Journal of Infection and Public Health 14 (2021) 169–178

Contents lists available at ScienceDirect

Journal of Infection and Public Health

journal homepage: http://www.elsevier.com/locate/jiph

Bioinformatics prediction of B and T cell epitopes within the spike and nucleocapsid proteins of SARS-CoV2

Reham M. Dawood[a], Mai A. El-Meguid[a], Ghada M. Salum[a], Khaled El-Wakeel[b], Mohamed Shemis[c], Mostafa K. El Awady[a,*]

[a] Department of Microbial Biotechnology, Genetic Engineering Division, National Research Centre, 33 EL Bohouth Street, Dokki, Giza 12622, Egypt
[b] Biological Anthropology Department, Medical Research Division, National Research Centre, Dokki, Giza, Egypt
[c] Department of Biochemistry and Molecular biology, Theodor Bilharz Research Institute, Egypt

ARTICLE INFO

Article history:
Received 26 June 2020
Received in revised form 1 September 2020
Accepted 2 December 2020

Keywords:
SARS CoV2
Furin cleavage site
Genetic distance
NH2 Terminal Domain
Phylogenetic tree

ABSTRACT

Background: The striking difference in severity of SARS CoV2 infection among global population is partly attributed to viral factors. With the spike (S) and nucleocapsid (N) are the most immunogenic subunits, genetic diversity and antigenicity of S and N are key players in virulence and in vaccine development.
Aim: This paper aims at identifying immunogenic targets for better vaccine development and/or immunotherapy of COVID 19 pandemic.
Methods: 18 complete genomes of SARS CoV2 ($n = 14$), SARS CoV ($n = 2$) and MERS CoV ($n = 2$) were examined. Bioinformatics of viral genetics and protein folding allowed functional tuning of NH2 Terminal Domain (NTD) of S protein and development of epitope maps for B and T cell responses.
Conclusion: A deletion of amino acid residues Y144 and G107 were discovered in NTD of S protein derived from Indian and French isolates resulting in altered pocket structure exclusively located in NTD and reduced affinity of NTD binding to endogenous nAbs and disrupted NTD mediated cell entry. We therefore, proposed a set of B and T cell epitopes based on Immune Epitope Database, homologous epitopes for nAbs in convalescent plasma post SARS CoV infection and functional domains of S (NTD, Receptor Binding domain and the unique polybasic Furin cleavage site at S1/S2 junction). Nevertheless, laboratory data are required to develop vaccine and immunotherapeutics.

Introduction

Currently seven pathogenic corona viruses have been isolated from humans. Four cause mild respiratory diseases; HKU1, NL63, OC43 and 229E [1,2], whereas SARS CoV, MERS CoV and the recently discovered SARS CoV2 are associated with severe pneumonia symptoms and higher rates of morbidity and mortality [3,4]. The alarmingly fast human to human transmission resulted in more than 23 million people infected and high fatality of 3.5% of infections in more than 200 countries [5].

The infection rate is extremely higher in four main areas of the globe; Asia, Middle East, Europe and North America [6]. SARS CoV2 has an RNA genome size of ~30 Kilobases, encoding, like other members of the family coronaviridae, structural and nonstructural proteins [7]. Structural proteins include the spike (S), the nucleocapsid (N), the membrane (M) and the envelope (E) [8]. The N protein holds the RNA genome, while M, E and S glycoproteins make up the viral envelope [9]. For coronaviruses family, there are four recognized protein receptors called peptidases: angiotensin converting enzyme 2 (ACE2), aminopeptidase N (APN), and dipeptidyl peptidase 4 (DPP4) [10]. The spike comprising 2 subunits S1 and S2 in each monomer to bind the permissive cell receptor (ACE2) [11]. Moreover, previous research showed that the corona viruses family would attach to certain molecules on the host cell surface, including glycosamineglycan (GAGs) as molecules, to promote the association with the target cell [12]. Recent studies using cryogenic electron microscopy (Cryo-EM) revealed that binding to cell receptor allows dissociation of S1 and S2 at the furin sensitive cleavage site PRRA at residues 683–686 located at S1-S2 junction discovered exclusively in SARS CoV2 [13]. The S subunit cleavage is followed by association of S1 and ACE2 [14] and a cascade of events leading to cell damage. Recent studies pointed to the significance of binding of spike to ACE2 receptors as a critical initial step toward the pathogenesis of SARS CoV2 [13,15]. Structural proteins of SARS CoV and

[*] Corresponding author.
E-mail address: mkawady49@yahoo.com (M.K. El Awady).

https://doi.org/10.1016/j.jiph.2020.12.006

MERS CoV are known to stimulate nAbs and cell mediated immune responses [16]. The S protein is the most exposed protein of SARS CoV so that antibody responses to S protein have been shown to protect mouse models from infection [17,18]. In addition, several studies on SARS CoV infected patients revealed that N protein is amply expressed during infection resulting in abundant immunogenicity and generation of plentiful amounts of anti N antibodies during the infection [19].

Given the level of genetic distance in the N and the S genes among various geographical isolates, it is likely that sequences of N and S proteins of SARS CoV2 viruses discovered in diverse geographical locations will determine, at least in part, the transmission rate and severity of SARS CoV2 infections in each population. Apart from the roles of host factors, it is expected that sequence variations in S and N the most exposed, immunogenic and abundantly expressed SARS CoV2 proteins are likely to clench disease virulence. Here we designed phylogenetic trees and genetic distances of the complete genome, S and the N sequences of SARS CoV2 isolates derived from populations ranging in severity of infection from mild to severe. As of August, 24 2020, the infection rates (IR) in the selected population ranged from the highest to the lowest as follows; Spain (0.87%), Italy (0.43%), France (0.37%), USA (1.79%), Israel (1.15%), Iran (0.44%),Japan (0.049%), Egypt (0.099%), and India (0.229%). The death rates/infections (DR) were in Italy (13.67%), France (12.56%). Spain (7.07%), Egypt (5.4%), Iran (5.75%), USA (3.07%), India (1.86%) and Israel (0.81%). Further, the convalescence rates/infections (CR) were: Iran (86%), Spain (39%), Egypt (67.7%), Italy (79%), France (35%), Israel (77.9%), India (75.3%) and USA (53.9%) [20]. The genetic distance of S and N nucleotide sequences and amino acid alignments among severe versus moderate infections as indicated by death and convalescence rates, e.g. France vs. India may presumably highlight mutational events which might define, at least partly, different levels of disease severity. The immunogenic responses and epitope maps account for severity of infection [21] and narrow down the strategic make up for successful vaccine against SARS CoV2.

Strategies for vaccine against SARS CoV2 attempt to elicit neutralizing antibodies as well as CD4$^+$ T cells [22]. The S1 subunit is more exposed at the viral surface than S2 subunit and therefore, is likely to be under selection pressure of the immune system thus facilitating less genetic conservation than S2 [23].

Bioinformatics based identification of SARS CoV2 derived B cell and T cell epitopes within S and N proteins were searched on the Immune Epitope Database IEDB (http://tools.iedb.org/tepitool/) [24]. To further narrow down the selection process to experimentally efficient B cell and T cell epitopes, we focused on those in silico determined epitopes that are identical to the available experimentally – determined SARS CoV B cell and T cell epitopes.

Methods

Phylogenetic tree

Fourteen full-length viral genomic sequences of SARS-CoV2, 2 MERS and 2 SARS CoV were obtained from the Genbank and analyzed by the MEGA program version 10.8 [25] to provide DNA-based phylogenetic tree; The evolutionary history was indicated by using the neighbour-joining method and the evolutionary distances were calculated using the Maximum Composite Likelihood method. The isolates include 14 SARS CoV2, 2 MERS CoV and 2 SARS CoV complete genomes. SARS CoV2 genomes were derived from USA (3 isolates), Italy (2 isolates, China (3 isolates) and one isolate from each of Spain, France, Japan, Iran, India and Israel. Details of the isolates are shown in Table 1. The phylogenetic trees for the nucleotide sequences of SARS CoV2 complete genomes, spike and nucleocapsid genes were carried out by using the molecular evolutionary genetics analysis program MEGA X. The genetic distance was performed only for the S and N genes.

Spike protein alignment

Multiple alignment of S protein sequences were performed using constraint-based multiple alignment tool through NCBI. The data obtained from the alignment of the S proteins derived from the selected SARS CoV2 isolates revealed complete similarity among the studied strains except for strains derived from France (Accession ID # MT320538.1, IR 0.37%; DR 12.56%; CR 35%; where IR = number of infections/country‘s population, DR = number of deaths/number of infections and CR = number of convalescence/number of infections) and India (Accession ID # MT012098.1, IR 0.23%; DR 1.86%; CR 75.3%), where S shows great genetic distance from the rest of SARS CoV2 S genes. Interestingly, 2 amino acid deletions were noted within the NTD of the S protein derived from the French and the Indian strains (G107 and Y144 respectively).

Protein modeling

The S protein sequences for SARS CoV2 isolates derived from Wuhan, India and France were obtained from GenBank with the following accession IDs NC_045512, MT012098.1, and MT320538.1 respectively. Phyre2 is a protein modeling program available on the web (http://www.sbg.bio.ic.ac.uk/phyre2). The phyre2 program can predict and analyze protein structure, function and mutations [26,27]. Phyre2 uses HMM-HMM (*h*idden *M*arkov *m*odel) for performing multiple protein alignments to compare a protein sequence to proteins in a fold library. If a match can be found, the query structure is modeled on the matching structure. Phyre2 uses advanced remote homology detection methods to build 3D models, predict ligand binding sites and analyze the effect of amino acid variants for a submitted protein sequence.

The Epitope mapping

B Cell Epitope selection (linear epitopes)

On the basis of the published genome sequence of the SARS CoV2 (GenBank: MT072688.1), we downloaded structural proteins S, and N into the Immune Epitope Database (IEDB) (http://tools.iedb.org/tepitool/) and analysis software to examine the physical parameters of the proteins, such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity and antigenic propensity of polypeptide chains. These have been correlated with the location of continuous epitopes (Fig. 3a and b).

T Cell epitopes

Peptides for CTL epitopes were selected based on bioinformatics provided by the Immune Epitope Database (IEDB) (http://tools.iedb.org/tepitool/) software. For the identified T cell epitopes, additional information about the associated MHC alleles were provided to seek maximization of global population coverage (Fig. 3).

Results

Phylogenetic trees and genetic distances

Total of 18 complete genome sequences were aligned and phylogenetic tree was constructed (Fig. 1a). Nucleotide sequences of S and N derived from various isolates of SARS-CoV2 were aligned and the phylogenetic tree was constructed (Fig. 1b and c) using the same parameters. Fig. 1a shows that Indian (MT012098.1) and Chinese (MN998531.1) isolates as well as Japanese (LC 534418.1) and French (MT320538.1) isolates are closely related and the four

Table 1
Designations, lengths and accession IDs of SARS CoV2, SARS CoV and MERS CoV complete genomes. Genome sequences of SARS CoV2 were derived from isolates of diverse populations covering four geographical areas worldwide; Asia, Middle East, Europe and North America.

	Accession number	S protein	N protein	Genome sequence
1	NC_045512	21,563.25384	28,274.29533	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
2	MT077125.1	21,507.25328	28,218.29477	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/INMI1/2020, complete genome
3	MT066156.1	21,563.25384	28,274.29533	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/INMI1/2020, complete genome
4	LC534418.1	21,559.25380	28,270.29529	Severe acute respiratory syndrome coronavirus 2 SARS-CoV-2/Hu/DP/Kng/19-031 RNA, complete genome
5	MN996531.1	21,550.25371	28,261.29520	Severe acute respiratory syndrome coronavirus 2 isolate WIV07, complete genome
6	MT012098.1	21,550.25368	28,258.29517	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/IND/29/2020, complete genome
7	MT020880.1	21,563.25384	28,274.29533	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-WA1-A12/2020, complete genome
8	MN938384.1	21,531.25352	28,242.29501	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-SZ-002a_2020, complete genome.
9	MT027064.1	21,563.25384	28,274.29533	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV/USA-CA5/2020, complete genome
10	MT292569.1	21,509.25330	28,220.29479	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ESP/Valencia13/2020, complete genome
11	MT304491.1	21,563.25384	28,274.29533	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/TX_2967/2020, complete genome
12	MT320538.1	21,562.25380	28,270.29529	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/FRA/KRA-ROB/2020, complete genome
13	MT320891.1	21,525.25346	28,236.29495	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/IRN/HGRC-1.1-IPI-8206/2020, complete genome
14	MT276598.1	21,567.25388	28,278.29537	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ISR/ISR_IT0320/2020, complete genome
15	KT225476.2	21450.25511	28,560.29801	Middle East respiratory syndrome coronavirus isolate **MERS**-CoV/THA/CU/17_06_2015, complete genome
16	KJ556336.1	21,240.25301	28,350.29591	Middle East respiratory syndrome coronavirus isolate **MERS** Jeddah_1_2013, complete genome
17	AY274119.3	21,492.25259	28,120.29388	**SARS**-related coronavirus isolate Tor2, complete genome
18	NC_006577.2	22,942.27012	28,320.29645	**SARS** Human coronavirus HKU1, complete genome

strains share common ancestor. Also the three USA (MT027064.1, MT 020880.1, MT 304491.1) isolates together with the Italian isolate (MT 066156.1) share great homology to the reference strain derived from Wuhan isolate (NC045512.2) (Fig. 1b and c).

The genetic distances of the S and N proteins in the selected populations are displayed in Tables 2 and 3 respectively. These showed that the N protein, unlike the S protein, appears less diverse than the S protein and shows great similarity among N proteins derived from SARS CoV2 isolates published from several global populations. The Japanese isolate (LC534418.1) is identical to all selected strains except for slight variations from Israeli (MT276598.1) and Persian (MT320891.1) strains. The greatest dissimilarity is noted with MERS (KJ556336.1, and KT225476.2) and SARS CoV (AY274119.3 and NC_006577.2) isolates described in this study. The Chinese WIV07 (MN996531.1) is identical to all selected strains except, like the Japanese isolate, for Israeli (MT276598.1) and Persian (MT320891.1) isolates.

On the other hand the S protein displays greater diversity in genetic distance between virulent infections (e.g. France) and others with less virulence (e.g. India). Further, genetic distance of the Indian strain is extremely far from the mild strain HKU1 (NC_006577.2) and SARS CoV (AY274119.3) sequences. On the other hand, S protein in Indian isolate displayed the same distance with S isolates from Iran (MT320891.1), Wuhan (NC_045512), the two Italian isolates (MT077125.and MT066156.1) and an isolate derived from USA (MT027064.1). Besides, similar distances were observed with the Japanese DP/Kng (LC534418.1) and the Chinese WIVO7 (MN996531.1) isolates. The S protein derived from the Italian isolate (MT066156.1) displayed very far distance from HKU1 (NC_006577.2) and SARS CoV (AY274119.3), followed by the French strain (MT320538.1). On the other hand, it showed close distance with all studied isolates described in this study, with the exception of HKU1 and SARS CoV which showed the furthest distance from Italian strain. The S protein derived from the Spanish (MT292569.1) isolate showed a short genetic distance from India, Iran, Wuhan and Japan. On the other hand the Spanish strain is far distant genetically from HKU1, SARS CoV and MERS Co. The Japanese strain is genetically far distant from both SARS CoV and MERS. Moderate genetic distance was noticed between S protein in Japanese isolate and each of the Indian, French, Spanish and CA5 USA (MT027064.1) isolates, whereas the closest strains are: WIVO7 China, WA1 USA (MT020880.1), Italy (MT066156.1 and MT077125), Iran and Wuhan.

Amino acid alignment and protein folding

Amino acid sequence of the N protein is identical among all SARS CoV2 N sequences (results not shown). Multiple amino acid alignment for NTD of S proteins in diverse populations are shown in (Fig. 2a). Contrary to the great similarity of N protein sequences among studied populations, identity of the S protein derived from the same diverse population provides support to the genetic findings derived from phylogenetic trees and genetic distance described herein. Amino acid alignment of SARS CoV2 isolates revealed that the S protein derived from France (MT320538.1) had a deletion of Glycine residue at position 107 (G 107). Further, a deletion of tyrosine residue at amino acid 144 (Y 144) was observed in the S protein derived from the Indian strain (MT012098.1). The impact of these deletions on S protein 3D structure is illustrated in Fig. 2b and c. For example, residues Y 144, Y 145 and V 146

Table 2

Genetic distance of SARS CoV2, SARS CoV and MERS CoV spike proteins between different populations. The numbers reflect the relative inter-relation between strains on the *X*-axis with 14 strains on the *Y* axis. The smaller the number indicates the greater sequence identity between any 2 isolates.

	LC534418.1, CV2 Japan	KJ556336.1_MERS_Jeddah	MN996531.1_CV2 Wuhan	MT012098.1_CV2_India	MT020880.1_CV2_USA WA	MT027064.1_CV2_USA CA	MT066156.1_CV2_Italy	MT077125.1_CV2_Italy	MT276598.1_CV2_Israel	MT292569.1_CV2_Espagne	MT320538.1_CV2_France	MT320891.1_CV2_Iran	NC_045512.2_CV2_Wuhan	AY274119.3_CV_Tor2
KJ556336.1_MERS_Jeddah	1.080													
MN996531.1_CV2 China	0.000	1.080												
MT012098.1_CV2_India	0.000	1.081	0.000											
MT020880.1_CV2_USA WA	0.000	1.080	0.000	0.000										
MT027064.1_CV2_USA CA	0.000	1.081	0.000	0.001	0.000									
MT066156.1_CV2_Italy	0.000	1.080	0.000	0.000	0.000	0.000								
MT077125.1_CV2_Italy	0.000	1.080	0.000	0.000	0.000	0.000	0.000							
MT276598.1_CV2_Israel	0.000	1.080	0.000	0.001	0.000	0.001	0.000	0.000						
MT292569.1_CV2_Espagne	0.000	1.080	0.000	0.001	0.000	0.001	0.000	0.000	0.000					
MT320538.1_CV2_France	0.001	1.080	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000				
MT320891.1_CV2_Iran	0.000	1.080	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001			
NC_045512.2_CV2_Wuhan	0.000	1.080	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000		
AY274119.3_CV_Tor2	0.470	1.342	0.470	0.469	0.470	0.469	0.470	0.470	0.470	0.470	0.470	0.470	0.470	
NC_006577.2_CV_HKU1	1.115	1.394	1.115	1.116	1.115	1.114	1.115	1.115	1.116	1.116	1.116	1.115	1.115	1.353

Table 3

Genetic distance of SARS CoV2, SARS CoV and MERS CoV nucleocapsid proteins between different populations. The numbers reflect the relative inter-relation between strains on the *X*-axis with 14 strains on the *Y* axis. The smaller the number indicates the greater sequence identity between any 2 isolates.

	LC534418.1 _CV2 Japan	KJ556336.1 _MERS Jeddah	MN996531.1 _CV2 Wuhan	MT012098.1 _CV2_India	MT020880.1 _CV2_USA WA	MT027064.1 _CV2_USA CA	MT066156.1 _CV2_Italy	MT077125.1 _CV2_Italy	MT276598.1 _CV2_Israel	MT292569.1 _CV2_Espagne	MT320538.1 _CV2_France	MT320891.1 _CV2_Iran	NC_045512.2 _CV2_Wuhan	AY274119.3 _CV_Tor2
KJ556336.1 _MERS_Jeddah	8.850													
MN996531.1_ CV2 Wuhan	0.000	8.850												
MT012098.1_ CV2_India	0.000	8.850	0.000											
MT020880.1_ CV2_USA WA	0.000	8.850	0.000	0.000										
MT027064.1_ CV2_USA CA	0.000	8.850	0.000	0.000	0.000									
MT066156.1_ CV2_Italy	0.000	8.850	0.000	0.000	0.000	0.000								
MT077125.1_ CV2_Italy	0.000	8.850	0.000	0.000	0.000	0.000	0.000							
MT276598.1_ CV2_Israel	0.002	8.887	0.002	0.002	0.002	0.002	0.002	0.002						
MT292569.1_ CV2_Espagne	0.000	8.850	0.000	0.000	0.000	0.000	0.000	0.000	0.002					
MT320538.1_ CV2_France	0.000	8.850	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000				
MT320891.1_ CV2_Iran	0.001	8.850	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.001	0.001			
NC_045512.2_ CV2_Wuhan	0.000	8.850	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.001		
AY274119.3_ CV_Tor2	0.192	9.113	0.192	0.192	0.192	0.192	0.192	0.192	0.199	0.192	0.192	0.194	0.192	
NC_006577.2_ CV_HKU1	11.335	13.175	11.335	11.335	11.335	11.335	11.335	11.335	11.358	11.335	11.335	11.335	11.335	12.393

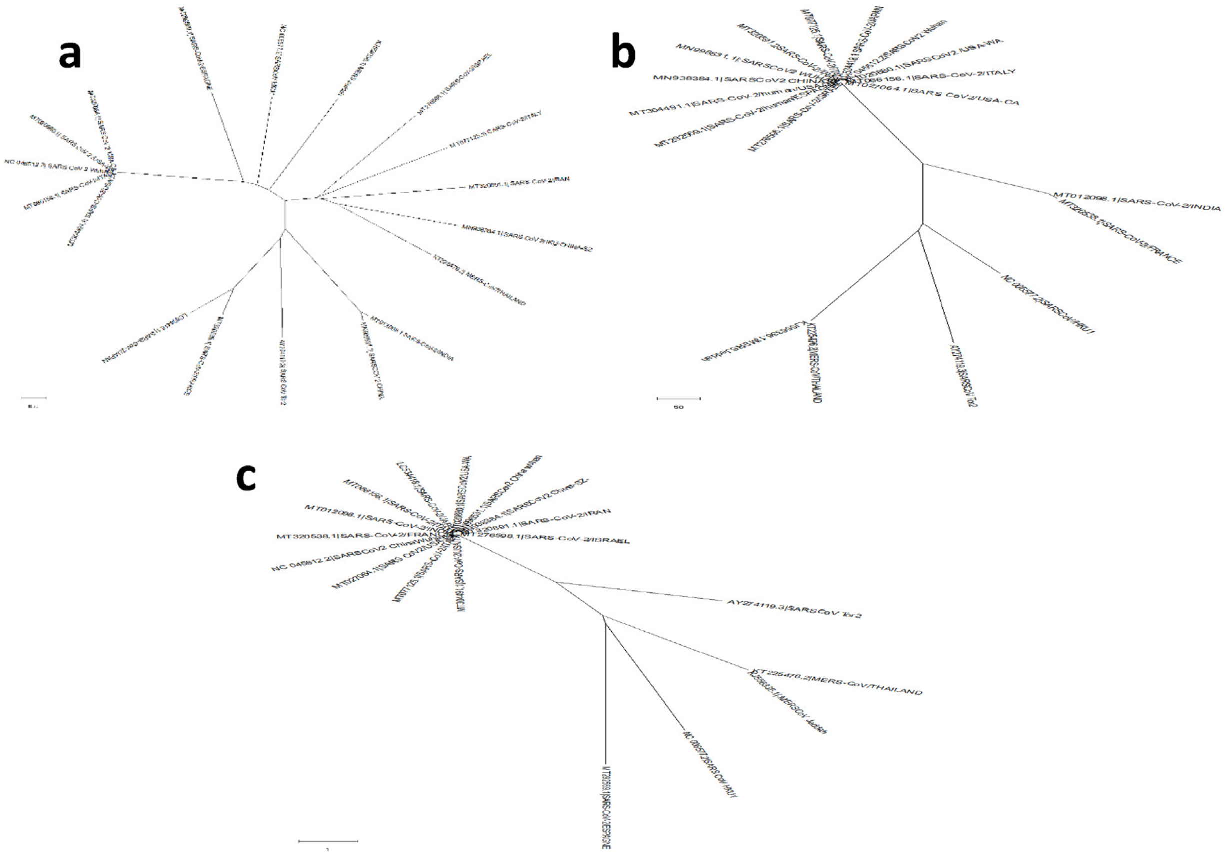

Figure 1. Radiation Phylogram of the phylogenetic tree showing comparison of the similarity, among diverse populations, of complete genomes (a) Spike (b) and nucleocapsid (c) of SARS CoV2. The tree was constructed based on the available diverse sequences using MEGA program (see Methods section).

form a conserved pocket in the NTD of S1 subunit in Wuhan reference strain (NC_045512). However, in the French isolate, the G107 deletion is linked with conformational changes in the pocket pattern in the entire NTD of the S protein. The YYV pocket at position (144–146) is replaced by a pocket made of only two residues, i.e. Y144 and H145. Also in the Indian strain with a deletion of Y144, similar to the French strain, a new two amino acid pocket YH at position 144 has become a part of the NTD specific large pocket. Besides, de novo pockets were created at several positions e.g. amino acid residue 19 that did not exist in the French or Wuhan isolates. The data of Fig. 2b showed that the above mutations in SARS CoV2 S from France and India created different changes in the large pocket structure that are apparently distinct from isolates derived from other studied populations. Fig. 2c shows the 3 D structure of the S protein in the reference Wuhan strain. The large pocket (red) is seen at the interface with the NTD (blue) and represents a target for binding to small molecules as well as for binding to nAb that might act on disrupting the cascades of events leading to interference with binding of the S to its cellular receptor ACE2.

Epitope mapping

Epitopes for B and T cell immunity are illustrated in Fig. 3. A total of 60 peptides ranging in size from 20 to 22 amino acid residues were determined by IEDB software (described in Methods section). Then 30 peptides out of the 60 were selected based on the similarity with the experimentally-determined epitopes and displaying high antigenicity in SARS CoV1 convalescent patients. Further, the 30 epitopes comprised amino acid sequences located at the S1/S2 boundaries known as polybasic domain. The latter is a cleavage site sensitive to furin and other host proteases and is considered in epitope selection process. Other functional epitopes within RBD and NTD were also selected. Also studies on MERS CoV revealed that NTD of the S protein contains specific loop structure for binding to G2, a potent murine nAb targeting MERS CoV S protein. This binding strongly disrupted the attachment of MERS-CoV S to its receptor dipeptidyl peptidase-4 (DPP4). However, the NTD of SARS CoV2 does not contain similar epitope sequence to the G2 antibody specific epitope described in MERS CoV virus.

The T cell epitopes are distributed all over the S protein. However 3 out of the 6 T cell epitopes are located in the near vicinity of B cell epitopes at amino acids 39 (S1 NTD), 412 (S1 RBD) and 1007 (S2). It is interesting to note that the T cell epitope at amino acid 140 lies very closely to the Y 144 deletion described in the Indian isolate and an important 3 amino acid pocket structure VYY characteristic of the Wuhan reference strain. The latter pocket has been modified to an altered pocket structure in Indian and French strains which points to the functional importance of this epitope.

Discussion

To examine the hypothesis that differences in amino acid sequences of the S reflect variations in the severity of SARS CoV2 infection among the studied populations, the genetic distance between S protein sequences in the reference Wuhan strain was compared to the less virulent population such as Japan or India (expressed as lower infection rate (IR) and death rate (DR) associated with higher convalescence rate (CR); Accession ID # MT012098.1, IR 0.23%; DR 1.86% and CR 75.3%) when compared with virulent strains as in Spain, Italy, USA and France (Accession ID # MT320538.1, IR 0.37%; DR 12.56%; CR 35%). The results shown in figure (1b) and Table 2 revealed that phylogenetic and genetic

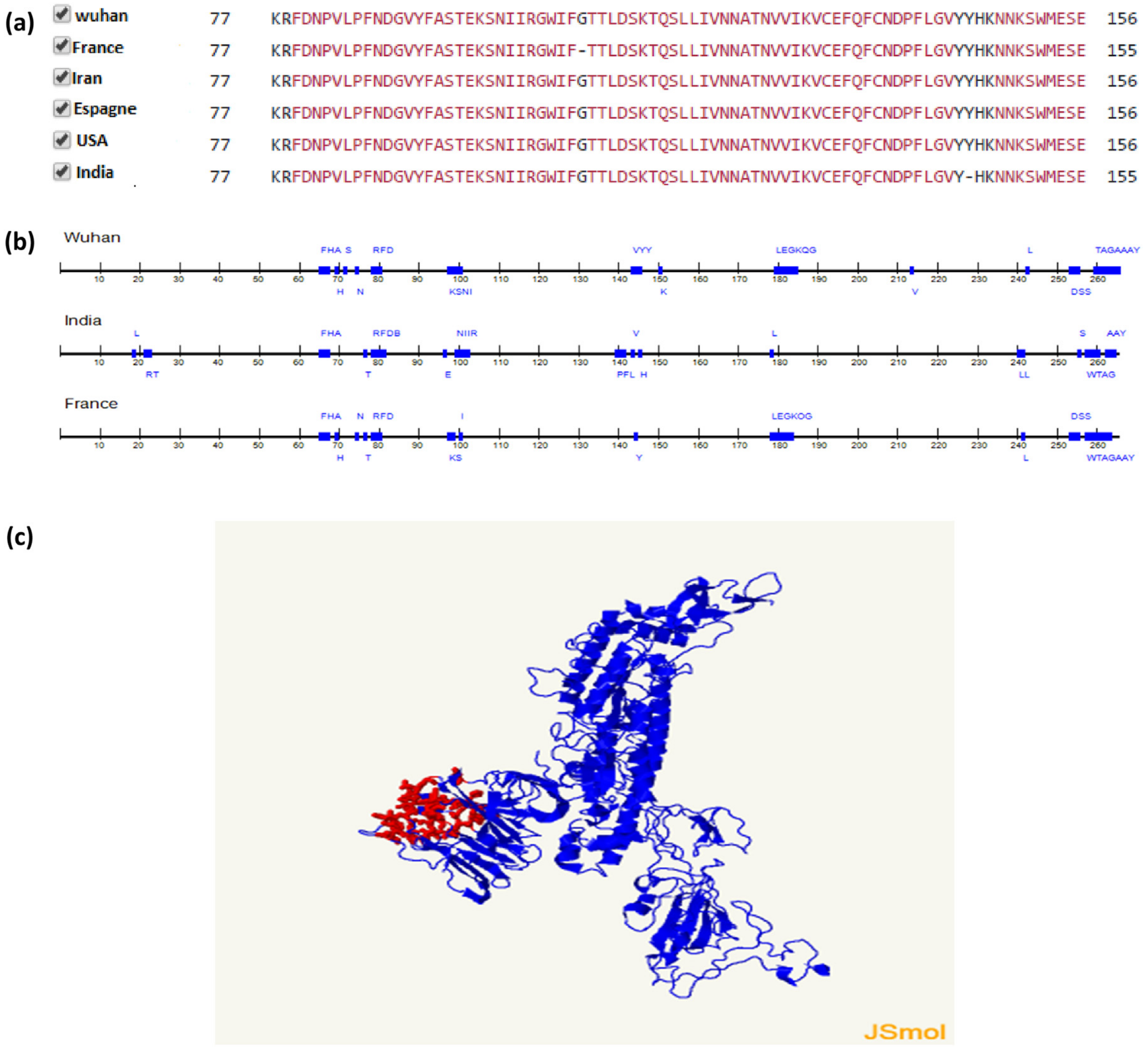
(a)
wuhan 77 KRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESE 156
France 77 KRFDNPVLPFNDGVYFASTEKSNIIRGWIF-TTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESE 155
Iran 77 KRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESE 156
Espagne 77 KRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESE 156
USA 77 KRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESE 156
India 77 KRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVY-HKNNKSWMESE 155
(b)
Wuhan
FHA S RFD VYY LEGKQG L TAGAAAY
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260
H N KSNI K V DSS
India
L FHA RFDB NIIR V L S AAY
RT T E PFL H LL WTAG
France
FHA N RFD I LEGKOG DSS
H T KS Y L WTAGAAY
(c)
JSmol

Fig. 2. (a) Multiple amino acid alignment between NH2-Terminal Domain of SARS CoV2 S proteins derived from 6 diverse populations covering the most stricken 3 continents Asia, Europe and North America. Homologous sequences are shown in red while mutations are indicated in black. Multiple alignments were performed using Cobalt software. (b) Pocket pattern of The NH2-Terminal Domain of the S protein. Pattern of amino acid pockets that make up the largest pocket at the farther end of the NTD derived from Wuhan reference strain as compared with the pocket pattern resulted from the Y144 and G107 deletions in the NTD of the India (accession ID # MT012098.1) and France (accession ID # MT320538.1) SARS CoV2 isolates respectively. The amino acid pockets were identified using Phyre2 program (see Methods section). (c) The 3 D models of S proteins in the reference strain derived from Wuhan China. Complete sequence of the S protein (blue) derived from the reference strain Wuhan (accession ID # NC_045512), was developed using Phyr2 program. The largest pocket (neutralizing antibody binding loop) is located at the end of the NH2-Terminal Domain (red).

distance of S gene in the Indian isolate is farthest from the French isolate followed by WA1 USA and Spanish isolates. Multiple alignment of amino acid sequence of S protein derived from the current diverse population supported the phylogenetic and genetic distance data and revealed that only the Indian and the French isolates proteins displayed single amino acid deletions at Y144 and G107 respectively. These deletions resulted in significant changes in the pattern and structure of the large pocket located exclusively in the NTD of the S protein. Such changes in the pocket structure most likely explain, at least in part, the difference in the severity and virulence of infection between the above two distinct populations. The exact role of difference in pocket structure of SARS CoV2 S protein in mutant strains derived from France and India requires excessive wet laboratory studies.

All the pockets of the S protein in Wuhan reference strain as well as the strains derived from other populations are identically located within the NTD between amino acid residues 65 and 265 and folded to form the largest pocket at the end of S protein NTD.

The S1-NTD of corona viruses comprise 3 layer structure; Core, top and bottom. The top layer binds proteins or glycan receptors; the bottom binds to the COOH-Terminal Domain (CTD). Both make a sandwich around a galectin-like core [28–29]. Studies on the function of MERS CoV S1-NTD have shown that NTD primarily mediates interaction with several host factors such as sialic acid prior to binding to its cell receptor DPP4 [30]. The Fab of murine nAb G2 was shown to interact with loops 1 and 2 on the interface of MERS CoV S1-NTD and to protect against a lethal infection of mouse model [28]. On the other hand, in SARS CoV2 two of the currently selected B cell epitopes at amino acid residues 69 and 244 are located near to amino acid pockets described in Wuhan, India and France with minor amino acid difference in the Indian strain. These results point to an association of the overall B cell epitope, amino acid pocket structure and level of disease severity. It is therefore, tempting to speculate that deletions G107 or Y144 represent escape variants against S1-NTD endogenous nAb. These escape mutants modify pocket pattern conformation, alter the binding affinity to nAb and allow NTD mediated viral entry. This

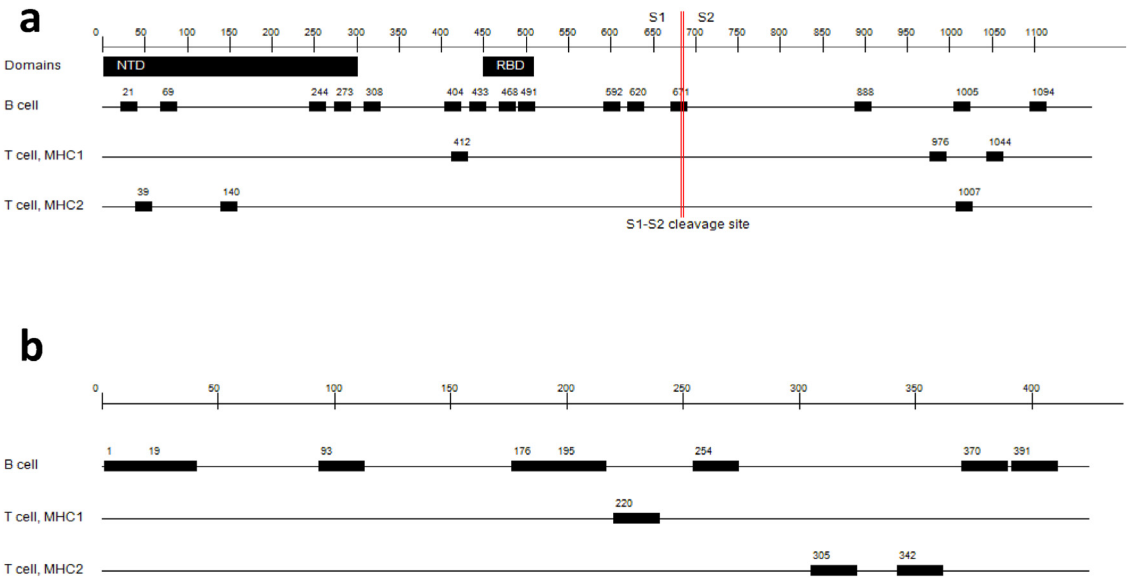

Fig. 3. Location of SARS CoV2-S (a) and N (b) proteins derived B cell and T cell epitopes (MCH1 and MCH2) on the protein structure. Functional domains in the S protein are marked as follows: NH2-Terminal Domain (a.a. 1–333), Receptor Binding Domain (a.a. 438–506) and S1/S2 cleavage site (a.a. 683–686).

speculation requires experimental evidence from wet laboratory studies. Further, a SARS-CoV RBD-specific human neutralizing mAb, CR3022, may recognize SARS CoV2 RBD specific epitope (s) that do not overlap with ACE 2 receptor binding [31]. Studies from SARS-CoV and MERS-CoV demonstrated that the principle domains of the S protein, i.e. S1-NTD, RBD and S2 are potential targets for nAbs. Given the exclusive function of the polybasic cleavage site at S1/S2 junction of SARS CoV2, the overall research on SARS CoV, MERS CoV and SARS CoV2 will provide major guidelines for designing SARS CoV2 nAbs and vaccines. Recently, plasma from convalescent SARS CoV2 were transfused to critically ill patients and was associated with clinical improvement in most patients [32]. However, the study was un-controlled and the patients had history of receiving other antiviral treatments and the nAb levels in each convalescent plasma was not justified. Therefore, we assume that identification of epitopes specific for nAbs is crucial for immunotherapeutic and prophylactic interventions. Structural studies revealed that SARS CoV2 appears to comprise a RBD within the S protein that binds to ACE2 in humans [33]. The S protein of SARS CoV2 contains a unique functional poly basic PRRA (residues 683-S 686) cleavage site at the junction between S1 and S2, the 2 spike subunits [34] that was never reported in other corona viruses. This site allowed the acquisition of 3 O-linked glycan shields for immune evasion [4]. A classical option is to develop S subunit vaccine cloned in a replicative defective-Adenovirus vectors; AdHu5 or AdC7 vectors driven by CMV or Chicken β actin promoters respectively [35]. However, using intact S subunit presumably preserves the poly basic cleavage sites and hence O-linked glycan that shield epitope or furin specific residues and allows immune evasion and reducing vaccine efficiency.

Although monoclonal antibodies against SARS CoV2 have not yet been reported, polyclonal antibodies from convalescent SARS CoV2 patients are currently used to treat succumbed patients [36]. Since SARS CoV2 is closely related to SARS CoV with great homology between S2 subunits and less homology in S1 subunits [37] several laboratories are working on producing such mAbs or other immunotherapeutic agents to protect from or cross neutralize SARS CoV2. Investigators are currently attempting to produce cross reactive or cross nAbs against SARS CoV2 infection. The most attractive target for nAb development is the RBD located within the sequence 439–506 in the S1 subunit and does not overlap with ACE2 binding site [31].

Another attractive epitope is the polybasic furin specific cleavage site; between S665 and S695 at the boundaries of S1 and S2. A third and promising target for immunotherapeutic or prophylactic intervention is located in the NTD in the form of a loop structure. This site was previously described in MERS CoV infection as it is composed of 2 adjacent loops that allow binding of heavy and light chains of the Fab portion of a murine nAb. Although our search in SARS CoV2 S protein for similar epitopes to MERS CoV loop structure failed to detect homologous sequence, our bio informatics data revealed identical locations of pockets with potential B cell epitopes at residues 244 and 69 within S NTD of Wuhan reference strain. During the pan epidemic crises of SARS CoV2 outbreak, there is a critical need to use sera from convalescent SARS CoV2 subjects for treating patients who succumbed to the disease [32,38]. The major hurdle facing this strategy is that the recorded convalescent subjects are relatively few. The bio-informatics approach used in this study helps defining an epitope panel for detection of nAbs in convalescent plasma. This strategy focuses on high risk population such as health care workers specially those working in quarantine hospitals, relatives of infected patients as well as subjects who contracted SARS CoV2 infection at subclinical or asymptomatic levels without even knowing they were infected. Besides CD4$^+$ T cell specific antigenicity, the exploration of S and N epitopes specific for cross nAbs are of utmost importance for vaccine design, particularly via reducing S protein-mediated SARS CoV2 entry. Recent studies emphasized the striking structural similarity between SARS Co V and SARS CoV2 S glycoproteins [13] and their roles in target cell entry via the human ACE2 receptor [39,40] and the finding that antibody responses potentially neutralize SARS CoV2 S mediated entry into cells. The majority of these nAbs are directed to the S1 protein particularly those directed against the RBD (438–506) and block S-receptor binding [13].

The T cell epitope located at a.a. 140 lies within the vicinity of the Y 144 deletion described in the Indian isolate. It is interesting to note that both G107 and Y144 deletions were associated with remarkable alteration in the VYY pocket characteristic of the Wuhan isolate. This change most likely plays important role in the S mediated immune dysregulation in SARS CoV2 severe infections (death rates 12.6% in France vs. 1.86% in India, as of 24 August, 2020). Such immune dysregulation involves lymphopenia, low CD4$^+$ counts and elevated cytokines (TNFα, IL1 and IL6) as well as chemokine (IL8) [41]. This picture of immune response i.e. higher expression of cytokines and chemokine associated with excessive consumption of helper T cells (Th) i.e. CD4^{+} and regulatory T cells as well as suppressor T cells (CD8^{+}) might result in outraged inflammation with a production of cytokine storm and further worsening of the disease.

This study highlights the striking difference in SARS CoV2 infection rates among diverse populations worldwide. The recorded rates of corona infection depend on the competence of patient enrollment, efficient diagnosis, and efficacy of healthcare protocols as well as on immunological and genetic host factors in specific populations. The current study implements 3 lines of research: (1) The viral factors associated with population specific degree of virulence. (2) Phylogenetic tree, genetic distance and protein sequence alignment as integrative tools to identify strain dissimilarities within populations and the role of mutational events in disrupting Spike mediated cell entry. (3) Maximizing the value of combining wet laboratory findings and bioinformatics in elucidating better designs for B and T epitopes for future immune-therapeutics and vaccine development.

The study showed that the selected populations from Asia, Europe and North America had great homology of the S protein sequences known at the earlier times of the outbreak, i.e. January 2020. Although 12 of the 14 strains had 100% homology of S sequences, only 2 strains from India and France had deletion mutations within the NTD of S1 protein. We therefore aimed to discuss the structure–function impacts of those 2 deletions, where both obviously revealed significant conformational changes of S1 glycoprotein. The main limitations of the current study are the relatively small number of presented population. Furthermore, the principle conclusion of the in silico analyses requires, indeed, extensive wet laboratory experiments employing site directed mutagenesis of S1 NTD.

Funding

No funding sources.

Competing interests

None declared.

Ethical approval

Not required.

References

[1] Zhang SF, Tuo JL, Huang XB, Zhu X, Zhang DM, Zhou K, et al. Epidemiology characteristics of human coronaviruses in patients with respiratory infection symptoms and phylogenetic analysis of HCoV-OC43 during 2010-2015 in Guangzhou. PLOS ONE 2018;13(1):e0191789.

[2] Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Di Napoli R. Features, evaluation and treatment coronavirus (COVID-19). In: StatPearls. Treasure Island (FL); 2020.

[3] Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and sources of endemic human coronaviruses. Adv Virus Res 2018;100:163–88.

[4] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26(4):450–2.

[5] COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE). Johns Hopkins University (JHU); 2020.

[6] Gilbert M, Pullano G, Pinotti F, Valdano E, Poletto C, Boelle PY, et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. Lancet 2020;395(10227):871–7.

[7] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579(7798):265–9.

[8] Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses 2020;12(3).

[9] Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B 2020.

[10] Li F. Receptor recognition mechanisms of coronaviruses: a decade of structural studies. J Virol 2015;89(4):1954–64.

[11] Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 2003;426(6965):450–4.

[12] Henderson R, Edwards RJ, Mansouri K, Janowska K, Stalls V, Kopp M, et al. Glycans on the SARS-CoV-2 spike control the receptor binding domain conformation. bioRxiv 2020.

[13] Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020;181(2):281–92, e6.

[14] Hasan A, Paray BA, Hussain A, Qadir FA, Attar F, Aziz FM, et al. A review on the cleavage priming of the spike protein on coronavirus by angiotensin-converting enzyme-2 and furin. J Biomol Struct Dyn 2020:1–9.

[15] Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181(2):271–80, e8.

[16] Li G, Fan Y, Lai Y, Han T, Li Z, Zhou P, et al. Coronavirus infections and immune responses. J Med Virol 2020;92(4):424–32.

[17] Deming D, Sheahan T, Heise M, Yount B, Davis N, Sims A, et al. Vaccine efficacy in senescent mice challenged with recombinant SARS-CoV bearing epidemic and zoonotic spike variants. PLoS Med 2006;3(12):e525.

[18] Graham RL, Becker MM, Eckerle LD, Bolles M, Denison MR, Baric RS. A live, impaired-fidelity coronavirus vaccine protects in an aged, immunocompromised mouse model of lethal disease. Nat Med 2012;18(12):1820–6.

[19] Lin Y, Shen X, Yang RF, Li YX, Ji YY, He YY, et al. Identification of an epitope of SARS-coronavirus nucleocapsid protein. Cell Res 2003;13(3):141–5.

[20] World Health Organization. Coronavirus disease (covid-2019) situation reports; 2020.

[21] Enayatkhani M, Hasaniazad M, Faezi S, Guklani H, Davoodian P, Ahmadi N, et al. Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: an in silico study. J Biomol Struct Dyn 2020:1–19.

[22] Shang W, Yang Y, Rao Y, Rao X. The outbreak of SARS-CoV-2 pneumonia calls for viral vaccines. NPJ Vaccines 2020;5:18.

[23] Walls AC, Tortorici MA, Bosch BJ, Frenz B, Rottier PJM, DiMaio F, et al. Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. Nature 2016;531(7592):114–7.

[24] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res 2019;47(D1):D339–43.

[25] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35(6):1547–9.

[26] Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10(6):845–58.

[27] Schmidtke P, Le Guilloux V, Maupetit J, Tuffery P. fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic Acids Res 2010;38(Web Server issue):W582–9.

[28] Wang N, Rosen O, Wang L, Turner HL, Stevens LJ, Corbett KS, et al. Structural definition of a neutralization-sensitive epitope on the MERS-CoV S1-NTD. Cell Rep 2019;28(13):3395–405, e6.

[29] Peng G, Sun D, Rajashankar KR, Qian Z, Holmes KV, Li F. Crystal structure of mouse coronavirus receptor-binding domain complexed with its murine receptor. Proc Natl Acad Sci USA 2011;108(26):10696–701.

[30] Li W, Hulswit RJG, Widjaja I, Raj VS, McBride R, Peng W, et al. Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. Proc Natl Acad Sci USA 2017;114(40):E8508–17.

[31] Tian X, Li C, Huang A, Xia S, Lu S, Shi Z, et al. Potent binding of 2019 novel coronavirus spike protein by a SARS coronavirus-specific human monoclonal antibody. Emerg Microbes Infect 2020;9(1):382–5.

[32] Shen C, Wang Z, Zhao F, Yang Y, Li J, Yuan J, et al. Treatment of 5 critically ill patients with COVID-19 with convalescent plasma. JAMA 2020.

[33] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367(6483):1260–3.

[34] Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 2020;176:104742.

[35] Kobinger GP, Figueredo JM, Rowe T, Zhi Y, Gao G, Sanmiguel JC, et al. Adenovirus-based vaccine prevents pneumonia in ferrets challenged with the SARS coronavirus and stimulates robust immune responses in macaques. Vaccine 2007;25(28):5220–31.

[36] Jiang S, Hillyer C, Du L. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. Trends Immunol 2020.

[37] Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect 2020;9(1):221–36.

[38] Duan K, Liu B, Li C, Zhang H, Yu T, Qu J, et al. Effectiveness of convalescent plasma therapy in severe COVID-19 patients. Proc Natl Acad Sci USA 2020;117(17):9490–6.

[39] Kuba K, Imai Y, Rao S, Gao H, Guo F, Guan B, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. Nat Med 2005;11(8):875–9.

[40] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579(7798):270–3.

[41] Qin C, Zhou L, Hu Z, Zhang S, Yang S, Tao Y, et al. Dysregulation of immune response in patients with COVID-19 in Wuhan, China. Clin Infect Dis 2020.