sequences. Values in each cell are computed by IEDB's Conservation Analysis Tool and it represents percentage of 114 non-redundant SARS Spike Glycoprotein sequences retrieved from UniProtKB database which matched with epitope sequence with the sequence similarity (percentage) described in the respective column field. Multiple sequence alignment (MSA) of 114 unique SARS Spike Glycoprotein sequences was performed with Clustal Omega and the results were analysed with ConSurf which grades conservation of amino acids in a scale of 1–9. Amino acids in not-bold font are variable regions (ConSurf score grade 1–3), in bold-only font are average conserved regions (ConSurf score grade 4–6) and in underlined-bold font are highly conserved regions (ConSurf score grade 7–9)

Epitope No.	B-cell and T-cell consensus epitope sequence	100%	90%	80%	70%	60%	50%	40%	30%	0–20%
1	PVA**I**H**A**D**QLTP**T**WR**V**YS**T**G**S**N**V**FQTRAGCLIGAEH**	1.75	6.14	77.19	98.25	98.25	98.25	98.25	99.12	100.00
2	LHRS**YLTP**GDSSSGWTAGAA**AY**Y**VG**Y**LQ**PR	1.75	1.75	1.75	1.75	1.75	1.75	1.75	78.95	100.00
3	VLP**F**N**DG**V**YFASTEKSN**I	1.75	1.75	2.63	87.72	98.25	99.12	99.12	99.12	100.00
4	**WM**ESEFRVYSS**A**N**NCT**FEYVSQP**F**LM**D**LEGKQ**GNF**KN**LR**	1.75	1.75	1.75	1.75	1.75	3.51	93.86	95.61	100.00

Table 4. The estimated binding free energy of HLA-epitope complexes after performing Molecular Docking with DOCK6. Grid$_{VDW}$ and ΔE_{EEL} are the interaction energies due to electrostatics and van der Waals interactions, and Dock Score is the binding free energy

Epitope No.	HLA alleles	Grid$_{vdw}$ (kcal/mol)	Grid$_{EEL}$ (kcal/mol)	Dock score (kcal/mol)
	MHC-I			
2.2.2	HLA-A*30:02	−77.11	−20.55	−97.67
2.2.3	HLA-A*01:01	−85.50	−9.43	−94.93
	HLA-A*26:01	−69.43	−21.54	−90.97
	HLA-A*30:02	−73.88	−10.43	−84.30
2.2.5	HLA-A*68:02	−77.53	−6.50	−84.02
	MHC-II			
2.2.1	HLA-DQA1*01:02/HLA-DQB1*06:02	−78.47	0.62	−77.86
	HLA-DQA1*04:01/HLA-DQB1*04:02	−104.62	−12.89	−117.51
	HLA-DQA1*05:01/HLA-DQB1*03:01	−104.93	−17.13	−122.06
	HLA-DRB1*09:01	−95.74	−12.88	−108.62
2.2.4	HLA-DPA1*01:03/HLA-DPB1*02:01	−114.10	−22.26	−136.36
	HLA-DQA1*01:01/HLA-DQB1*05:01	−82.91	−0.63	−83.54
	HLA-DQA1*05:01/HLA-DQB1*03:01	−102.13	−10.39	−112.52

Table 5. The estimated binding free energy of HLA-epitope complexes using MM-PBSA after performing MD simulations. ΔE_{VDW} and ΔE_{EEL} are the changes in interaction energy due to electrostatics and van der Waals interactions, ΔG_{POL} and ΔG_{NP} are the changes in the polar and non-polar part of the solvation free energy and ΔG_{bind} is the change in the binding free energy

Epitope No.	HLA alleles	ΔE_{VDW}(kcal/mol)	ΔE_{EEL} (kcal/mol)	ΔG_{POL} (kcal/mol)	ΔG_{NP} (kcal/mol)	ΔG_{bind} (kcal/mol)
	MHC-I					
2.2.2	HLA-A*30:02	−65.34 ± 5.29	−125.39 ± 18.93	141.01 ± 17.72	−7.85 ± 0.4	−57.56 ± 4.22
2.2.3	HLA-A*01:01	−45.8 ± 5.07	−57.52 ± 20.79	73.34 ± 18.80	−6.53 ± 0.37	−36.51 ± 4.57
	HLA-A*26:01	−43.82 ± 6.69	−50.71 ± 15.20	56.43 ± 14.50	−6.29 ± 0.81	−44.39 ± 6.47
	HLA-A*30:02	−80.38 ± 6.21	−138.93 ± 10.81	154.10 ± 9.18	−9.58 ± 0.21	−74.80 ± 5.02
2.2.5	HLA-A*68:02	−62.65 ± 4.22	−64.73 ± 17.62	83.47 ± 14.83	−7.48 ± 0.24	−51.39 ± 4.36
	MHC-II					
2.2.1	HLA-DQA1*01:02/HLA-DQB1*06:02	−80.76 ± 5.15	−21.02 ± 18.03	42.68 ± 17.13	−9.19 ± 0.47	−68.28 ± 5.05
	HLA-DQA1*04:01/HLA-DQB1*04:02	−95.32 ± 7.53	−4.07 ± 16.18	29.75 ± 17.70	−11.08 ± 0.94	−80.72 ± 6.05
	HLA-DQA1*05:01/HLA-DQB1*03:01	−66.17 ± 6.14	−11.64 ± 14.42	32.88 ± 13.91	−8.22 ± 0.76	−53.16 ± 5.35
	HLA-DRB1*09:01	−90.06 ± 4.98	3.23 ± 9.13	15.09 ± 9.47	−10.08 ± 0.34	−81.81 ± 6.06
2.2.4	HLA-DPA1*01:03/HLA-DPB1*02:01	−103.46 ± 4.84	−158.79 ± 18.29	181.21 ± 17.09	−11.38 ± 0.39	−92.41 ± 5.37
	HLA-DQA1*01:01/HLA-DQB1*05:01	−68.90 ± 6.39	−53.92 ± 17.25	71.80 ± 16.31	−8.50 ± 0.56	−59.52 ± 5.53
	HLA-DQA1*05:01/HLA-DQB1*03:01	−61.47 ± 5.17	−91.13 ± 14.50	102.30 ± 13.98	−8.49 ± 0.67	−58.78 ± 3.86

groove has a preference towards certain amino acids to assure a stable interaction between the MHC molecule and the peptide. Overall, our results indicate that the five constituent T-cell epitopes of consensus epitope number 2 effectively bind to different HLA molecules, which is consistent with our earlier results. We also note the preference of MHC molecules towards certain peptides and *vice-versa*, which is in-line with previous studies [16, 25].

Conformational stability of HLA-epitope complexes

In above analysis, we show that the five constituent T-cell epitopes of consensus epitope number 2 have good binding affinity towards HLA molecules, but given the approximations made in molecular docking such as, the 'target receptor' is considered rigid, absence of water molecules, etc., it does not guarantee the correct binding mode for a ligand and therefore, to further confirm the results of molecular docking, we performed binding free energy calculations using MM-PBSA in combination with MD simulation.

First, we tried to understand the conformational stability of the HLA-epitope complexes in terms of three structural order parameters: (i) root-mean square deviation (RMSD), (ii) radius of gyration (Rg) and (iii) solvent-accessible surface area (SASA), and the results are shown in Supplementary Figure 1 (see Supplementary Data available online at http://bib.oxfordjournals.org/). On comparing the Cα RMSD of the HLA proteins, it can be clearly seen that all the systems attained equilibrium in the first 40 ns and remained stable thereafter, except for HLA-A*68:02-epitope-2.2.5 (orange line in the top-left panel of Supplementary Figure 1) and HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1 (orange line in the top-right panel of Supplementary Figure 1, see Supplementary Data available online at http://bib.oxfordjournals.org/) complexes. The RMSD plot of HLA-A*68:02-epitope-2.2.5 showed slightly larger deviations in the initial 60 ns. The RMSD rose up to 3.7 Å around 43 ns, but thereafter, a gradual drop is seen until 65 ns, after which a stable trajectory is seen till 100 ns. Similar behaviour was observed for the HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1 complex. The RMSD of the HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.4 is slightly higher than the other HLA-epitope complexes and fluctuates around an average value of ∼2.5 Å (blue line in the top-right panel of

Supplementary Figure 1, see Supplementary Data available online at http://bib.oxfordjournals.org/). The initial fluctuation in the RMSD indicates the spatial adjustment of the epitope in the binding site of the HLA molecules. Overall, the results suggest that all the systems achieved a steady equilibrium after 65 ns, suggesting an equilibrated and stabilized HLA-epitope interaction.

To further understand the stability of the HLA-epitope complexes, we computed the radius of gyration (Rg) and SASA of the HLA proteins, as a measure of the compactness of the protein structure upon epitope binding. Except for HLA-A*68:02-epitope-2.2.5 (orange line in the center-left panel of Supplementary Figure 1, see Supplementary Data available online at http://bib.oxfordjournals.org/) which showed major fluctuations in the initial 60 ns (Rg values ranges between 23.3 and 24.5 Å), all the complexes showed fairly stable Rg values since the beginning of the simulations till 100 ns. Interestingly, we observed that the SASA values for all the HLA-epitope complexes remained stable throughout the length of the simulation time. All these results indicate a stable conformational dynamics of the HLA-epitope interaction and substantiate our previous results.

MM-PBSA re-scoring and molecular interactions stabilizing HLA-epitope complexes

To further confirm the results of molecular docking and understand the molecular interactions stabilizing the HLA-epitope complex, binding free energy calculation using MM-PBSA was performed wherein both receptor flexibility as well as effect of water molecules are taken care of, which is usually ignored in molecular docking. The calculated binding free energies for the HLA-epitope complexes averaged over the snapshots extracted from the last 20 ns MD trajectories are listed in Table 5. We note that the exact order of the ranking of binding affinities of the epitopes towards both MHC-I and MHC-II protein molecules have changed, but consistent with the results of molecular docking, all the epitopes bind favourably to the HLA molecules. The binding free energy ranges from −74.80 ± 5.02 (HLA-A*30:02-epitope-2.2.3) to −36.51 ± 4.57 kcal/mol (HLA-A*01:01-epitope-2.2.3) for MHC-I molecules and −92.41 ± 5.37 (HLA-DPA1*01:03/HLA-DPB1*02:01-epitope-2.2.4) to −53.16 ± 5.35 kcal/mol (HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1) for MHC-II molecules. Consistent with molecular docking results, the best binding affinity was observed for epitope number 2.2.4 against HLA-DPA1*01:03/HLA-DPB1*02:01 (−92.41 ± 5.37 kcal/mol) for MHC-II molecules; however, the results changed for epitope number 2.2.1 against HLA-DQA1*01:02/HLA-DQB1*06:02 and HLA-DQA1*05:01/HLA-DQB1*03:01. The epitope number 2.2.1 now possesses the least binding affinity towards HLA-DQA1*05:01/HLA-DQB1*03:01 not HLA-DQA1*01:02/HLA-DQB1*06:02. The results also changed for the three constituent T-cell epitopes of consensus epitope number 2 against class I MHC molecules. In addition to this, we also observed a disparity between the contribution of the gas phase interactions, i.e. both *van der* Waals interaction as well as the electrostatic interactions between molecular docking and MM-PBSA results. Both the components now equally contribute to the stability of the MHC-I-epitope complexes, except there are cases, where electrostatic component overpowers the *van der* Waals interactions, as is the case with HLA-A*30:02-epitope-2.2.2 and HLA-A*30:02-epitope-2.2.3 complexes (Table 5). In the case of MHC-II-epitope complexes, *van der* Waals interaction dominates in the majority of the cases with a few exceptions (Table 5). The increase in the contribution of the electrostatic component points towards a spatial adjustment that favours the polar interactions indicating formation of mainchain–mainchain, mainchain–sidechain and sidechain–sidechain polar interactions (Tables 4 and 5**,** Figures 3 and 4). The other factor that determines the binding affinity is the solvation free energy ($\Delta G_{POL} + \Delta G_{NP}$), which is neglected in molecular docking, but was found to be always positive for the HLA-epitope complexes in MM-PBSA calculations. The polar part of the solvation free energy has opposing effects much larger in magnitude than the non-polar part of solvation free energy suggesting that the solvation free energy opposes the formation of the HLA-epitope complex. But, since in all the cases, the gas phase interaction energy ($\Delta E_{VDW} + \Delta E_{EEL}$) combats the opposing effects of solvation free energy, we observe stable HLA-epitope complexes. While we did observe some discrepancies between the molecular docking results and MM-PBSA calculations, overall the results indicate that the epitopes possess good binding affinity towards HLA molecules, usually overwhelmed by *van der* Waals interaction in case of MHC-II molecules, while both *van der* Waals and electrostatic interactions contribute fairly to the stability of MHC-I-epitope complexes. The molecular interactions stabilizing the HLA-epitope complexes are shown in Figures 3 and 4, which also indicate towards the *van der* Waals interaction to be one of the major stabilizing forces.

In addition, we also calculated the average number of hydrogen bonds formed between the HLA and epitope molecules as hydrogen bonds play a critical role in stabilizing the protein-ligand interaction. The values are listed in Supplementary Table 4 (see Supplementary Data available online at http://bib.oxfordjournals.org/). The average number of hydrogen bonds formed between the HLA molecule's antigen presenting pocket and epitopes ranges from 3 ± 1 (HLA-A*01:01-epitope-2.2.3) to 9 ± 2 (HLA-A*30:02-epitope-2.2.3) for MHC-I molecules and from 5 ± 1 (HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.4) to 11 ± 2 (HLA-DRB1*09:01-epitope-2.2.1) for MHC-II molecules, which also indicate towards the stability of the epitopes in the antigen presenting groove of the MHC molecules.

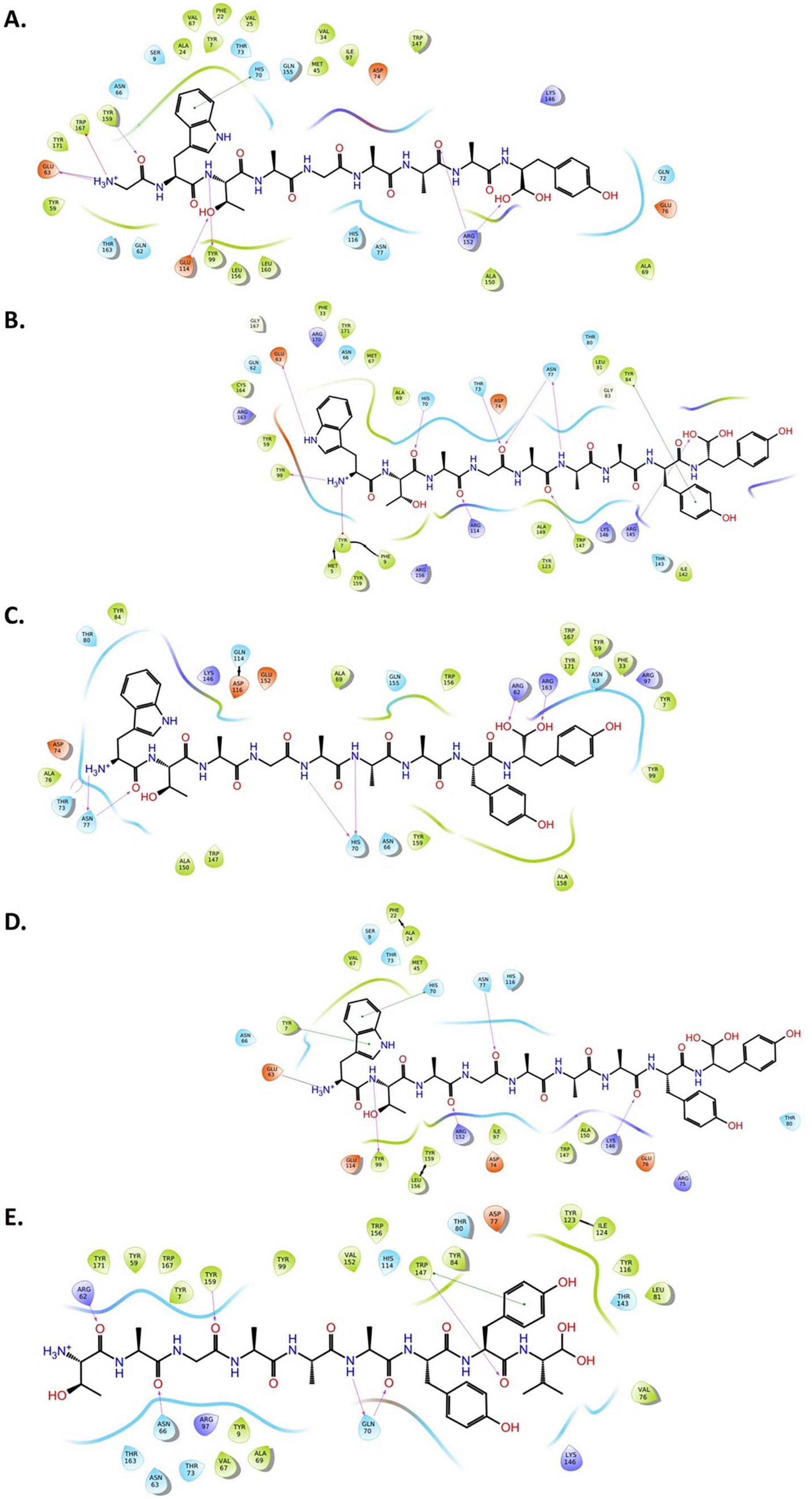

Figure 3. Molecular interactions of MHC-I-epitope complexes: (**A**) HLA-A*30:02-epitope-2.2.2, (**B**) HLA-A*01:01-epitope-2.2.3, (**C**) HLA-A*26:01-epitope-2.2.3, (**D**) HLA-A*30:02-epitope-2.2.3 and (**E**) HLA-A*68:02-epitope-2.2.5.

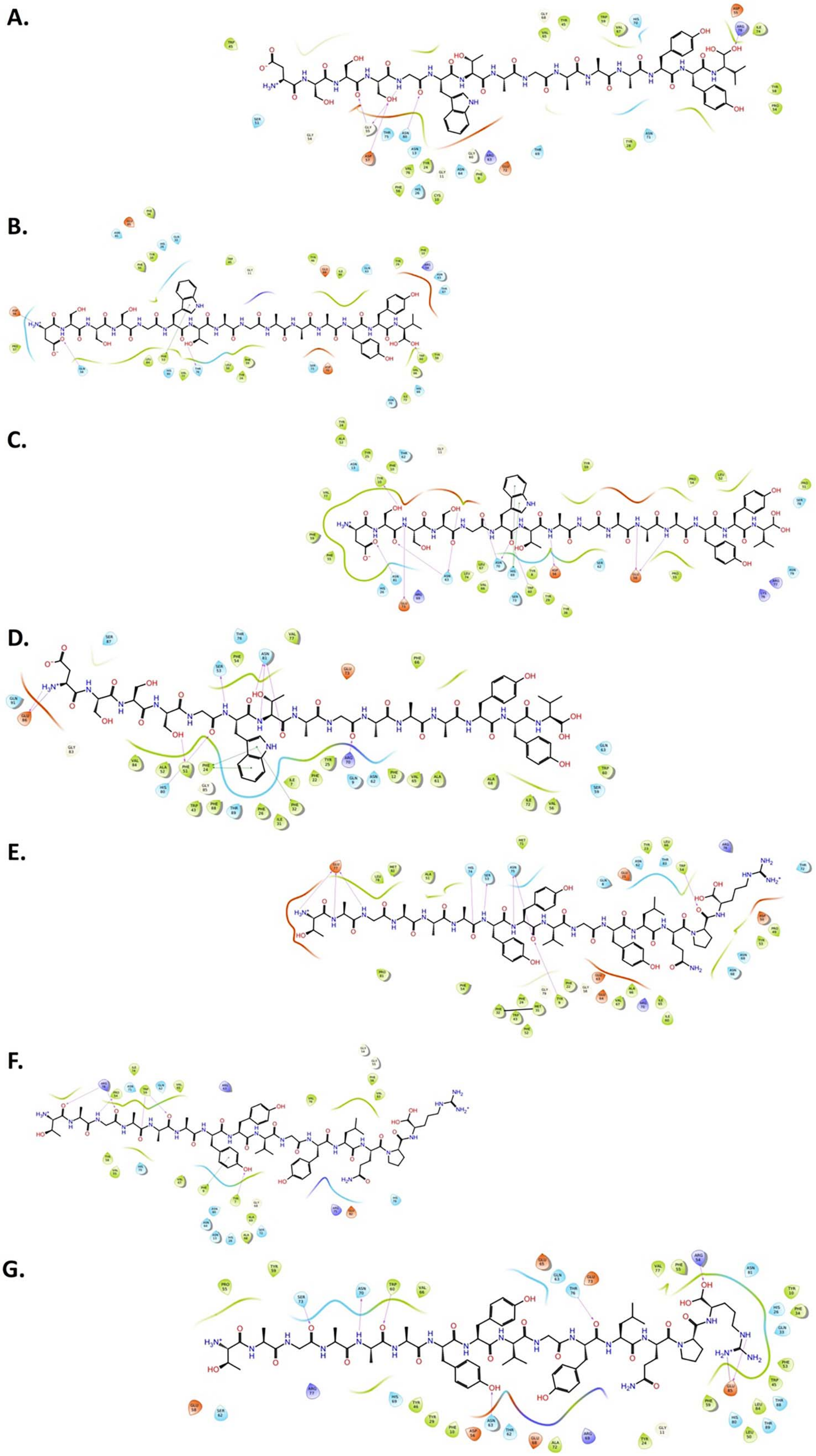

Figure 4. Molecular interactions of MHC-II-epitope complexes: (**A**) HLA-DQA1*01:02/HLA-DQB1*06:02-epitope-2.2.1, (**B**) HLA-DQA1*04:01/HLA-DQB1*04:02-epitope-2.2.1, (**C**) HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1, (**D**) HLA-DRB1*09:01-epitope-2.2.1, (**E**) HLA-DPA1*01:03/HLA-DPB1*02:01-epitope-2.2.4, (**F**) HLA-DQA1*01:01/HLA-DQB1*05:01-epitope-2.2.4 and (**G**) HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.4.

Structural motions and conformational redistribution of HLA-epitope complexes

To further examine the dominant motions and conformational sampling of the HLA protein molecules upon binding of the epitopes, principal component analysis was performed. It is generally assumed that the first 10 principal components account for more than 90% motion of the protein responsible for their function. Therefore, we calculated the first 10 principal components, and the conformational sampling of the HLA-epitope complexes in the essential subspace illustrating the global motions along PC1 and PC2 are shown in Supplementary Figure 2 (see Supplementary Data available online at http://bib.oxfordjournals.org/). It is apparent from the figure that majority of the HLA-epitope complexes show a global collective dynamics except for HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1 (Supplementary Figure 2H, see Supplementary Data available online at http://bib.oxfordjournals.org/), HLA-DQA1*01:01/HLA-DQB1*05:01–2.2.4 (Supplementary Figure 2K, see Supplementary Data available online at http://bib.oxfordjournals.org/) and HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.4 (Supplementary Figure 2L, see Supplementary Data available online at http://bib.oxfordjournals.org/). The widespread conformational subspace indicates that the HLA molecules manoeuvre through a broad conformational space before achieving an equilibrated state. As stated earlier, in some cases (DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1 and DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.4), we do observe a smaller cluster of conformations, which indicates higher flexibility of the HLA molecules and is consistent with the RMSD plot (orange and blue lines in top-right panel of Supplementary Figure 1, see Supplementary Data available online at http://bib.oxfordjournals.org/).

To further understand the effect of epitope binding on the conformational redistribution and the energetics of the HLA molecules, FELs were determined as the function of the first two principal components, PC1 and PC2. The 3D and 2D FEL plots for the HLA-epitope complexes are shown in Figure 5 and Supplementary Figure 3 (see Supplementary Data available online at http://bib.oxfordjournals.org/), respectively. As can be seen from these figures, the FEL of the MHC-I molecules consists of a broad minima with multiple conical ends suggesting a widespread distribution of low-energy conformations. Similar is the case with three MHC-II molecules, i.e. HLADQA1*04:01/HLA-DQB1*04:02 (Figure 5G), HLA-DRB1*09:01 (Figure 5I) and HLA-DPA1*01:03/HLA-DPB1*02:01 (Figure 5J). However, in case of other MHC-II molecules such as, HLA-DQA1*05:01/HLA-DQB1*03:01 (Figure 5H) and HLA-DQA1*01:01/HLA-DQB1*05:01 (Figure 5K), we observe multiple minima separated by small energy barriers in a broad basin, which indicates that epitope binding induces selection of multiple conformations of the HLA molecules, but only one minima consists of low-energy conformation. Thus, both the essential dynamics and FEL analysis indicate, although different but stable binding stability of the HLA-epitope complexes.

Discussion and conclusions

While the world continues to suffer from the COVID-19 disease, the causal organism, i.e. SARS-CoV-2 that has jumped off or spilled over humans from an animal reservoir, continues to evolve over time like any other virus to incorporate mutations which increases its survival and infection rate. The COVID-19 outbreak, which was declared as pandemic by WHO within 3 months of the first report of the disease, perdures to remain, arguably, one of the biggest threat and mystery to mankind and also science, as still a lot of questions such as, how the virus infects, spreads, survives, mutates, etc. remains unclear. However, given the collective efforts made by the scientific community across the world who have been working indefatigably to understand these questions to come up with preventive measures as well as therapeutic agents has led to the development of some RNA and vector-based vaccines [26–28], which have been approved under Emergency Use Authorization.

These approved vaccines from Moderna, Pfizer-BioNTech and Oxford-AstraZeneca have shown 62–95% efficacy in phase-3 or phase-2/3 clinical trials [26–28]; however, they still possess several challenges including storage at ultra-low temperature, stability, scalability, high costs and allergic reactions. Synthetic vaccines based on peptides can minimize these challenges. Peptide-based vaccines have the ability to generate epitope-specific immune response and are more stable and easily accessible under normal storage conditions [29, 30]. Since they are chemically synthesized, they can be manufactured at large scale with low cost [30]. Unfortunately, none of the SARS-CoV-2 vaccines rolled-out to date in the market are peptide-based, as the majority of the focus of the scientific and industrial community has been on vector-based, whole pathogen, DNA, RNA and recombinant protein-based vaccines [7].

However, like any other vaccine development approaches, peptide vaccines also possess some limitations including reduced immunogenicity and enzymatic degradation. Such weakness could be improved by combining an adjuvant or particulate delivery carriers [31].

Given the advantages associated with peptide-based vaccines, in this study, we have tried to design B-cell and T-cell consensus epitopes (peptides) against SARS-CoV-2 by deploying various immuno-informatics tools and bioinformatics databases and have also confirmed the binding affinity of the designed epitopes against various HLA (MHC-I and II) proteins using molecular docking, MD simulation and binding free energy estimation with MM-PBSA. We have meticulously selected epitope prediction tools from a pool of software available in the immuno-informatics research literature based on their documented performance on the area under the

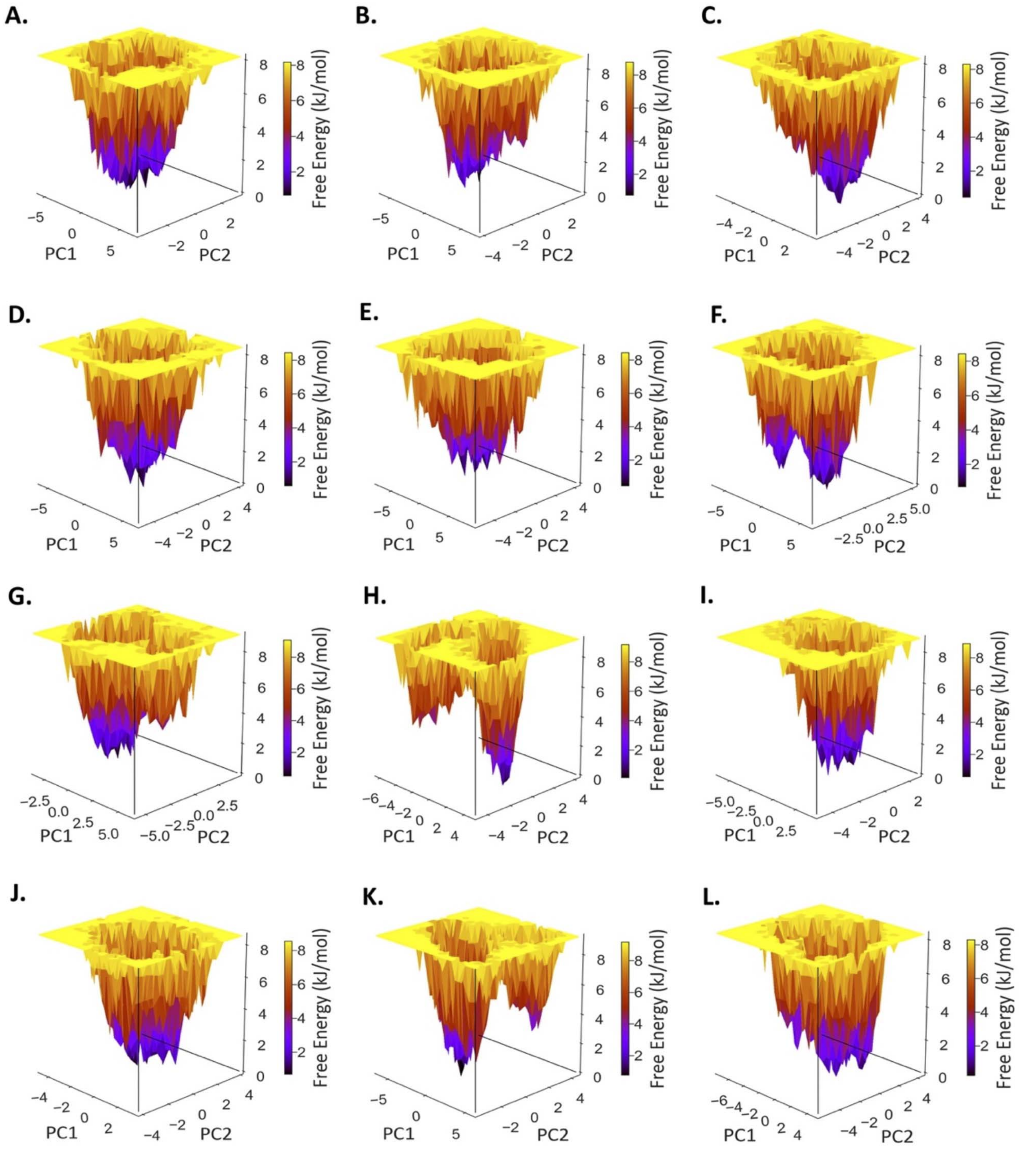

Figure 5. Free energy landscape (FEL) of HLA-epitope complexes in 3D space: PC1 and PC2 are the first and second principal components of the projection of the motion of the HLA-epitope complex in phase space. (**A**) HLA-A*30:02-epitope-2.2.2, (**B**) HLA-A*01:01-epitope-2.2.3, (**C**) HLA-A*26:01-epitope-2.2.3, (**D**) HLA-A*30:02-epitope-2.2.3, (**E**) HLA-A*68:02-epitope-2.2.5, (**F**) HLA-DQA1*01:02/HLA-DQB1*06:02-epitope-2.2.1, (**G**) HLA-DQA1*04:01/HLA-DQB1*04:02-epitope-2.2.1, (**H**) HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.1, (**I**) HLA-DRB1*09:01-epitope-2.2.1, (**J**) HLA-DPA1*01:03/HLA-DPB1*02:01-epitope-2.2.4, (**K**) HLA-DQA1*01:01/HLA-DQB1*05:01-epitope-2.2.4 and (**L**) HLA-DQA1*05:01/HLA-DQB1*03:01-epitope-2.2.4.

receiver operating characteristics curve (AUC) metric. Knowing the variability of results given by B-cell and T-cell epitope predictors and the recommendation of Moutaftsi *et al.*'s study [32], we ensured to build consensus of results by deploying multiple methods. To further refine our results and maximize our chances of identifying the best vaccine candidates, we also applied a battery of immunological properties' tests using well established state-of-the-art immuno-informatics and bioinformatics tools and databases (Figure 1 and Supplementary File 1, see Supplementary Data available online at http://bib.oxfordjournals.org/).

SARS-CoV-2 spike glycoprotein was predicted to be antigenic as per VaxiJen at the start of the analysis;

hence, epitope candidates were expected to be identified throughout the protein. Nevertheless, it was necessary to refine the analysis to identify the best candidates which fit essential properties of a good peptide-based vaccine candidate such as, immunogenicity, allergenicity, population coverage, etc.

Our best four B-cell and T-cell epitope candidates were found in S1 regions of SARS-CoV-2 spike protein, which aligns with the fact the most of the recent epitope prediction results against SARS-CoV-2 have been focused on S1 region [1, 9, 10]. However, our epitopes were concentrated around NTD and CTD2 rather than the receptor binding domain (RBD), which interacts with the human ACE2 receptor to facilitate the entry of SARS-CoV-2 in human target cells [33]. In a recent study by Seyran *et al.* [34], it has been proposed that the flat sialic acid-binding domain at the NTD of the S1 subunit plays a crucial role in fast motion over respiratory epithelium and ACE2 receptor scanning that allow SARS-CoV-2 rapid cellular entry. And also, no high-frequency mutations have been detected so far in the CTD of the S1 subunit [34]. In light of the above and given the small sizes and immunogenic properties of our final four epitopes, they remain viable and better candidates for vaccine development when compared with the whole spike glycoprotein or its S1 and S2 subunits, which unfortunately in the case of SARS-CoV had shown potential to cause lung pathology [35].

SARS-CoV-2 virus is continuing to mutate and evolve [36] like any other virus. As of August 2021, a total of 32 mutations have been reported on Spike glycoprotein by Covid-Miner [37] which is enabled by data from the Global Initiative on Sharing Avian Influenza Data [38]. Out of these 32 mutations, four mutations were found in our three out of final four predicted epitopes (Supplementary Table 5, see Supplementary Data available online at http://bib.oxfordjournals.org/). These four mutations mostly had very low frequency. Epitope number 2, which was the best candidate given its maximum global population coverage of 98.22%, had no mutations reported in Covid-Miner. However, this does not guarantee that epitope number 2 is immortal, but given the robust immuno-informatics pipeline PVPredPip we have produced (Figure 1), new epitopes based on the evolving sequences of SARS-CoV-2 spike glycoprotein can be rapidly predicted, as and when needed.

The effectiveness of a vaccine from a public health program point of view depends on its specificity and the level of population coverage among many other factors. Conservation analysis of the final four epitopes confirmed that they are unique and thus specific for SARS-CoV-2 found in humans as a host. Population coverage analysis confirmed that when the final four epitopes are used as a set, they are able to cover 99.82% of the overall world population (Table 2), indicating that the development of a multi-epitope vaccine may provide better protection against the SARS-CoV-2 virus. Considering the technical constraints, expertise and infrastructure needed for developing multi-epitope vaccine, especially in low income countries, and thus from the health economics and value-for-money perspective, epitope number 2 with its broad allele specificity (Table 1) would be the preferred vaccine candidate. It can cover 98% of the overall world population on its own (Table 2) and, at the same time, is very unique to SARS-CoV-2 isolated from human as a host species (Table 3).

Interestingly, all the final four epitopes showed less population coverage ranging between 0.00 and 59.15% in the South African population (Table 2), which has also been previously demonstrated for other vaccines [20]. Furthermore, SARS-CoV-2 vaccines from Moderna, Pfizer-BioNTech, Novavax and Oxford-AstraZeneca have also been found to be less efficacious against South African population [21–24]. With the recurrent evidence of poor efficacy of vaccines against South African population, it has become inevitably important that more systematic and focused approaches are needed for vaccine development for South African population.

We further validated our results with molecular docking and MD simulation experiments. Complemented with MM-PBSA calculations, essential dynamics analysis and FEL analysis, our results indicate a remarkable binding affinity of the five constituent T-cell epitopes of consensus epitope number 2 towards their corresponding HLA (MHC-I and II) proteins. The *van der* Waals interactions appeared to be the dominant factor responsible for the stability of the HLA-epitope complexes. To conclude, our results provide a profound biophysical insight into the factors and energetics stabilizing the HLA-epitope complexes of the five constituent T-cell epitopes of consensus epitope number 2, which is needed for triggering epitope-specific immune response against SARS-CoV-2. Going forward, the immuno-informatics pipeline (PVPredPip) described and the proposed candidate epitopes could have significant impact upon development of globally effective multi-epitope vaccines against SARS-CoV-2.

Key points

- Considering the variable nature of the SARS-CoV-2 which is continuing to mutate and evolve like any other virus, persistent efforts are needed to develop better vaccine candidates.
- In this study, we have identified four potential epitopes which have the capability to initiate both antibody and cell-mediated immune responses, are non-allergenic and do not trigger autoimmunity.
- These peptide sequences were also evaluated to show 99.82% of global population coverage based on the genotypic frequencies of HLA binding alleles for both MHC class-I and class-II and are unique for SARS-CoV-2 isolated from human as a host species.

- Epitope number 2 alone had a global population coverage of 98.2%; therefore, we further validated binding and interaction of its constituent T-cell epitopes with their corresponding HLA proteins using molecular docking and molecular dynamics simulation experiments, followed by binding free energy calculations with molecular mechanics Poisson–Boltzmann surface area, essential dynamics analysis and free energy landscape analysis.
- The immuno-informatics pipeline (PVPredPip) described and the candidate epitopes discovered herein could have significant impact upon efforts to develop globally effective SARS-CoV-2 vaccines.

Supplementary data

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

Authors' contributions

P.S. conceived the project, led the data analysis and visualization along with P.P., B.P. and T.R. and wrote the first draft of the manuscript. R.P., L.D.G., M.T., A.M., J.H., T.S.R, E.M., D.G. and A.J.B. helped in data interpretation, reviewing and editing of the manuscript. All the authors have revised and approved the submitted version.

Acknowledgements

We are grateful for access to the Tier 2 High Performance Computing resources provided by the Northern Ireland High Performance Computing (NI-HPC) facility. We would also like to thank Dr Arun Kumar from Coalition for Epidemic Preparedness Innovations (CEPI), Oslo, Norway, for critically reading this manuscript and providing valuable feedback.

Conflict of interest: A UK-wide patent application has been filed by the Ulster University; UK Application No. 2102598.6, patent pending. All the aspects of this manuscript are covered in this patent application.

Funding

UKRI funded National Core Study: Immunity (NCSi4P programme), 'Optimal cellular assays for SARS-CoV-2 T-cell, B-cell and innate immunity' to P.S. and D.S.G.; programme grant jointly from Science Foundation Ireland (SFI), Republic of Ireland and Department for the Economy (DfE), Northern Ireland, UK, 'COVRES: Understanding the host-virus response in patients with mild versus serious disease' to P.S., T.S.R., E.M., D.S.G. and A.J.B.; research grant from the Northern Ireland Public Health Agency (HSC R&D Division), 'COVRES2: Identifying temporal immune responses associated with Covid-19 severity' (Grant No. COM/5631/20) to P.S., T.S.R., E.M., D.S.G. and A.J.B.; research grant from the Northern Ireland Public Health Agency (HSC R&D Division), 'Senescence biomarkers for predicting risk in Covid-19 patients' (Grant No. COM/5618/20) to T.S.R. and D.S.G.; Vice-Chancellor's Research Scholarship (VCRS), Ulster University to B.P.; postgraduate studentship by Department for the Economy (DfE), Northern Ireland to T.R.; programme grant jointly from the European Union (EU) Regional Development Fund (ERDF) EU Sustainable Competitiveness Programme for Northern Ireland, the Northern Ireland Public Health Agency (HSC R&D Division) and Ulster University to A.J.B.; Kelvin-2 (Grant No. EP/T022175/1) by UK Engineering and Physical Sciences Research Council (EPSRC) to A.J.B.

References

1. Grifoni A, Sidney J, Zhang Y, *et al.* A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe* 2020;**27**(4): 671–680.e2.
2. Cai Y, Zhang J, Xiao T, *et al.* Distinct conformational states of SARS-CoV-2 spike protein. *Science* 2020;**369**(6511):1586–92.
3. Abdelmageed MI, Abdelmoneim AH, Mustafa MI, *et al.* Design of a multiepitope-based peptide vaccine against the E protein of human COVID-19: an immunoinformatics approach. *Biomed Res Int* 2020;**2020**:1–12.
4. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat Rev Microbiol* 2009;**7**(6):439–50.
5. The WHO Mers-Cov Research Group. State of knowledge and data gaps of middle east respiratory syndrome coronavirus (MERS-CoV) in humans. *PLoS Curr* 2013;**5**: https://pubmed.ncbi.nlm.nih.gov/24270606/.
6. Zhou P, Yang XL, Wang XG, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**(7798):270–3.
7. le TT, Cramer JP, Chen R, *et al.* Evolution of the COVID-19 vaccine development landscape. *Nat Rev Drug Discov* 2020;**19**(10):667–8.
8. Altmann DM, Boyton RJ, Beale R. Immunity to SARS-CoV-2 variants of concern. *Science* 2021;**371**(6534):1103–4.
9. Kiyotani K, Toyoshima Y, Nemoto K, *et al.* Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *J Hum Genet* 2020;**65**(7):569–75.
10. Baruah V, Bose S. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. *J Med Virol* 2020;**92**(5):495–500.
11. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007;**8**:4.
12. Saha S, Raghava GP. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 2006;**65**(1): 40–8.
13. Saha S, Raghava GPS. Bce Pred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: Nicosia G, Cutello V, Bentley PJ *et al.* (eds). *ICARIS*2004, Vol. **3239**. Berlin, Heidelberg: LNCS, Springer, 2004, 197–204.
14. Dhanda SK, Vaughan K, Schulten V, *et al.* Development of a novel clustering tool for linear peptide sequences. *Immunology* 2018;**155**(3):331–45.

15. Trolle T, McMurtrey CP, Sidney J, *et al.* The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J Immunol* 2016;**196**(4):1480–7.
16. Calis JJ, Maybeno M, Greenbaum JA, *et al.* Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol* 2013;**9**(10):e1003266.
17. Dimitrov I, Bangov I, Flower DR, *et al.* Aller TOP v. 2–a server for in silico prediction of allergens. *J Mol Model* 2014;**20**(6):2278.
18. Wu CH, Yeh LS, Huang H, *et al.* The protein information resource. *Nucleic Acids Res* 2003;**31**(1):345–7.
19. Pettersen EF, Goddard TD, Huang CC, *et al.* UCSF chimera–a visualization system for exploratory research and analysis. *J Comput Chem* 2004;**25**(13):1605–12.
20. Vanderslott S, Dadonaite B. Vaccination. 2013. (5 July 2021, date last accessed).
21. Wang Z, Schmidt F, Weisblum Y, *et al.* mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *Nature* 2021;**592**(7855):616–22.
22. Xie X, Liu Y, Liu J, *et al.* Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera. *Nat Med* 2021;**27**(4):620–1.
23. Mahase E. Covid-19: Novavax vaccine efficacy is 86% against UK variant and 60% against South African variant. *BMJ* 2021; **372**:n296.
24. Mahase E. Covid-19: South Africa pauses use of Oxford vaccine after study casts doubt on efficacy against variant. *BMJ* 2021;**372**:n372.
25. Alexander J, Sidney J, Southwood S, *et al.* Development of high potency universal DR-restricted helper epitopes by modification of high affinity DR-blocking peptides. *Immunity* 1994;**1**(9):751–61.
26. Baden LR, el Sahly HM, Essink B, *et al.* Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N Engl J Med* 2021;**384**(5): 403–16.
27. Polack FP, Thomas SJ, Kitchin N, *et al.* Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;**383**(27): 2603–15.
28. Voysey M, Clemens SAC, Madhi SA, *et al.* Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* **397**(10269):99–111.
29. Skwarczynski M, Toth I. Peptide-based synthetic vaccines. *Chem Sci* 2016;**7**(2):842–54.
30. di Natale C, la Manna S, de Benedictis I, *et al.* Perspectives in peptide-based vaccination strategies for syndrome coronavirus 2 pandemic. *Front Pharmacol* 2020;**11**:578382.
31. Nevagi RJ, Toth I, Skwarczynski M. Peptide-based vaccines. In: Koutsopoulos S (ed). *Peptide Applications in Biomedicine, Biotechnology and Bioengineering*. Woodhead Publishing, 2018, 327–58. https://www.sciencedirect.com/book/9780081007365/peptide-applications-in-biomedicine-biotechnology-and-bioengineering
32. Moutaftsi M, Peters B, Pasquetto V, *et al.* A consensus epitope prediction approach identifies the breadth of murine T(CD8+)-cell responses to vaccinia virus. *Nat Biotechnol* 2006;**24**(7): 817–9.
33. Walls AC, Park YJ, Tortorici MA, *et al.* Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020;**181**(2):281–292.e6.
34. Seyran M, Takayama K, Uversky VN, *et al.* The structural basis of accelerated host cell entry by SARS-CoV-2. *FEBS J* 2021;**288**(17):5010–5020. 10.1111/febs.15651.
35. Ma C, Su S, Wang J, *et al.* From SARS-CoV to SARS-CoV-2: safety and broad-spectrum are important for coronavirus vaccine development. *Microbes Infect* 2020;**22**(6–7):245–53.
36. Lauring AS, Hodcroft EB. Genetic variants of SARS-CoV-2-what do they mean? *JAMA* 2021;**325**(6):529–31.
37. Massacci A, Sperandio E, D'Ambrosio L, *et al.* Design of a companion bioinformatic tool to detect the emergence and geographical distribution of SARS-CoV-2 spike protein genetic variants. *J Transl Med* 2020;**18**(1):494.
38. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 2017;**1**(1):33–46.