Received: 26 July 2020 | Revised: 18 September 2020 | Accepted: 7 October 2020

DOI: 10.1002/jmv.26596

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Vaccine design based on 16 epitopes of SARS-CoV-2 spike protein

Jinlei He[1] | Fan Huang[2] | Jianhui Zhang[1] | Qiwei Chen[1] | Zhiwan Zheng[1] | Qi Zhou[1] | Dali Chen[1] | Jiao Li[1] | Jianping Chen[1,3]

[1]Department of Pathogenic Biology, West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu, China

[2]Department of First Surgical, Chengdu Shuangliu Hospital of Traditional Chinese Medicine, Chengdu, China

[3]Animal Disease Prevention and Food Safety Key Laboratory of Sichuan Province, Chengdu, China

Correspondence
Jianping Chen and Jiao Li, Department of Pathogenic Biology, West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, The third part of south renmin road, Chengdu 610041, China.
Email: jpchen007@163.com (J.C.) and joyleeql2019@163.com (J.L.)

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 31572240, 31802184, 31872959, 81672048; China Scholarship Council, Grant/Award Number: 201706240018

Abstract

The global outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) urgently requires an effective vaccine for prevention. In this study, 66 epitopes containing pentapeptides of SARS-CoV-2 spike protein in the IEDB database were compared with the amino acid sequence of SARS-CoV-2 spike protein, and 66 potentially immune-related peptides of SARS-CoV-2 spike protein were obtained. Based on the single-nucleotide polymorphisms analysis of spike protein of 1218 SARS-CoV-2 isolates, 52 easily mutated sites were identified and used for vaccine epitope screening. The best vaccine candidate epitopes in the 66 peptides of SARS-CoV-2 spike protein were screened out through mutation and immunoinformatics analysis. The best candidate epitopes were connected by different linkers in silico to obtain vaccine candidate sequences. The results showed that 16 epitopes were relatively conservative, immunological, nontoxic, and nonallergenic, could induce the secretion of cytokines, and were more likely to be exposed on the surface of the spike protein. They were both B- and T-cell epitopes, and could recognize a certain number of HLA molecules and had high coverage rates in different populations. Moreover, epitopes 897-913 were predicted to have possible cross-immunoprotection for SARS-CoV and SARS-CoV-2. The results of vaccine candidate sequences screening suggested that sequences (without linker, with linker GGGSGGG, EAAAK, GPGPG, and KK, respectively) were the best. The proteins translated by these sequences were relatively stable, with a high antigenic index and good biological activity. Our study provided vaccine candidate epitopes and sequences for the research of the SARS-CoV-2 vaccine.

KEYWORDS

epitope, non-synonymous mutation, SARS-CoV-2, spike protein, vaccine

1 | INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a worldwide pandemic, seriously threatening the health of entire humankind, and an effective vaccine is urgently needed to help people resist the infection of the virus. The spike protein on the virion surface is thought to bind to the host receptor angiotensin-converting enzyme 2 (ACE2) and plays an important role in cell adhesion and virulence similar to SARS-CoV.[1] As the spike protein plays key roles in inducing neutralizing antibodies and protective immunity during SARS-CoV infection,[2,3] the spike protein is considered to be an important vaccine research target for SARS-CoV-2.

According to a report by Lucchese G from Universitätsmedizin Greifswald, the proteome of SARS-CoV-2 was compared with that of

humans, focusing on searching for pentapeptides that are unique to SARS-CoV-2, especially pentapeptides of SARS-CoV-2 in spike protein.[4] They found 107 unique pentapeptides in SARS-CoV-2 spike protein, corresponding to 66 antigen epitopes containing pentapeptides of SARS-CoV-2 spike protein in the Immune Epitope Database (IEDB, http://www.iedb.org). These epitopes in the IEDB database had been experimentally proven to have immunologic relevance[5] and their templates were mainly derived from SARS-CoV. As a novel coronavirus closely related to SARS-CoV,[6] the corresponding peptides of SARS-CoV-2 may also be immunologically relevant. Therefore, based on the above studies, our study aligned the 66 epitope sequences in the IEDB database with the amino acid sequences of SARS-CoV-2 spike protein to obtain the corresponding 66 peptides of SARS-CoV-2 spike protein, which may be candidate epitopes for vaccine research. Mutation analysis and immunoinformatics analysis were used to screen the best vaccine candidate epitopes from the 66 peptides of SARS-CoV-2 spike protein. Then, the best candidate epitopes were connected by different linkers in silico to obtain vaccine candidate sequences. Through structure, surface properties, and function analysis, the best vaccine candidate sequences were finally screened out.

2 | MATERIALS AND METHODS

2.1 | Sequence alignment of 66 epitopes in IEDB database to SARS-CoV-2 spike protein

We downloaded the spike protein amino acid sequence of SARS-CoV-2 isolate Wuhan-Hu-1 from GenBank (GenBank ID: QHD43416.1). The sequences of the 66 epitopes containing pentapeptides of SARS-CoV-2 spike protein were from Lucchese G's report and checked in the IEDB database.[4] Then, the sequences of these epitopes were aligned with the amino acid sequence of SARS-CoV-2 spike protein to obtain 66 peptides at the corresponding sequence position of SARS-CoV-2 spike protein, which might be candidate epitopes of a vaccine.

2.2 | Detection of nonsynonymous mutation sites of SARS-CoV-2 spike protein

As nonsynonymous mutation sites in the viral amino acid sequence may affect the recognition of vaccine antigens, vaccine candidate antigens are generally more inclined to choose conservative sequences.[7,8] Therefore, the inclusion of mutation sites in candidate epitopes of SARS-CoV-2 should be avoided as much as possible. We searched the 2019 Novel Coronavirus Resource (2019nCoVR, https://bigd.big.ac.cn/ncov) from the China National Center for Bioinformation (CNCB) to obtain high-quality genomic data of SARS-CoV-2 clinical isolates. A total of 1218 isolates from 34 countries around the world sampled from June 1, 2020 to June 30, 2020 were selected for analysis. The detailed countries are shown in Table S1. We focused on counting nonsynonymous mutations that cause amino acid changes in spike protein single-nucleotide polymorphism (SNPs). The amino acid sites with nonsynonymous mutations that appeared twice or more in 1218 isolates were considered to be easily mutated. The obtained 66 peptides of SARS-CoV-2 spike protein were checked for the presence of the easily mutated amino acid sites, and peptides containing the easily mutated sites should be noted in subsequent screening.

2.3 | Screening candidate vaccine epitopes in spike protein

The immune protective antigens in the peptides of SARS-CoV-2 spike protein were predicted using immunoinformatics tool Vaxijen v2.0,[9] the toxic peptides were predicted using ToxinPred[10] and the allergenic peptides were predicted using AllergenFP v.1.0.[11] The ability of the epitopes to induce interferon-γ (IFN-γ), interleukin-4 (IL-4), and IL-10 secretion was predicted using IFNepitope,[12] IL4Pred,[13] and IL-10Pred,[14] respectively. The peptides with nonantigenic protection, toxicity, or allergenicity were removed, and the remaining peptides were used as antigen epitopes for subsequent screening. The solvent accessibility of each amino acid of spike protein (template 6xr8.1[15]) was predicted by SWISS-MODEL[16] to screen the epitopes that were more likely to be exposed on the surface of the spike protein. ABCpred[17] and IEDB Bepipred Linear Epitope Prediction 2.0[18] were used to predict B-cell epitopes. NetMHC 4.0 Sever,[19] Rankpep,[20] and SYFPEITHI[21] were used to predict T-cell epitopes and HLA molecules. As different HLA types are expressed at dramatically different frequencies in different ethnicities,[22] after obtaining the results of HLA class I and class II molecules recognized by these epitopes, we predicted the coverage rate of each epitope in different populations using Population Coverage in IEDB Analysis Resource.[22] Although some epitopes contained easily mutated sites, some of them might be strong neutralizing epitopes which might induce strong protections and should also be considered in vaccine design. Therefore, according to the above analysis, the selected vaccine candidate epitopes for SARS-CoV-2 were predicted to be relatively conservative, immunoprotective, nontoxic, and nonallergenic, and could promote the secretion of cytokines and more likely to be exposed on the surface of the spike protein. They were both B- and T-cell epitopes, which could identify a certain number of HLA molecules and had high coverage rates in different populations.

2.4 | Acquisition, analysis, and screening of vaccine candidate sequences

The selected vaccine candidate epitopes were connected by different linkers (no linker, GGGGS, GGGSGGG, EAAAK, GPGPG, AAY, and KK, respectively) to obtain vaccine candidate sequences. Bioinformatics tools were used to analyze and screen the vaccine candidate sequences. PredictProtein was used to predict the amino acid composition, secondary structure composition, solvent accessibility, and gene ontology terms of the candidate sequences.[23] The flexibility and antigenic index of the candidate sequences were predicted using

DNAStar software.[24] Expasy ProtParam tool was used to predict the half-life and stability of the candidate proteins.[25] Finally, through a comprehensive analysis, the best candidate vaccine sequences were selected and will be prepared into vaccines and their immune effects verfied through animal experiments.

3 | RESULTS

3.1 | Epitope sequence alignment

After comparing the amino acid sequences of 66 epitopes in the IEDB database with those of corresponding positions of SARS-CoV-2 spike protein, 66 peptides belonging to SARS-CoV-2 spike protein were obtained and shown in Table 1. Among the 66 epitopes in the IEDB database, 60 epitopes were from the spike protein of SARS-CoV, four epitopes were from hemagglutinin of influenza A virus, and two epitopes were from ribonucleoside-diphosphate reductase large subunit-like protein of human herpesvirus 6B. Among the obtained 66 peptides of SARS-CoV-2 spike protein, six peptides (310-317,[26] 757-764,[26] 891-907,[27] 897-913,[27] 899-906,[26] and 1025-1041[27]) were completely consistent with the sequences of epitopes in the IEDB database, which are bolded in Table 1. Moreover, there were seven peptides (356-372, 356-373, 365-381, 371-387, 373-389, 379-395, and 418-434) partially overlapped with CR3022 epitope of SARS-CoV-2 published in Science by Yuan et al.,[28] which are underlined in Table 1. CR3022 is a neutralizing antibody previously isolated from a convalescent SARS patient and targets a highly conserved epitope that enables cross-reactive binding between SARS-CoV and SARS-CoV-2.[28,29] CR3022 related epitopes may produce cross-protective antibody responses against SARS-CoV and SARS-CoV-2. Therefore, these peptides need to be focused on in subsequent experiments.

3.2 | Detection of nonsynonymous mutation sites of SARS-CoV-2 spike protein

After analyzing the SNPs of 1218 SARS-CoV-2 clinical isolates of spike protein, we found a total of 52 nonsynonymous mutation sites that occurred twice or more, which were considered to be easily mutated and are marked in Figure 1A. The D614G mutation occurred the most and appeared in 1101 SARS-CoV-2 clinical isolates. The D614G mutation was also discovered by Korber et al.,[30] and might lead to the change of SARS-CoV-2 virulence, but further research is needed. We checked the obtained 66 peptide sequences of SARS-CoV-2 to determine whether they contained easily mutated sites, and the peptides containing easily mutated sites should be noted in subsequent screening. Finally, 21 peptides containing easily mutated sites were found and are shown in Table 1. Peptides 15-44, 195-226, 683-699, 690-706, and 690-707 even contained more than two easily mutated sites, and should not be considered as vaccine epitopes.

3.3 | Prediction of protective antigen, toxicity, allergenicity, and cytokine secretion of the 66 peptides

The prediction results of protective antigen, toxicity, allergenicity, and cytokine secretion of the 66 peptides are shown in Table 2. There were 26 peptides without immune protection (score lower than 0.4 in analysis tool), 6 peptides with toxicity (score higher than 0 in analysis tool), and 19 peptides with allergenicity. There were 28 epitopes that had the ability to induce IFN-γ secretion, 42 epitopes had the ability to induce IL-4 secretion, and 24 epitopes had the ability to induce IL-10 secretion. After removing the non-immunoprotective, toxic, or allergenic peptides, there were 28 remaining peptides as candidate epitopes for further screening. Among the 28 epitopes, only 897-913, 899-906, and 1025-1041 epitopes were completely consistent with the sequences in the IEDB database, and only 371-387 and 379-395 epitopes partially overlapped with CR3022 epitope of SARS-CoV-2.[28] Moreover, 371-387, 379-395, and 410-426 epitopes exist in the binding region of spike protein and ACE2,[28] which might be the important candidate vaccine targets. These six epitopes would be noted in the subsequent screening.

3.4 | Prediction of solvent accessibility, B- and T-cell epitopes, and population coverage rates

The solvent accessibility prediction results of spike protein and the remaining 28 epitopes are shown in Figure S1, and the average solvent accessibility scores of amino acids for the 28 epitopes are shown in Table 3. There were 15 epitopes with an average solvent accessibility score higher than 20, which might be considered as vaccine candidates. The prediction results of B-, T-cell epitopes, and HLA class I and class II molecules identified by the 28 epitopes are shown in Table 3. Except that the amino acid sequence of 899-906 epitope was too short to predict, all the other 27 epitopes were predicted to contain B-cell epitopes, which might induce the production of neutralizing antibodies. The analysis results also suggested that the 28 epitopes belonged to T-cell epitopes, 25 of which could recognize HLA class I and class II molecules, two of which could only recognize HLA class I molecules, and one of which could only recognize HLA class II molecules. However, among the six epitopes we focused on, only 371-387, 379-395, and 897-913 could recognize a certain number of HLA class I and class II molecules. The epitope 410-426, 899-906, and 1025-1041 could only recognize HLA class I molecules or class II molecules. The population coverage rates of HLA class I and class II molecules recognized by the 28 epitopes in different populations around the world are shown in Figure 1B. The highest population coverage rate of each epitope was found in Europe and North America, followed by East Asia and Oceania, and the population coverage rates of all epitopes in Africa populations were lower than in other populations. Among the 28 epitopes, 19 epitopes had a world population coverage rate of more than 50%. They were 15-44, 194-210, 195-226, 291-325, 307-323, 371-387, 410-426, 525-566,

TABLE 1 Sequence alignment of 66 epitopes in IEDB database to SARS-CoV-2 spike protein

IEDB ID number	Epitope sequence	Organism	Position in spike protein	Amino acid sequence in spike protein	Easily mutated site
307	aalvsgtatagWTFGAg	SARS-CoV	875-891	SALLAGTITSGWTFGAG	N/A
462	aatkMSECVlgqskrvd	SARS-CoV	**1025-1041**	**AATKMSECVLGQSKRVD**	N/A
1460	agcIIGAEHvdtsyecd	SARS-CoV	647-663	AGCLIGAEHVNNSYECD	653, 660
3176	aMQMAYRF	SARS-CoV	**899-906**	**AMQMAYRF**	N/A
6011	canlllqygsFCTQLnralsgia	SARS-CoV	749-771	CSNLLLQYGSFCTQLNRALTGIA	769
6333	cgpklstdliknqCVNFNfngltgtgvltpsskrfqpfqqfg	SARS-CoV	525-566	CGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG	N/A
6334	cgpklstdliknqCVNFNfngltgtgvltpsskrfqpfqqfgr-dvsdftd	SARS-CoV	525-574	CGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTD	574
7066	csqnplaelkcsvksfeidkGIYQTsnfrvvpsgd	SARS-CoV	291-325	CALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTES	N/A
7217	cttfddvqapnytqhtssmRGVYYPDeifr	SARS-CoV	15-44	CVNLTTRTQLPPAYTNSFTRGVYYPDKVFR	17, 21, 22, 29
7383	CYGVSatklndlcfsnv	SARS-CoV	379-395	CYGVSPTKLNDLCFTNV	382
8239	dfcgkGYHLMSfpqaap	SARS-CoV	1041-1057	DFCGKGYHLMSFPQSAP	N/A
12417	eidkGIYQTsnfrvvps	SARS-CoV	307-323	TVEKGIYQTSNFRVQPT	N/A
15903	ffSTFKCYGVSatklnd	SARS-CoV	373-389	SFSTFKCYGVSPTKLND	382
18161	fvfngtswfiTQRNFfs	SARS-CoV	1095-1110	FVSNGTHWFVTQRNFY	N/A
18515	gaalqipFAMQMAYRFn	SARS-CoV	**891-907**	**GAALQIPFAMQMAYRFN**	N/A
21464	gnliaprGYFKIrsgkssim	Influenza A virus	192-211	FVFKNIDGYFKIYSKHTPIN	211
22321	gsFCTQLn	SARS-CoV	**757-764**	**GSFCTQLN**	N/A
24978	htssmRGVYYPDeifrs	SARS-CoV	29-45	TNSFTRGVYYPDKVFRS	29
25250	IADYNYKLpddfmgcvl	SARS-CoV	418-434	IADYNYKLPDDFTGCVI	N/A
25293	iaglIAIVMvtillccm	SARS-CoV	1221-1237	IAGLIAIVMVTIMLCCM	1237
25378	iapgqtgvIADYNYKLp	SARS-CoV	410-426	IAPGQTGKIADYNYKLP	N/A
25382	iaprGYFKIrngkssimrsdapigtcssecit	Influenza A virus	195-226	KNIDGYFKIYSKHTPINLVRDLPQGFSALEPL	211, 215, 220
29728	iywtivkpgdillinstgnliaprGYFKIrn	Influenza A virus	175-205	FLMDLEGKQGNFKNLREFVFKNIDGYFKIYS	180, 181
30987	kGIYQTsn	SARS-CoV	**310-317**	**KGIYQTSN**	N/A
30988	kGIYQTsnfrvvpsgdvvrf	SARS-CoV	310-329	KGIYQTSNFRVQPTESIVRF	N/A
31581	kkisnCVADYsvlynst	SARS-CoV	356-372	KRISNCVADYSVLYNSA	N/A
31582	kkisnCVADYsvlynstf	SARS-CoV	356-373	KRISNCVADYSVLYNSAS	N/A

TABLE 1 (Continued)

IEDB ID number	Epitope sequence	Organism	Position in spike protein	Amino acid sequence in spike protein	Easily mutated site
33305	ksfeidkGIYQTsnfrvv	SARS-CoV	304-321	KSFTVEKGIYQTSNFRVQ	N/A
33358	ksivAYTMSlgadssia	SARS-CoV	690-706	QSIIAYTMSLGAENSVA	690, 691, 701
33874	kTSVDCnMYICGDSTEC	SARS-CoV	733-749	KTSVDCTMYICGDSTEC	N/A
36579	liknqCVNFNfngItgt	SARS-CoV	533-549	LVKNKCVNFNFNGLTGT	N/A
36815	lkcsvksfeidkGIYQT	SARS-CoV	299-315	TKCTLKSFTVEKGIYQT	N/A
36856	lkgacscgsCCKFDedd	SARS-CoV	1244-1260	LKGCCSCGSCCKFDEDD	N/A
37758	llrstsqksivAYTMSl	SARS-CoV	683-699	RARSVASQSIIAYTMSL	688, 690, 691
39023	lqygsFCTQLnralsgi	SARS-CoV	754-770	LQYGSFCTQLNRALTGI	769
41177	MAYRFNGIgvtqnvlye	SARS-CoV	902-918	MAYRFNGIGVTQNVLYE	N/A
42999	mvtilLCCMTSCCsclk	SARS-CoV	1229-1245	MVTIMLCCMTSCCSCLK	1237
43145	nafnCTFEYisdafsld	SARS-CoV	162-178	SANNCTFEYVSQPFLMD	N/A
46379	nvfqtqagclIGAEHvd	SARS-CoV	641-657	NVFQTRAGCLIGAEHVN	653
46822	PAICHegkayfpregvfvfngtswfitqrnffs	SARS-CoV	1079-1111	PAICHDGKAHFPREGVFVSNGTHWFVTQRNFYE	N/A
47479	pFAMQMAYRFNGIgvtq	SARS-CoV	**897-913**	**PFAMQMAYRFNGIGVTQ**	N/A
49968	pvsmakTSVDCnMYICGds	SARS-CoV	728-746	PVSMTKTSVDCTMYICGDS	N/A
50058	pwyvwlgfiaglIAIVM	SARS-CoV	1213-1229	PWYIWLGFIAGLIAIVM	N/A
53202	rasanlaatkMSECVlg	SARS-CoV	1019-1035	RASANLAATKMSECVLG	N/A
54989	rnfttaPAICHegkayf	SARS-CoV	1073-1089	KNFTTAPAICHDGKAHF	1078
58143	sgncdvvigiinNTVYD	SARS-CoV	1123-1139	SGNCDVVIGIVNNTVYD	N/A
58730	sivAYTMSl	SARS-CoV	691-699	SIIAYTMSL	691
61554	stdliknqCVNFNfn	SARS-CoV	530-544	STNLVKNKCVNFNFN	N/A
61598	stffSTFKCYGVSatkl	SARS-CoV	371-387	SASFSTFKCYGVSPTKL	382
62872	tagWTFGAgaalqipfa	SARS-CoV	883-899	TSGWTFGAGAALQIPFA	N/A
63309	tecanlllqygsFCTQL	SARS-CoV	747-763	TECSNLLLQYGSFCTQL	N/A
68971	vigiinNTVYDplqpel	SARS-CoV	1129-1145	VIGIVNNTVYDPLQPEL	N/A
72205	VYYPDeifrsdtlyltqd	SARS-CoV	36-53	VYYPDKVFRSSVLHSTQD	N/A
74173	yicgDSTECanlllqyg	SARS-CoV	741-757	YICGDSTECSNLLLQYG	N/A
75920	ysvlynstffSTFKCYG	SARS-CoV	365-381	YSVLYNSASFSTFKCYG	N/A

(Continues)

TABLE 1 (Continued)

IEDB ID number	Epitope sequence	Organism	Position in spike protein	Amino acid sequence in spike protein	Easily mutated site
99918	CTFEYisdafsld	SARS-CoV	166-178	CTFEYVSQPFLMD	N/A
100048	gaalqipFAMQMAYRF	SARS-CoV	891-906	GAALQIPFAMQMAYRF	N/A
100230	ksivAYTMSlgadssiay	SARS-CoV	690-707	QSIIAYTMSLGAENSVAY	690, 691, 701
100300	MAYRFNGIgvtqnvly	SARS-CoV	902-917	MAYRFNGIGVTQNVLY	N/A
100316	nafnCTFEYisdafsldv	SARS-CoV	162-179	SANNCTFEYVSQPFLMDL	N/A
100537	swfiTQRNFfspqii	SARS-CoV	1101-1115	HWFVTQRNFYEPQII	N/A
100711	agcllGAEHvdtsyecdi	SARS-CoV	647-664	AGCLIGAEHVNNSYECDI	653, 660
129239	liaprGYFKIrsgkssi	Influenza A virus	194-210	FKNIDGYFKIYSKHTPI	N/A
532052	gtswfiTQRNFfspq	SARS-CoV	1099-1113	GTHWFVTQRNFYEPQ	N/A
873061	mmcehiyytcvrTSVDCc	Human herpes virus 6B	722-739	VTTEILPVSMTKTSVDCT	N/A
874104	ytcvrTSVDCcmkgaep	Human herpes virus 6B	729-745	VSMTKTSVDCTMYICGD	N/A

Note: In the table, the capitalized amino acid sequences were the sequences in SARS-CoV-2 spike protein. The underlined peptides partially overlaped with CR3022 epitope published in Science by Yuan et al.[28] The bolded peptides were completely consistent with the corresponding epitope sequences in the IEDB database.

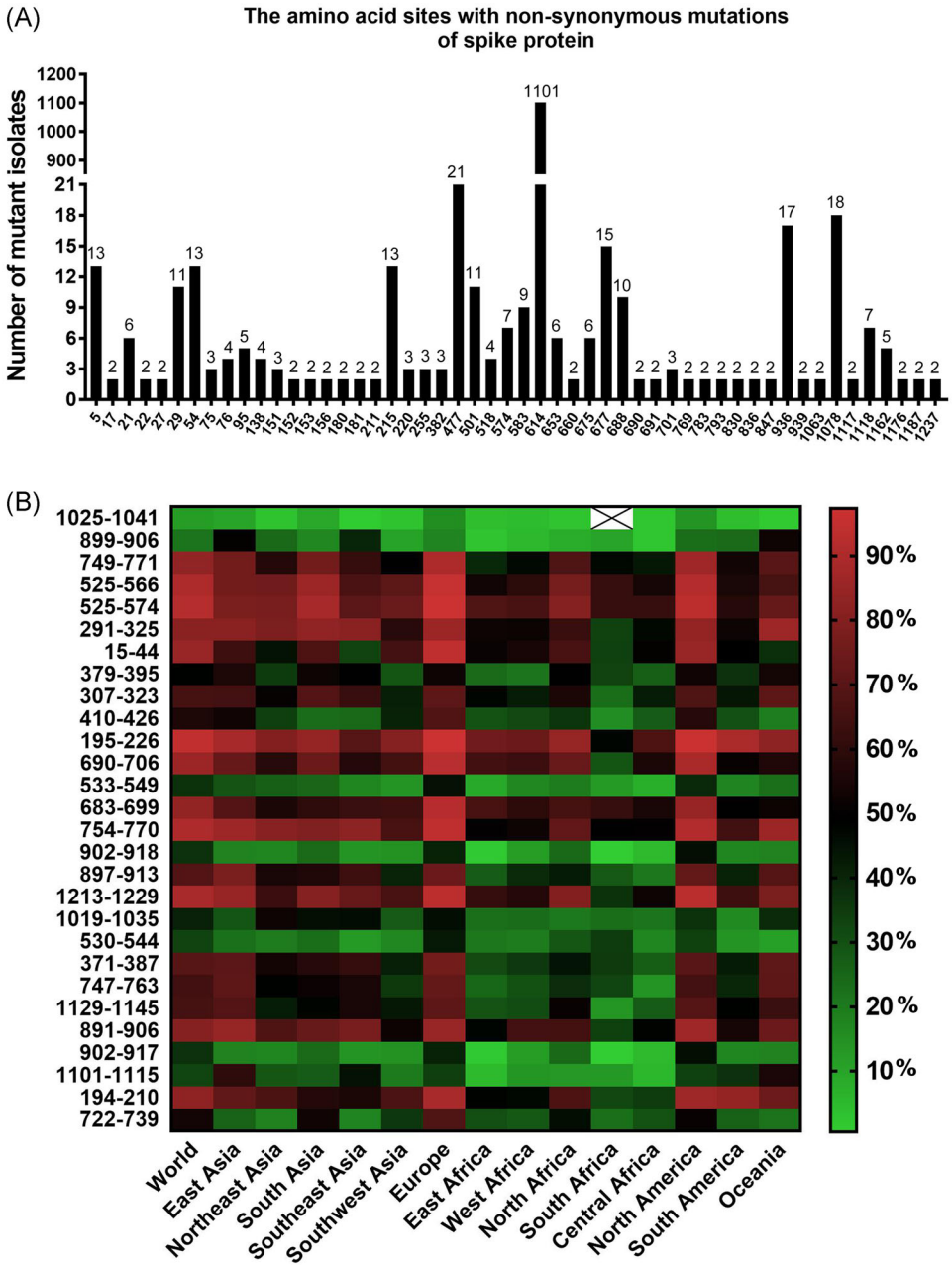

FIGURE 1 Mutation analysis of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein and prediction of epitope population coverage. (A) The amino acid sites with nonsynonymous mutations of spike protein in 1218 clinical isolates of SARS-CoV-2. The amino acid sites with nonsynonymous mutations appeared twice or more in 1218 isolates, which were considered to be easily mutated and marked in the figure. We totally found 52 easily mutated sites. (B) Prediction of population coverage rates of 28 epitopes in SARS-CoV-2 spike protein

525-574, 683-699, 690-706, 722-739, 747-763, 749-771, 754-770, 891-906, 897-913, 1129-1145, and 1213-1229.

3.5 | Screening vaccine candidate epitopes

Combined with the prediction results, among the 28 epitopes, epitopes with an average accessibility score of more than 20 or a world population coverage rate of more than 50% were selected. Therefore, a total of 21 epitopes were selected. However, among the 21 epitopes, eight of them (15-44, 195-226, 371-387, 525-574, 683-699, 690-706, 749-771, and 754-770) had easily mutated sites. Considering the importance of the 371-387 epitope and the relatively few mutations of 749-771 and 754-770 epitopes, these three epitopes were retained. Finally, 16 epitopes were selected for vaccine preparation, they were 194-210, 291-325, 307-323, 371-387, 410-426, 525-566, 530-544, 722-739, 747-763, 749-771, 754-770, 891-906, 897-913, 1101-1115, 1129-1145, and 1213-1229. The 16 epitopes were relatively conservative, immunological,

TABLE 2 Prediction of protective antigen, toxicity, allergenicity, and cytokine secretion of 66 peptides in SARS-CoV-2 spike protein

Position in spike protein	Amino acid sequence in spike protein	Protective antigen prediction	Toxicity prediction	Allergenicity prediction	IFN-γ prediction	IL-4 prediction	IL-10 prediction
875-891	SALLAGTITSGWTFGAG	Nonantigen	Nontoxin	Allergen	Inducer	Noninducer	Noninducer
1025-1041	**AATKMSECVLGQSKRVD**	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
647-663	AGCLIGAEHVNNSYECD	Antigen	Toxin	Nonallergen	Inducer	Inducer	Inducer
899-906	**AMQMAYRF**	Antigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
749-771	CSNLLLQYGSFCTQLNRALTGIA	Antigen	Nontoxin	Nonallergen	Inducer	Noninducer	Inducer
525-566	CGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG	Antigen	Nontoxin	Nonallergen	N/A	Noninducer	Inducer
525-574	CGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGR-DIADTTD	Antigen	Nontoxin	Nonallergen	N/A	Noninducer	Inducer
291-325	CALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTES	Antigen	Nontoxin	Nonallergen	N/A	Inducer	Inducer
15-44	CVNLTTRTQLPPAYTNSFTRGVYYPDKVFR	Antigen	Nontoxin	Nonallergen	Inducer	Inducer	Inducer
379-395	CYGVSPTKLNDLCFTNV	Antigen	Nontoxin	Nonallergen	Non-Inducer	Inducer	Noninducer
1041-1057	DFCGKGYHLMSFPQSAP	Nonantigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
307-323	TVEKGIYQTSNFRVQPT	Antigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
373-389	SFSTFKCYGVSPTKLND	Antigen	Nontoxin	Allergen	Noninducer	Inducer	Noninducer
1095-1110	FVSNGTHWFVTQRNFY	Nonantigen	Nontoxin	Allergen	Noninducer	Inducer	Noninducer
891-907	**GAALQIPFAMQMAYRFN**	Antigen	Nontoxin	Allergen	Noninducer	Inducer	Noninducer
192-211	FVFKNIDGYFKIYSKHTPIN	Antigen	Nontoxin	Allergen	Inducer	Inducer	Inducer
757-764	**GSFCTQLN**	Antigen	Nontoxin	Allergen	Inducer	Noninducer	Noninducer
29-45	TNSFTRGVYYPDKVFRS	Nonantigen	Nontoxin	Allergen	Inducer	Noninducer	Inducer
418-434	IADYNYKLPDDFTGCVI	Antigen	Nontoxin	Allergen	Non-Inducer	Inducer	Noninducer
1221-1237	IAGLIAIVMVTIMLCCM	Antigen	Toxin	Allergen	Inducer	Inducer	Inducer
410-426	IAPGQTGKIADYNYKLP	Antigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
195-226	KNIDGYFKIYSKHTPINLVRDLPQGFSALEPL	Antigen	Nontoxin	Nonallergen	N/A	Noninducer	Inducer
175-205	FLMDLEGKQGNFKNLREFVFKNIDGYFKIYS	Nonantigen	Nontoxin	Nonallergen	N/A	Inducer	Inducer
310-317	**KGIYQTSN**	Nonantigen	Nontoxin	Allergen	Inducer	Inducer	Noninducer
310-329	KGIYQTSNFRVQPTESIVRF	Nonantigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
356-372	KRISNCVADYSVLYNSA	Nonantigen	Nontoxin	Nonallergen	Non-Inducer	Inducer	Inducer
356-373	KRISNCVADYSVLYNSAS	Nonantigen	Nontoxin	Nonallergen	Non-Inducer	Inducer	Inducer

TABLE 2 (Continued)

Position in spike protein	Amino acid sequence in spike protein	Protective antigen prediction	Toxicity prediction	Allergenicity prediction	IFN-γ prediction	IL-4 prediction	IL-10 prediction
304-321	KSFTVEKGIYQTSNFRVQ	Nonantigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
690-706	QSIIAYTMSLGAENSVA	Antigen	Nontoxin	Nonallergen	Inducer	Noninducer	Noninducer
733-749	KTSVDCTMYICGDSTEC	Nonantigen	Toxin	Nonallergen	Non-Inducer	Noninducer	Noninducer
533-549	LVKNKCVNFNFNGLTGT	Antigen	Nontoxin	Nonallergen	Non-Inducer	Inducer	Noninducer
299-315	TKCTLKSFTVEKGIYQT	Nonantigen	Nontoxin	Allergen	Inducer	Inducer	Noninducer
1244-1260	LKGCCSCGSCCKFDEDD	Nonantigen	Toxin	Nonallergen	Noninducer	Inducer	Noninducer
683-699	RARSVASQSIIAYTMSL	Antigen	Nontoxin	Nonallergen	Inducer	Noninducer	Noninducer
754-770	LQYGSFCTQLNRALTGI	Antigen	Nontoxin	Nonallergen	Inducer	Noninducer	Noninducer
902-918	MAYRFNGIGVTQNVLYE	Antigen	Nontoxin	Nonallergen	Noninducer	Noninducer	Noninducer
1229-1245	MVTIMLCCMTSCCSCLK	Nonantigen	Toxin	Nonallergen	Noninducer	Inducer	Inducer
162-178	SANNCTFEYVSQPFLMD	Nonantigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
641-657	NVFQTRAGCLIGAEHVN	Antigen	Nontoxin	Allergen	Noninducer	Noninducer	Inducer
1079-1111	PAICHDGKAHFPREGVFVSNGTHWFVTQRNFYE	Nonantigen	Nontoxin	Nonallergen	N/A	Inducer	Inducer
897-913	**PFAMQMAYRFNGIGVTQ**	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
728-746	PVSMTKTSVDCTMYICGDS	Nonantigen	Nontoxin	Allergen	Noninducer	Noninducer	Noninducer
1213-1229	PWYIWLGFIAGLIAIVM	Antigen	Nontoxin	Nonallergen	Noninducer	Noninducer	Inducer
1019-1035	RASANLAATKMSECVLG	Antigen	Nontoxin	Nonallergen	Inducer	Noninducer	Noninducer
1073-1089	KNFTTAPAICHDGKAHF	Nonantigen	Nontoxin	Allergen	Inducer	Inducer	Inducer
1123-1139	SGNCDVVIGIVNNTVYD	Antigen	Nontoxin	Allergen	Noninducer	Noninducer	Noninducer
691-699	SIIAYTMSL	Antigen	Nontoxin	Allergen	Noninducer	Noninducer	Noninducer
530-544	STNLVKNKCVNFNFN	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
371-387	SASFSTFKCYGVSPTKL	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
883-899	TSGWTFGAGAALQIPFA	Nonantigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
747-763	TECSNLLLQYGSFCTQL	Antigen	Nontoxin	Nonallergen	Inducer	Noninducer	Inducer
1129-1145	VIGIVNNTVYDPLQPEL	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Inducer
36-53	VYYPDKVFRSSVLHSTQD	Nonantigen	Nontoxin	Nonallergen	Inducer	Noninducer	Inducer
741-757	YICGDSTECSNLLLQYG	Nonantigen	Nontoxin	Allergen	Inducer	Noninducer	Noninducer

(Continues)

TABLE 2 (Continued)

Position in spike protein	Amino acid sequence in spike protein	Protective antigen prediction	Toxicity prediction	Allergenicity prediction	IFN-γ prediction	IL-4 prediction	IL-10 prediction
365-381	YSVLYNSASFSTFKCYG	Nonantigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
166-178	CTFEYVSQPFLMD	Nonantigen	Nontoxin	Allergen	Noninducer	Inducer	Noninducer
891-906	GAALQIPFAMQMAYRF	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
690-707	QSIIAYTMSLGAENSVAY	Antigen	Nontoxin	Allergen	Inducer	Noninducer	Noninducer
902-917	MAYRFNGIGVTQNVLY	Antigen	Nontoxin	Nonallergen	Noninducer	Noninducer	Noninducer
162-179	SANNCTFEYVSQPFLMDL	Nonantigen	Nontoxin	Nonallergen	Inducer	Inducer	Noninducer
1101-1115	HWFVTQRNFYEPQII	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
647-664	AGCLIGAEHVNNSYECDI	Antigen	Toxin	Nonallergen	Inducer	Inducer	Inducer
194-210	FKNIDGYFKIYSKHTPI	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Inducer
1099-1113	GTHWFVTQRNFYEPQ	Nonantigen	Nontoxin	Nonallergen	Noninducer	Inducer	Noninducer
722-739	VTTEILPVSMTKTSVDCT	Antigen	Nontoxin	Nonallergen	Noninducer	Inducer	Inducer
729-745	VSMTKTSVDCTMYICGD	Nonantigen	Nontoxin	Nonallergen	Noninducer	Noninducer	Noninducer

Note: In the table, underlined peptides partially overlaped with CR3022 epitope published in Science by Yuan et al.[28] The bolded peptides were consistent with the corresponding epitope sequences in IEDB database. N/A meant undetectable because the peptide length was beyond the range of the analysis system (≤30 amino acids).

TABLE 3 Prediction of B cell and T cell epitopes of 28 epitopes in SARS-CoV-2 spike protein

Position in spike protein	Amino acid sequence in spike protein	Average SOA score	B cell epitope prediction		T cell epitope prediction	
			ABCpred	IEDB	HLA class I molecule	HLA class II molecule
1025-1041	**AATKMSECVLGQSKRVD**	12.33	2-17	7-14	N/A	HLA-DRB1 0101
899-906	**AMQMAYRF**	4.06	N/A	0	HLA-A2402	N/A
749-771	CSNLLLQYGSFCTQLNRALTGIA	22.79	2-17	17	HLA-A0301 HLA-A2402 HLA-B0801 HLA-B1501 HLA-B2705 HLA-B3901	HLA-DRB1 0101 HLA-DRB1 0301 HLA-DRB1 0401 HLA-DRB1 0701 HLA-DRB1 1501
525-566	CGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFG	25.59	1-37	5-39	HLA-A0101 HLA-A0301 HLA-A1101 HLA-A2601 HLA-B0702 HLA-B0801 HLA-B1501 HLA-B3901 HLA-B3902 HLA-B5101	HLA-DRB1 0101 HLA-DRB1 0301 HLA-DRB1 0401 HLA-DRB1 0404 HLA-DRB1 0701 HLA-DRB1 1501
525-574	CGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGR-DIADTTD	24.71	2-21, 24-43	7-46	HLA-A0101 HLA-A0301 HLA-A1101 HLA-A2601 HLA-B0702 HLA-B0801 HLA-B1501 HLA-B3901 HLA-B3902 HLA-B5101	HLA-DRB1 0101 HLA-DRB1 0301 HLA-DRB1 0401 HLA-DRB1 0404 HLA-DRB1 0701 HLA-DRB1 1101 HLA-DRB1 1501
291-325	CALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTES	32.24	2-28	18-32	HLA-A0301 HLA-A1101 HLA-A2402 HLA-A2601 HLA-B5801	HLA-DRB1 0101 HLA-DRB1 0401 HLA-DRB1 0701 HLA-DRB1 1101 HLA-DRB1 1501
15-44	CVNLTTRTQLPPAYTNSFTRGVYYPDKVFR	31.38	5-24	5-26	HLA-A0101 HLA-A0201 HLA-A0301 HLA-A2601 HLA-B0702 HLA-B1501 HLA-B1516 HLA-B2705	HLA-DRB1 0101 HLA-DRB1 0401 HLA-DRB1 0701
379-395	CYGVSPTKLNDLCFTNV	15.48	1-16	6-13	HLA-A2402	HLA-DRB1 0301 HLA-DRB1 0701
307-323	TVEKGIYQTSNFRVQPT	36.14	1-16	6-14	HLA-A0301 HLA-A2402 HLA-B5801	HLA-DRB1 0701 HLA-DRB1 1101 HLA-DRB1 1501
410-426	IAPGQTGKIADYNYKLP	8.61	2-17	6-13	HLA-A0201 HLA-B0702 HLA-B1501 HLA-B5101	N/A
195-226	KNIDGYFKIYSKHTPINLVRDLPQGFSALEPL	30.24	3-22	8-28	HLA-A0101 HLA-A0201 HLA-A2601 HLA-B0702 HLA-B0801 HLA-B1501 HLA-B3501 HLA-B3901 HLA-B5802	HLA-DRB1 0101 HLA-DRB1 0401 HLA-DRB1 0404 HLA-DRB1 0701 HLA-DRB1 1101 HLA-DRB1 1501 HLA-DQA1 0301 HLA-DQB1 0302
690-706	QSIIAYTMSLGAENSVA	30.17	1-16	6-13	HLA-A0201 HLA-A2601 HLA-B0702 HLA-B0801 HLA-B1501 HLA-B1516 HLA-B3901 HLA-B5801	HLA-DRB1 0101 HLA-DRB1 0401 HLA-DRB1 0402 HLA-DRB1 0701 HLA-DRB1 1101 HLA-DRB1 1501

(Continues)

TABLE 3 (Continued)

Position in spike protein	Amino acid sequence in spike protein	Average SOA score	B cell epitope prediction		T cell epitope prediction	
			ABCpred	IEDB	HLA class I molecule	HLA class II molecule
533-549	LVKNKCVNFNFNGLTGT	19.67	4-17	8-13	HLA-B0801 HLA-B1501	HLA-DRB1 0404 HLA-DRB1 1501
683-699	RARSVASQSIIAYTMSL	N/A	2-15	6-13	HLA-A0101 HLA-A0201 HLA-A2601 HLA-B0702 HLA-B0801 HLA-B1501 HLA-B1516 HLA-B2705 HLA-B3901 HLA-B4001 HLA-B5801 HLA-B5802	HLA-DRB1 0101 HLA-DRB1 0301
754-770	LQYGSFCTQLNRALTGI	25.03	2-15	0	HLA-A0201 HLA-A0301 HLA-A1101 HLA-B1501 HLA-A2402 HLA-B2705 HLA-B3901	HLA-DRB1 0101 HLA-DRB1 0301 HLA-DRB1 0401 HLA-DRB1 0701
902-918	MAYRFNGIGVTQNVLYE	12.31	2-15	6-13	HLA-B2705	HLA-DRB1 0401 HLA-DRB1 0402 HLA-DRB1 0404 HLA-DRB1 1501
897-913	**PFAMQMAYRFNGIGVTQ**	12.19	2-17	10, 12-13	HLA-A2402 HLA-A2601 HLA-B0801 HLA-B1501 HLA-B2705 HLA-B3901 HLA-B5801	HLA-DRB1 0101 HLA-DRB1 0401 HLA-DRB1 0402 HLA-DRB1 1501
1213-1229	PWYIWLGFIAGLIAIVM	N/A	1-16	0	HLA-A0201 HLA-A2402 HLA-A2601 HLA-B3901	HLA-DRB1 0101 HLA-DRB1 0301 HLA-DRB1 0401 HLA-DRB1 0402 HLA-DRB1 0701 HLA-DRB1 1101 HLA-DRB1 1501
1019-1035	RASANLAATKMSECVLG	14.31	2-17	8-13	HLA-A0301 HLA-A1101 HLA-B5801	HLA-DRB1 0101
530-544	STNLVKNKCVNFNFN	28.43	3-14	6-11	HLA-B0801 HLA-B1501	HLA-DRB1 0301
371-387	SASFSTFKCYGVSPTKL	24.57	3-16	8-13	HLA-A0101 HLA-A0301 HLA-A2402 HLA-B1501 HLA-B3901 HLA-B5801	HLA-DRB1 0401 HLA-DRB1 1501
747-763	TECSNLLLQYGSFCTQL	26.54	1-16	7-13	HLA-A0101 HLA-A2402 HLA-B0801 HLA-B1501 HLA-B3901	HLA-DRB1 0101 HLA-DRB1 1501
1129-1145	VIGIVNNTVYDPLQPEL	37.57	1-16	5-13	HLA-A0201 HLA-A2402	HLA-DRB1 0701 HLA-DRB1 0801
891-906	GAALQIPFAMQMAYRF	9.67	1-14	0	HLA-A2402 HLA-A2601 HLA-B0801 HLA-B1501 HLA-B3501 HLA-B3901 HLA-B4001 HLA-B5301 HLA-B5801	HLA-DRB1 0101 HLA-DRB1 0401 HLA-DRB1 0404 HLA-DRB1 0701 HLA-DRB1 1501

TABLE 3 (Continued)

Position in spike protein	Amino acid sequence in spike protein	Average SOA score	B cell epitope prediction		T cell epitope prediction	
			ABCpred	IEDB	HLA class I molecule	HLA class II molecule
902-917	MAYRFNGIGVTQNVLY	11.38	2-15	5-12	HLA-B2705	HLA-DRB1 0401 HLA-DRB1 0402 HLA-DRB1 0404 HLA-DRB1 1501
1101-1115	HWFVTQRNFYEPQII	23.55	2-13	5-11	HLA-A2402 HLA-A2601 HLA-B2705	HLA-DRB1 0801
194-210	FKNIDGYFKIYSKHTPI	22.72	1-16	9-13	HLA-A0101 HLA-A0201 HLA-B0702 HLA-B0801 HLA-B3901	HLA-DRB1 1501 HLA-DQA1 0301 HLA-DQB1 0302
722-739	VTTEILPVSMTKTSVDCT	13.03	1-16	9-14	HLA-A0101 HLA-A0301 HLA-B1516 HLA-B3901	HLA-DRB1 0101 HLA-DRB1 0701

Note: In the table, the underlined epitopes overlaped with CR3022 epitope published in Science by Yuan et al.[28] The bolded epitopes were consistent with the corresponding epitope sequences in IEDB database. The prediction results of T-cell epitope and MHC binding combined the results of three analysis tools NetMHC 4.0 Sever, Rankpep and SYFPEITHI. N/A meant undetectable. HLA, human lymphocyte antigen; SOA, solvent accessibility.

nontoxic, and nonallergenic, and could induce the secretion of cytokines, and more likely to be exposed on the surface of the spike protein. They were both B- and T-cell epitopes, could recognize a certain number of HLA molecules, and their population coverage rates in the world were more than 50%.

3.6 | Vaccine candidate sequences acquisition and general analysis

The 16 candidate epitopes were eventually merged into 11 peptides and connected with different linkers to obtain vaccine candidate sequences. The schematic diagram of tandem sequences of the 11 peptides is shown in Figure 2A. After the analysis of the candidate sequences by PredictProtein, DNAStar, and Expasy ProtParam tool, their secondary structure and surface properties were obtained. The number of amino acids with no linker, linker GGGGS, GGGSGGG, EAAAK, GPGPG, AAY, and KK were 243aa, 293aa, 313aa, 293aa, 293aa, 273aa, and 263aa, respectively. Their molecular weights were 27046.39 Da, 30199.25 Da, 31340.29 Da, 31751.62 Da, 30700.29 Da, 30099.73 Da, and 29609.79 Da, respectively. The isoelectric points of the sequences were 8.84, 8.84, 8.84, 8.76, 8.84, 8.78, and 9.85, respectively. As the N-terminal amino acids of the sequences were all phenylalanine (F), their half-lives were the same. Their estimated half-life were: 1.1 h (mammalian reticulocytes, in vitro), 3 min (yeast, in vivo) and 3 min (*Escherichia coli*, in vivo). Therefore, a methionine (M) was considered to add at the N-terminus of each of the sequences to extend the half-life of the protein. Moreover, the instability index of the seven sequences was 27.39, 40.80, 38.52, 27.54, 23.62, 25.33, and 22.02, which suggested that the protein with linker GGGGS was classified as unstable and the other proteins were classified as stable.

The analysis results of amino acid composition, secondary structure composition, and solvent accessibility of the candidate sequences are shown in Figure 2B. We found that the addition of different linkers changed the secondary structure composition of the proteins, especially the GPGPG and AAY linkers increased the Loop structure. Loops are irregular structures that connect two secondary structure elements in proteins, and they often play important roles in function, including enzyme reactions and ligand binding.[31] Moreover, the addition of linkers also changed the solvent accessibility of the proteins. The addition of these six linkers increased the solvent accessibility of the proteins, and all exposed more amino acids on the protein surface. The flexibility and antigenic index results of the sequences were shown in Figure 3. Compared with the results of no linker sequence, except for linker AAY, the addition of other linkers all increased the flexibility and antigenic index of the sequences.

3.7 | Functional analysis of vaccine candidate sequences

The prediction results of gene ontology terms of the sequences were shown in supplementary material Figure S2. The number of

molecular function ontology and cell composition ontology of the protein sequences was changed by different linkers, and the specific results of molecular function ontology prediction for the sequences are shown in supplementary material Figure S3-S9. However, protein sequences connected by different linkers had almost no effect on biological process ontology. Interestingly, the number of molecular function ontology or cell composition ontology of sequence with linker GGGGS and GPGPG was more than that of the other sequences, indicating that the use of GGGGS or GPGPG linker might increase some biological activities of the protein. However, the previous protein stability prediction showed that the protein with the GGGGS linker was unstable, so the GGGGS linker was not a good choice.

3.8 | Comprehensive selection of vaccine candidate sequences

Based on the above analysis, the protein sequence with linker GGGGS was predicted to be unstable and the sequence with linker AAY was predicted to reduce the flexibility and antigenic index of the protein. Therefore, considering the secondary structure, flexibility, antigenic index, solvent accessibility, stability, and function prediction results of the sequences, we finally selected five sequences (without linker, with linker GGGSGGG, EAAAK, GPGPG, and KK, respectively) as the SARS-CoV-2 vaccine candidate sequences. These vaccine candidate sequences contained T- and B-cell epitopes exposed on the surface of spike protein and the HLA molecules recognized by the epitopes had high population coverage rates. These sequences were predicted to be stable, with high antigenic index and good biological activity, especially the sequence linked by GPGPG.

4 | DISCUSSION

At present, scientists all over the world are stepping up the research into the COVID-19 vaccine. According to the Draft landscape of COVID-19 candidate vaccines-7 July 2020 published by the World Health Organization (WHO),[32] 21 candidate vaccines of COVID-19 had been approved for clinical trials. These included five RNA vaccines, four inactivated vaccines, four DNA vaccines, four protein subunit vaccines, three viral vector vaccines, and one plant-derived virus-like particle vaccine. The vaccine in this study belonged to the class of epitope vaccine and peptide vaccine, increasing the diversity of COVID-19 vaccine types. Previous studies on vaccines for other infectious diseases showed that different types of vaccines had their own limitations.[33] In most cases, the immune effects of combining different types of vaccines were stronger than that of a single vaccine alone,[34,35] and this situation had also appeared in SARS-CoV vaccine research.[36,37] Therefore, we recommend in future research and development of COVID-19 vaccines, considering the diversity of vaccine types, combining the advantages and disadvantages of different types of vaccines, and using different vaccines for immunization, and carrying out research on heterologous prime-boost vaccines.

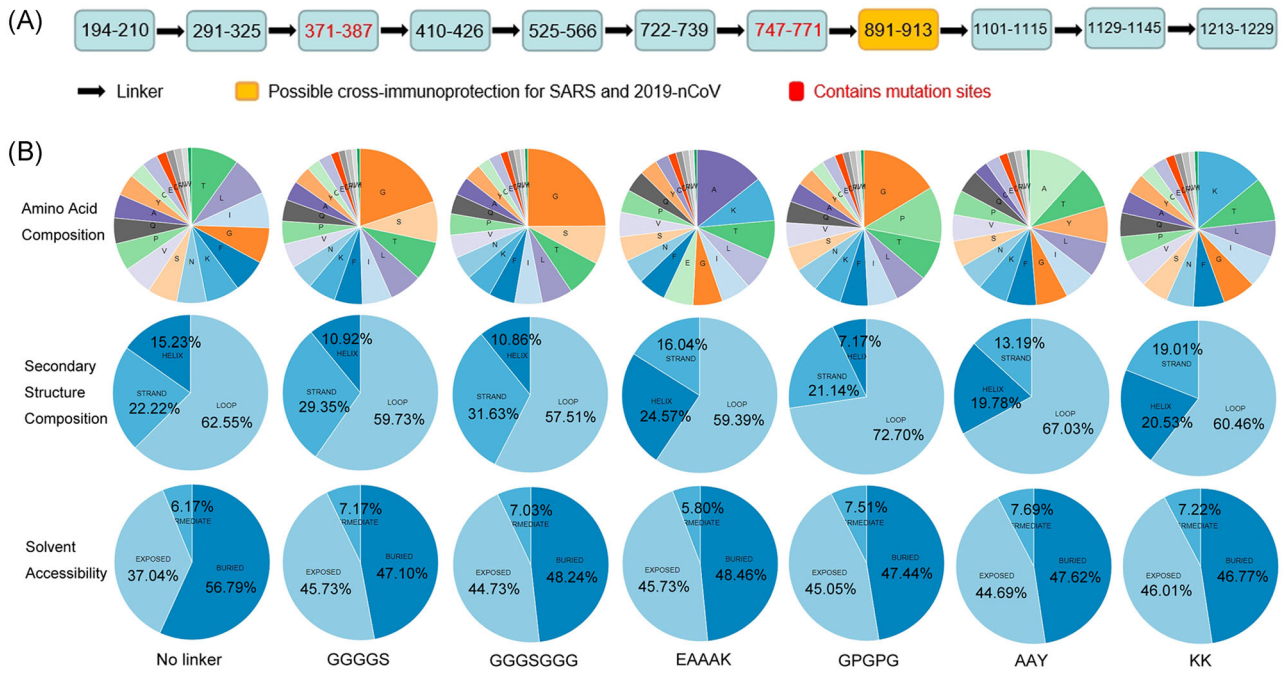

FIGURE 2 Schematic diagram of the tandem sequence of 11 vaccine candidate peptides and sequence analysis after connecting with different linkers. (A) Schematic diagram of tandem sequences of 11 vaccine candidate peptides. Peptides 897-913 were predicted possible cross-immunoprotection for SARS-CoV and SARS-CoV-2. Peptides 371-387 and 747-771 contained easily mutated sites. (B) Amino acid composition, secondary structure composition, and solvent accessibility analysis of different vaccine candidate sequences

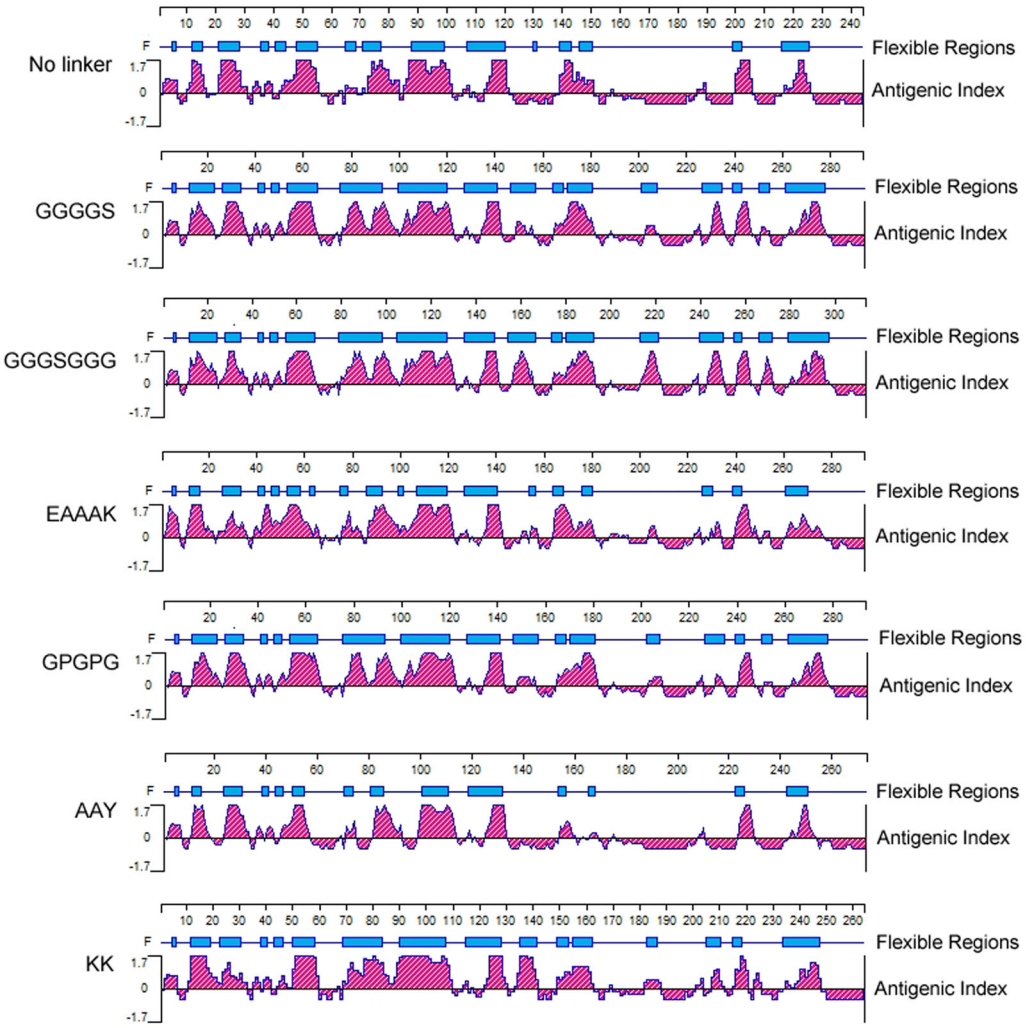

FIGURE 3 Analysis of flexibility and antigenic index of different vaccine candidate sequences

Vaccine design is a complex issue with many factors to consider, the most important of which is the safety and effectiveness of the vaccine.[38] When screening candidate epitopes in our study, non-synonymous mutation sites in the sequence were considered to ensure that the candidate epitopes did not contain easily mutated sites to avoid affecting antigen recognition.[7,8] The toxicity and allergenicity of epitopes were considered to ensure the safety of the epitopes.[39,40] The immunogenicity of antigens, the secretion of cytokines, the solvent accessibility of amino acids, and the recognition of MHC molecules were considered to ensure the effectiveness of the epitopes.[38,41–43] The coverage of epitopes in different populations was also considered to ensure the effectiveness of the epitopes in most populations.[44] Moreover, when expressing the fusion protein, choosing the appropriate linker is very important for the design of the vaccine candidate sequence. Different linkers have impacts on the correct folding, stability, biological activity, and immunogenicity of proteins.[45] These studies need a lot of experiments to verify. However, the application of immunoinformatics tools to help design vaccine has greatly improved the efficiency and accuracy of epitope screening and the rationality of vaccine design and has been applied to many vaccine research.[46,47]

In this study, 16 epitopes of spike protein were predicted to be B- and T-cell epitopes and selected as vaccine candidate epitopes for vaccine design. Among them, the epitope 371-387 partially overlapped with the CR3022 epitope of SARS-CoV-2 published in Science by Yuan et al.[28] CR3022 can neutralize SARS-CoV and is also able to interact with SARS-CoV-2.[28,29] Moreover, containing the 371-387 epitope in this study, epitope 375-394 was observed to stimulate robust secretion of IFN-γ from splenocytes.[48] Epitopes 375-394, 525-646, and 902-926 with an average positive rate of ≥ 50% (the percentage of convalescent sera from COVID-19 patients having positive reactions to the epitopes) among all 39 patients[48] contained or overlapped with epitope 371-387, 525-566, and 891-913 in this study. The study of Ferretti et al.[49] reported the epitopes of spike protein recognized by memory CD8+ T cells of patients with COVID-19 recovery. Among them, the 378-386 and 1208-1216 epitopes partially overlap with the 371-387 and 1213-1229 epitopes screened in this study. Therefore, some of the epitopes selected in

this study had been confirmed to be antigenic epitopes in other studies and showed immune effects. As the multi-epitope vaccine proposed in this study consists of less than 500 amino acids, we will consider connecting another strong immunogenicity peptide and choosing the appropriate vaccine vectors (such as adenovirus vectors), drug delivery systems (such as PAGL microspheres or liposomes, etc.) and adjuvants (such as TLR receptor adjuvants, etc.) to improve the immune effects of the vaccine.

In the previous results, we believed that six epitopes were important. Epitope 897-913, 899-906, and 1025-1041 epitopes were completely consistent with the sequences in the IEDB database. Epitope 371-387 and 379-395 partially overlapped with CR3022 epitope of SARS-CoV-2.[28] Epitope 371-387, 379-395, and 410-426 exist in the binding region of spike protein and ACE2.[28] However, only 371-387, 410-426, and 879-913 were finally selected as vaccine candidate epitopes. The reasons were that epitope 379-395 contained the easily mutated sites 382, the average solvent accessibility score of amino acids was only 15.48, and it recognized few HLA molecules. Epitope 899-906 had an average solvent accessibility score of only 4.06 and cannot recognize HLA class II molecules. Epitope 1025-1041 also had a low average solvent accessibility score (12.33) and cannot recognize HLA class I molecule. Therefore, these three epitopes were not selected as vaccine candidates.

Another interesting finding was that in the population coverage results of 28 epitopes, the coverage rate of each epitope was high in Europe, North America, East Asia, and Oceania, but low in East Africa, West Africa, South Africa, and Central Africa. We thought this was due to the differences in recognition of HLA molecules by different populations.[50] However, this difference might lead to people in Africa being less protected by the same vaccine than people in Europe, North America, East Asia, and Oceania. Whether it is necessary to prepare a specific vaccine based on the recognition ability of the African population to HLA subclasses in the future remains to be studied.

5 | CONCLUSIONS

According to the results of mutation and immunoinformatics analysis, we finally recommend 16 epitopes (194-210, 291-325, 307-323, 371-387, 410-426, 525-566, 530-544, 722-739, 747-763, 749-771, 754-770, 891-906, 897-913, 1101-1115, 1129-1145, and 1213-1229) of spike protein as SARS-CoV-2 vaccine candidate epitopes. In particular, epitope 897-913 was predicted to have possible cross-immunoprotection for SARS-CoV and SARS-CoV-2. The vaccine candidate sequences (without linker, with linker GGGSGGG, EAAAK, GPGPG, and KK, respectively) were predicted to be relatively stable, with a high antigenic index and good biological activity. We recommended the five sequences as candidate sequences for SARS-CoV-2 vaccine. Our next project is to synthesize the gene sequences for cloning and expression to prepare vaccines for SARS-CoV-2 and verify their immune effects. The bioinformatics analysis method in our study will greatly improve the accuracy and effectiveness of vaccine epitopes screening and the rationality of vaccine design, and can also be applied to vaccine design for other infectious diseases.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China to Jianping Chen (grant number 81672048), to Dali Chen (grant number 31872959 and 31572240), and to Jiao Li (grant number 31802184). Jinlei He is the recipient of the State Scholarship Fund supported by the China Scholarship Council (grant number 201706240018).

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

Conceptualization: Jinlei He, Jianping Chen, and Jiao Li. *Data curation*: Jianhui Zhang. *Formal analysis*: Jinlei He and Fan Huang. *Investigation*: Qiwei Chen and Dali Chen. *Methodology*: Jinlei He, Fan Huang, and Jianhui Zhang. *Project administration*: Zhiwan Zheng and Qi Zhou. *Supervision*: Jianping Chen and Jiao Li. *Writing-original draft*: Jinlei He and Fan Huang. *Writing-review & editing*: Jianping Chen and Jiao Li.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ORCID

Jianping Chen http://orcid.org/0000-0001-7617-2800

REFERENCES

1. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol*. 2020;94:e00127-20.
2. Buchholz UJ, Bukreyev A, Yang L, et al. Contributions of the structural proteins of severe acute respiratory syndrome coronavirus to protective immunity. *Proc Natl Acad Sci U S A*. 2004;101:9804-9809.
3. Xu X, Gao X. Immunological responses against SARS-coronavirus infection in humans. *Cell Mol Immunol*. 2004;1:119-122.
4. Lucchese G. Epitopes for a 2019-nCoV vaccine. *Cell Mol Immunol*. 2020;17:539-540.
5. Vita R, Mahajan S, Overton JA, et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. 2019;47(D1):D339-D343.
6. Xu X, Chen P, Wang J, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci China: Life Sci*. 2020;63:457-460.
7. Karimzadeh H, Kiraithe MM, Oberhardt V, et al. Mutations in hepatitis D virus allow it to escape detection by CD8+ T cells and evolve at the population level. *Gastroenterology*. 2019;156:1820-1833.
8. Magnusson SE, Altenburg AF, Bengtsson KL, et al. Matrix-M™ adjuvant enhances immunogenicity of both protein- and modified vaccinia virus Ankara-based influenza vaccines in mice. *Immunol Res*. 2018;66:224-233.
9. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*. 2007;8:4.

10. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R, Raghava GPS. In silico approach for predicting toxicity of peptides and proteins. *PLOS One*. 2013;8:e73957.
11. Dimitrov I, Naneva L, Doytchinova I, Bangov I. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics*. 2014;30:846-851.
12. Dhanda SK, Vir P, Raghava GP. Designing of interferon-gamma inducing MHC class-II binders. *Biol Direct*. 2013;8:30.
13. Dhanda SK, Gupta S, Vir P, Raghava GPS. Prediction of IL4 inducing peptides. *Clin Dev Immunol*. 2013;2013:263952-263959.
14. Nagpal G, Usmani SS, Dhanda SK, et al. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci Rep*. 2017;7:42851.
15. Cai Y, Zhang J, Xiao T, et al. Distinct conformational states of SARS-CoV-2 spike protein. *Science*. 2020;369(6511):1586-1592.
16. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296-W303.
17. Saha S, Raghava GP. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*. 2006;65:40-48.
18. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*. 2017;45(W1):W24-W29.
19. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*. 2016;32:511-517.
20. Reche P, Glutting JP, Zhang H, Reinherz E. Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics*. 2004;56:405-419.
21. Schuler MM, Nastke MD, Stevanovikć S. SYFPEITHI: database for searching and T-cell epitope prediction. *Methods Mol Biol*. 2007;409:75-93.
22. Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics*. 2006;7:153.
23. Yachdav G, Kloppmann E, Kajan L, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res*. 2014;42(Web Server issue):W337-W343.
24. Burland TG. DNASTAR's Lasergene sequence analysis software. *Methods Mol Biol*. 2000;132:71-91.
25. Gasteiger E, Hoogland C, Gattiker A, et al. Protein identification and analysis tools on the ExPASy server. In: Walker JM, ed. *The Proteomics Protocols Handbook*. Geneva: Humana Press; 2005:571-607.
26. Guo JP, Petric M, Campbell W, McGeer PL. SARS corona virus peptides recognized by antibodies in the sera of convalescent cases. *Virology*. 2004;324:251-256.
27. He Y, Zhou Y, Wu H, et al. Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (SARS) coronavirus: implication for developing SARS diagnostics and vaccines. *J Immunol*. 2004;173:4050-4057.
28. Yuan M, Wu NC, Zhu X, et al. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. *Science*. 2020;368:630-633.
29. ter Meulen J, van den Brink EN, Poon LLM, et al. Human monoclonal antibody combination against SARS coronavirus: synergy and coverage of escape mutants. *PLOS Med*. 2006;3:e237.
30. Korber B, Fischer WM, Gnanakaran S, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *BioRxiv*. 2020.
31. Choi Y, Agarwal S, Deane CM. How long is a piece of loop? *PeerJ*. 2013;1:e1.
32. WHO. Draft landscape of COVID-19 candidate vaccines. 2020. https://www.who.int/who-documents-detail/draft-landscape-of-covid-19-candidate-vaccines
33. Vetter V, Denizer G, Friedland LR, Krishnan J, Shapiro M. Understanding modern-day vaccines: what you need to know. *Ann Med*. 2018;50:110-120.
34. Kardani K, Bolhassani A, Shahbazi S. Prime-boost vaccine strategy against viral infections: mechanisms and benefits. *Vaccine*. 2016;34:413-423.
35. Lu S. Heterologous prime-boost vaccination. *Curr Opin Immunol*. 2009;21:346-351.
36. Schulze K, Staib C, Schätzl HM, Ebensen T, Erfle V, Guzman CA. A prime-boost vaccination protocol optimizes immune responses against the nucleocapsid protein of the SARS coronavirus. *Vaccine*. 2008;26:6678-6684.
37. Kobinger GP, Figueredo JM, Rowe T, et al. Adenovirus-based vaccine prevents pneumonia in ferrets challenged with the SARS coronavirus and stimulates robust immune responses in macaques. *Vaccine*. 2007;25:5220-5231.
38. Zepp F. Principles of vaccination. *Methods Mol Biol*. 2016;1403:57-84.
39. Forster R. Study designs for the nonclinical safety testing of new vaccine products. *J Pharmacol Toxicol Methods*. 2012;66:1-7.
40. Konstantinou GN. T-cell epitope prediction. *Methods Mol Biol*. 2017;1592:211-222.
41. Deng H, Yu S, Guo Y, et al. Development of a multivalent enterovirus subunit vaccine based on immunoinformatic design principles for the prevention of HFMD. *Vaccine*. 2020;38:3671-3681.
42. Zheng W, Ruan J, Hu G, Wang K, Hanlon M, Gao J. Analysis of conformational B-cell epitopes in the antibody-antigen complex using the depth function and the convex hull. *PLOS One*. 2015;10(8):e0134835.
43. Bartlett BL, Pellicane AJ, Tyring SK. Vaccine immunology. *Dermatol Ther*. 2009;22:104-109.
44. Oyarzún P, Kobe B. Recombinant and epitope-based vaccines on the road to the market and implications for vaccine design and production. *Hum Vaccin Immunother*. 2016;12:763-767.
45. Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. *Adv Drug Deliv Rev*. 2013;65(10):1357-1369.
46. 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;92:495-500.
47. He J, Huang F, Li J, Chen Q, Chen D, Chen J. Bioinformatics analysis of four proteins of *Leishmania donovani* to guide epitopes vaccine design and drug targets selection. *Acta Trop*. 2019;191:50-59.
48. Zhang B, Hu Y, Chen L, et al. Mining of epitopes on spike protein of SARS-CoV-2 from COVID-19 patients. *Cell Res*. 2020;30(8):702-704.
49. Ferretti AP, Kula T, Wang Y, et al. COVID-19 patients form memory CD8+ T cells that recognize a small set of shared immunodominant epitopes in SARS-CoV-2. *MedRxiv*. 2020.
50. Dos Santos Francisco R, Buhler S, Nunes JM, et al. HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms. *Immunogenetics*. 2015;67:651-663.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: He J, Huang F, Zhang J, et al. Vaccine design based on 16 epitopes of SARS-CoV-2 spike protein. *J Med Virol*. 2021;93:2115-2131. https://doi.org/10.1002/jmv.26596