Indian J Med Res, Epub ahead of print
DOI: 10.4103/ijmr.IJMR_663_20

Quick Response Code:

Full-genome sequences of the first two SARS-CoV-2 viruses from India

Pragya D. Yadav[1,†], Varsha A. Potdar[2,†], Manohar Lal Choudhary[2], Dimpal A. Nyayanit[1], Megha Agrawal[2], Santosh M. Jadhav[2], Triparna D. Majumdar[1], Anita Shete-Aich[1], Atanu Basu[3], Priya Abraham[#] & Sarah S. Cherian[4]

[1]Maximum Containment Laboratory, [2]Influenza Group, [3]Electron Microscopy & [4]Bioinformatics & Data Manangement Group, [#]ICMR-National Institute of Virology, Pune, Maharashtra, India

Received March 13, 2020

***Background & objectives*: Since December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has globally affected 195 countries. In India, suspected cases were screened for SARS-CoV-2 as per the advisory of the Ministry of Health and Family Welfare. The objective of this study was to characterize SARS-CoV-2 sequences from three identified positive cases as on February 29, 2020.**

***Methods*: Throat swab/nasal swab specimens for a total of 881 suspected cases were screened by *E* gene and confirmed by *RdRp* (1), *RdRp* (2) and *N* gene real-time reverse transcription-polymerase chain reactions and next-generation sequencing. Phylogenetic analysis, molecular characterization and prediction of B- and T-cell epitopes for Indian SARS-CoV-2 sequences were undertaken.**

***Results*: Three cases with a travel history from Wuhan, China, were confirmed positive for SARS-CoV-2. Almost complete (29,851 nucleotides) genomes of case 1, case 3 and a fragmented genome for case 2 were obtained. The sequences of Indian SARS-CoV-2 though not identical showed high (~99.98%) identity with Wuhan seafood market pneumonia virus (accession number: NC 045512). Phylogenetic analysis showed that the Indian sequences belonged to different clusters. Predicted linear B-cell epitopes were found to be concentrated in the S1 domain of spike protein, and a conformational epitope was identified in the receptor-binding domain. The predicted T-cell epitopes showed broad human leucocyte antigen allele coverage of A and B supertypes predominant in the Indian population.**

***Interpretation & conclusions*: The two SARS-CoV-2 sequences obtained from India represent two different introductions into the country. The genetic heterogeneity is as noted globally. The identified B- and T-cell epitopes may be considered suitable for future experiments towards the design of vaccines and diagnostics. Continuous monitoring and analysis of the sequences of new cases from India and the other affected countries would be vital to understand the genetic evolution and rates of substitution of the SARS-CoV-2.**

Key words Epitope - genomes - India - Kerala - next-generation sequencing - phylogeny - real-time reverse transcription-polymerase chain reaction - severe acute respiratory syndrome coronavirus 2

[†]Equal contribution

Supplementary material available from *http://www.ijmr.org.in/preprintarticle.asp?id=281471*

The *Coronaviridae* family encompasses viruses with a single-stranded, positive-sense RNA genome of size approximately 26-32 kb. Initially, the virus was associated with human and animal infections that caused intestinal as well as respiratory infections[1,2]. In 2002, the severe acute respiratory syndrome (SARS) coronavirus (CoV) outbreak that claimed the lives of many people in China raised the alarm towards these viruses[2]. Further, after a decade, another human pathogenic virus emerged, Middle East respiratory syndrome CoV (MERS-CoV) that affected the Middle Eastern countries[2]. Current knowledge identifies six virus groups that can infect humans[3] in the *Coronaviridae* family, which includes SARS-CoV (now termed as SARS-CoV-1) and MERS-CoV.

Recently in December 2019, China reported cases with pneumonia of unknown aetiology in the Hubei province, Wuhan city[4]. Further analysis of these cases was carried out to identify the causative agent of pneumonia[5]. Virus isolation and genomic characterization of the complete sequence of the virus through next-generation sequencing (NGS), identified it as a novel CoV, named 2019-nCoV[3]. The virus characterization revealed that it is an enveloped RNA virus with a genome size of 29,903 bp. The phylogenetic analysis of the sequence showed that it belonged to the *Sarbecovirus* subgenus of genus *Betacoronavirus* and the family *Coronaviridae*. The sequence was closely related (~87.5% sequence similarity) to two bat-derived SARS-like CoV strains (bat-SL-CoVZC45 and bat-SL-CoVZXC21) that are known to infect humans, including the virus which led to the 2003 SARS-CoV-1 outbreak[6]. The 2019-nCoV is now named as SARS-CoV-2[7]. Further, based on SimPlot analyses, it was demonstrated that SARS-CoV-2 was more closely related to the BatCoV RaTG13 sequence (~96.3% similarity) throughout the genome. The bat-SL-CoVZC45 and bat-SL-CoVZXC21 strains clustered differently from the group formed by SARS-CoV-2 and BatCoV RaTG13 in the region spanning the 3′-end of open reading frame (ORF)1a, the ORF1b and almost half of the spike region[8].

The receptor-binding domain (RBD) of the spike protein mediates interaction with the host cell receptor[9], and the angiotensin-converting enzyme 2 (ACE2) has been identified as the receptor for the SARS-CoVs[10]. Specific mutations in the RBD of the SARS-CoV-2 spike glycoprotein were found to have enhanced binding to the ACE2[11].

The human-to-human transmission of the SARS-CoV-2 created an alert with the increasing number of cases[12]. The WHO report dated February 28, 2020 confirmed 83,652 cases of SARS-CoV-2, with a total of 2,858 deaths from 52 countries[12]. After the first report of SARS-CoV-2 from Wuhan, China, the Government of India reviewed and initiated multisectoral measures for the mitigation of this emerging public health crisis. These include point-of-entry surveillance at 21 international airports, enhanced State-level surveillance programmes and preparedness for handling clinical cases in designated hospitals. Till date, the Integrated Disease Surveillance Programme (IDSP), a national health programme, Government of India, has collected samples from symptomatic travellers in liaison with the State-level Viral Research and Diagnostic Laboratories (VRDLs), Department of Health Research. These VRDLs respond for timely diagnosis during outbreaks.

The suspected samples were collected and transported to the Indian Council of Medical Research-National Institute of Virology (ICMR-NIV), Pune, for the diagnosis of SARS-CoV-2. The specimens of the positive cases were diagnosed with real-time reverse transcription-polymerase chain reaction (RT-PCR)-specific for SARS-CoV-2 using the protocol published by the WHO[13] and characterized by complete genome sequencing and epitope prediction analyses. These sequences were also compared with the available GenBank sequences to monitor the mutations and understand their relation with other known SARS-CoV-2 available in the public database. Here, we report molecular characterization of SARS-CoV-2 sequences from three positive cases.

Material & Methods

The clinical samples were referred by the hospital authorities through the Kerala State Health Services for diagnostic purposes. Further samples were received from different parts of India for establishing the presence of SARS-CoV-2.

Detection of SARS-CoV-2 in suspected samples: Blood and throat swab (TS) specimens were collected from the suspected cases that complied with the case definition of SARS-CoV-2 infection as per the guidelines of the Ministry of Health and Family Welfare[14]. The TS was collected in viral transport medium. These samples were referred to the ICMR-NIV, Pune, India (which is the national reference laboratory for India, also referred as the government's apex laboratory). As

of February 29, 2020, 881 samples of suspected cases referred from different States, with a travel history to Wuhan, China, and other SARS-CoV-2-affected countries, were screened.

The viral RNA was extracted from the TS sample using the Magmax RNA extraction kit (Applied Biosystems, USA) as per the manufacturer's instructions. The extracted RNA was immediately used for testing the presence of SARS-CoV-2 using the real-time RT-PCR protocol published by the WHO[12] for the detection of *RdRp* (1), *RdRp* (2), *E* gene and *N* gene. *RNase P* gene was used as the internal control for the analysis. Confirmatory laboratory tests were performed as per the WHO-recommended test protocols[13]. These samples were also sequenced using the NGS approach to retrieve the complete genome of the virus.

NGS of SARS-CoV-2 from India - Phylogenetic analysis and molecular characterization: The total RNA of three positive TS specimens from Kerala, was extracted from 250-300 µl of the SARS-CoV-2 real-time RT-PCR positive samples. QIAamp Viral RNA extraction kit (QIAGEN, Hilden, Germany) was used according to the manufacturer's instructions. The extracted RNA was further quantified using a Qubit RNA High-Sensitivity kit (Invitrogen, USA). RNA libraries were prepared as per the earlier-defined protocol and quantified using KAPA Library Quantification Kit (Kapa Biosystems, Roche Diagnostics Corporation, USA) as per the manufacturer's protocol. Further, individual libraries were neutralized and loaded on the Miniseq platform (Illumina, USA). The detailed protocols for the steps undertaken have been published earlier[15,16]. The data generated from the machine were analyzed using CLC genomics workbench version 11.0 (CLC, QIAGEN, Germany). Reference-based mapping was performed to retrieve the sequence of the SARS-CoV-2.

Full-length genome sequences of SARS-CoV-2 were downloaded from the GISAID database[17] (Supplementary Table I). Multiple sequence alignment was performed using the MEGA software version 7.0[18] with retrieved sequences from two of the three positive cases and the available GISAID sequences. A phylogenetic tree was generated using the neighbour joining method and the Kimura-2-parameter as the nucleotide (nt) substitution model with 1000 bootstrap replications as implemented in MEGA software[18]. Per cent nucleotide divergence and amino acid (aa) divergence were calculated using the p-distance method[18]. Mutations specific to the Indian SARS-CoV-2 viruses were identified by comparing the coding regions with respect to the SARS-CoV-2, Wuhan, China (Wuhan hu-1).

Three-dimensional (3D) model of the spike protein and epitope prediction: The pre-fusion structure of the Indian case 1 SARS-CoV-2 spike (S) glycoprotein was modelled using the Swiss-Model server (*https://swissmodel.expasy.org/interactive*) and the corresponding S protein of Wuhan-Hu-1 (6VSB.PDB) as the template (99.97% identity). Sequential (linear) B-cell epitopes were predicted using BepiPred-2.0 server (*http://www.cbs.dtu.dk/services/BepiPred/*). The ABCpred prediction tool (*http://crdd.osdd.net/raghava/abcpred/*) was also used to identify the B-cell epitopes in the Indian SARS-CoV-2 sequence. The epitope prediction probability of >0.8 was set to increase the specificity of the peptide stretch. The overlapping epitopes predicted by BepiPred-2.0 online server and the ABCpred prediction tool were identified. The antigenicity of the shortlisted peptide sequences was further predicted using the Vaxijen online server (*http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html*) with a default threshold of 0.4.

Discontinuous epitopes on the modelled structure of the Indian case 1 SARS-CoV-2 spike protein were predicted using the online servers, Ellipro (*http://tools.iedb.org/ellipro/*) and DiscoTope 2.0 (*http://tools.iedb.org/discotope/*), integrated in the Immune Epitope Database. Ellipro predicts epitopes based on the protusion index (PI), wherein the protein shape is approximated as an ellipsoid (Ref for Ellipro and DiscoTope). An ellipsoid with the PI value of 0.8 indicates that 80 per cent of the residues are within the ellipsoid and 20 per cent are outside. All residues that are outside the 80 per cent ellipsoid will have a score of 0.8. Residues with larger scores are associated with greater solvent accessibility. The PI value was set to a score of 0.8. DiscoTope predicts epitopes using 3D structure and half-sphere exposure as a surface measure in a novel spatial neighbourhood definition method. Default values were set for sensitivity (0.47) and specificity (0.75) for selecting the amino acids forming discontinuous epitopes. A sensitivity of 0.47 means that 47 per cent of the epitope residues are predicted as part of the epitopes, while a specificity of 0.75 means that 25 per cent of the non-epitope residues are predicted as part of the epitopes. Outputs from both the methods were combined, and the final regions

were mapped on the modelled 3D-structure as the most probable conformational epitopes. In addition, we also predicted N-linked glycosylation sites in the S protein using NetNGlyc 1.0 Server (*http://www.cbs.dtu.dk/services/NetNGlyc/)*. The spike proteins were also screened for the presence of potential epitopes presented by major histocompatibility complex (MHC) class I molecules to cytotoxic T lymphocytes (CTLs). The online NetCTL1.2 server (*http://www.cbs.dtu.dk/services/NetCTL/)* based on machine learning techniques such as artificial neural network (ANN) and support vector machine (SVM) was used to predict the T-cell epitopes. The prediction was made for all the human leucocyte antigen (HLA) supertypes and the available human alleles. The C terminal cleavage, weight of transport-associated protein (TAP) efficiency and threshold for identification were kept as default. VaxiJen v2.0 tool was used to predict the antigenicity of the predicted epitopes (*http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html)*. The sequences were further screened to be potential epitopes using the CTLPred online server (*http://crdd.osdd.net/raghava/ctlpred/*).

The ability of the predicted linear B-cell and the T-cell epitopes to mount interferon-gamma (IFN-γ) response was assessed using the IFNepitope (*http://crdd.osdd.net/raghava/ifnepitope/index.php*).

Results

Detection of SARS-CoV-2 in suspected samples: Three of the 881 TS/nasal swab (NS) specimens from the suspected cases, tested positive for the SARS-CoV-2 using the real-time RT-PCR specific to *E* gene, *RdRp* (1), *RdRp* (2) and *N* gene. The Ct value of the *E* gene ranged from 19.8 to 34.5 for the TS/NS specimens. Detailed Ct values for the real-time RT-PCRs specific to the above-mentioned genes of the positive specimens are given in Table I. Blood samples were found to be negative for the SARS-CoV-2.

Case 1 travelled from Wuhan, China, reached India on January 23, 2020 and further travelled to the final destination of Kerala on January 24. This individual developed cough on January 25 and further experienced a sore throat and mild fever and was admitted to the General Hospital, Thrissur, Kerala. The second case travelled from Wuhan and had close contact with case 1 during the travel to the final destination in India. Case 2 developed similar symptoms along with fever and diarrhoea on January 26, and the collected TS specimens were referred to the ICMR-NIV on January 28. The second case was hospitalized on January 30, in a medical college, Alappuzha, Kerala. The clinical sample (TS) was collected on January 31, 2020. Case 3 travelled from China to India, developed a runny nose on January 30 and was admitted to the General Hospital, Kasaragod, Kerala, on January 31, 2020. TS specimens were collected on January 31, 2020.

NGS of SARS-CoV-2 from India - Phylogenetic analysis and molecular characterization: NGS analysis from the TS specimens retrieved two complete genome sequences from case 1 and case 3. The complete genomic sequence data for case 2 could not be recovered due to the lower kappa concentration of the sample and hence not included in the study for analysis. The FastQ files were reference mapped with the available Wuhan seafood pneumonia virus (Wuhan Hu-1) complete SARS-CoV-2 genome (accession number: NC 045512.2). The total reads which were mapped and the percentage of the genome recovered for the two cases are summarized in Table I.

Analysis of the complete genome sequences of SARS-CoV-2 from the positive cases in India revealed that the percentage nt and aa differences between case 1 and case 3 were 0.038 and 0.10 per cent, respectively. The sequences of case 1 and case 3 diverged from the Wuhan-Hu1 sequence by 0.017 per cent nt and 0.041 per cent aa respectively. Indian SARS-CoV-2 clustered with the *Sarbecovirus* subgenus of the *Betacoronavirus* genus and was closest to the BatCoV RaTG13 sequence (96.09% nt)[8]. The phylogenetic comparison showed the clustering of the genome sequences of case 1 and case 3 with the existing sequences of the SARS-CoV-2 sequences (Fig. 1). The phylogeny revealed emerging heterogeneity within the SARS-CoV-2 sequences globally. The Indian SARS-CoV-2 viruses were positioned in different clusters.

Indian SARS-CoV-2 sequences showed two changes 408 Arg→Ile and 930 Ala→Val in the spike protein compared to the Wuhan Hu-1 sequence. The mutations were further mapped on the spike protein model of the Indian sequence (Supplementary Fig. 1). Deletion of a three-nucleotide stretch, encoding tyrosine residue at position 144, of the spike gene was also observed in the Indian SARS-CoV-2 from case 1 when compared to the other SARS-CoV-2 sequences. As noted in the earlier SARS-CoV-2 sequences, both the Indian sequences possessed the polybasic cleavage site (RRAR) in the spike protein at the junction of S1 and S2, the two subunits of the spike protein[19].

Table I. Real-time reverse transcription-polymerase chain reaction (RT-PCR) values for *RdRp* (1), *RdRp* (2), *E* gene and *N* gene, per cent genome coverage recovered and reads mapped for the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) positive cases

Positive cases	Ct values for real-time RT-PCR for the confirmation of SARS-CoV-2					Relevant reads	Total reads	Genome length recovered (bp)	Per cent genome coverage
	RdRp (1)	*RdRp* (2)	*E* gene	*N* gene	*Rnase P* internal control				
Case 1	33.33	27.93	34.5	33.90	Positive	20,096	5,615,846	29,854	99.83
Case 2	24.6	29	19.8	38	Positive	610	8,587,146	16,047	53.66
Case 3	34.17	32.64	28.98	36.35	Positive	11,296	1,405,038	29,851	99.83

Epitope predictions: Thirty one linear B-cell epitopes were predicted by Bepipred in the Indian SARS-CoV-2, of which three were found to have a length of <6 amino acids and hence not considered. Linear epitopes were also predicted using the ABCpred prediction tool, which predicted 47 epitopes based on the threshold of 0.8. Regions common to both the prediction methods (n=17) were identified manually. The 17 epitopes were screened for their antigenicity using the VaxiJen v2.0 tool (*http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html*), and nine of these epitopes were shortlisted. These epitopes were further screened for their ability to elicit an IFN-γ response, which was predicted using the IFNepitope tool. Finally, five epitopes, four in the S1 domain and one in the S2 domain, were predicted, which could possibly generate an immune response and suppress the IFN-γ response (Table II). N-linked glycosylation site prediction revealed that two putative glycosylation sites (with a low value for jury agreement) were present within the epitope stretch 328-344.

The discontinuous epitopes in the spike protein of the Indian SARS-CoV-2 were further identified using multiple methods, Ellipro and DiscoTope. Conformational epitopes based on these methods were mapped on the pre-fusion structure of the modelled Indian SARS-CoV-2 spike protein. The newly released structure of the SARS-CoV-2 spike protein was used as the template for modelling the Indian spike protein. Ramachandran plot statistics revealed 83.7 per cent of the residues to be in the core region, 14.4 per cent in the additionally allowed region and 0.5 per cent in the disallowed region. Four epitopes were predicted by Ellipro based on the PI threshold of 0.8 (Supplementary Table II). The result from the DiscoTope is presented in Supplementary Table III. The mapped conformational epitopes are depicted in Figure 2. For the purpose of comparison, the Indian S protein sequence was also modelled using the pre-fusion structure of SARS-CoV-1 (6ACC.PDB; 87.29% identity), and the results for the conformational epitopes predicted are in Supplementary Table IV and Supplementary Figure 2.

T-cell epitope prediction revealed 105 strong binding epitopes capable of binding to different HLA types using the NetCTL1.2 software based on the threshold of 0.4. Twelve of these were shortlisted, considering a binding efficiency of >0.5 nM and capable of eliciting IFN-γ response (Table III).

Discussion

Till February 29, 2020, three positive cases of SARS-CoV-2 were reported from India from 881 suspected cases tested at ICMR-NIV, Pune. All the three cases had a travel history from Wuhan, China, during January 2020. Although NGS was performed on the specimens for all the three positive cases, the complete genome sequence could be retrieved only from case 1 and case 3. The three cases were recovered after hospitalization and were home quarantined as per the guidelines of the Ministry of Health and Family Welfare, Government of India[14].

The low viral copy number of the TS specimen from case 2 could be the possible reason for lesser viral reads being retrieved during the NGS run, leading to a fragmented genome. The recent study from China on serial samples (TSs, sputum, urine and stool) from two patients followed days 3-12 and days 4-15 post onset[20]. *N* gene-specific real-time RT-PCR assay showed that the viral loads in TS and sputum samples peaked at around 5-6 days after symptom onset, ranging from around 10^4-10^7 copies per ml during this time[20]. In another study, the virus was detected in the saliva specimens of 11 of the 12 patients, and serial saliva testing showed declines of viral RNA levels[21].

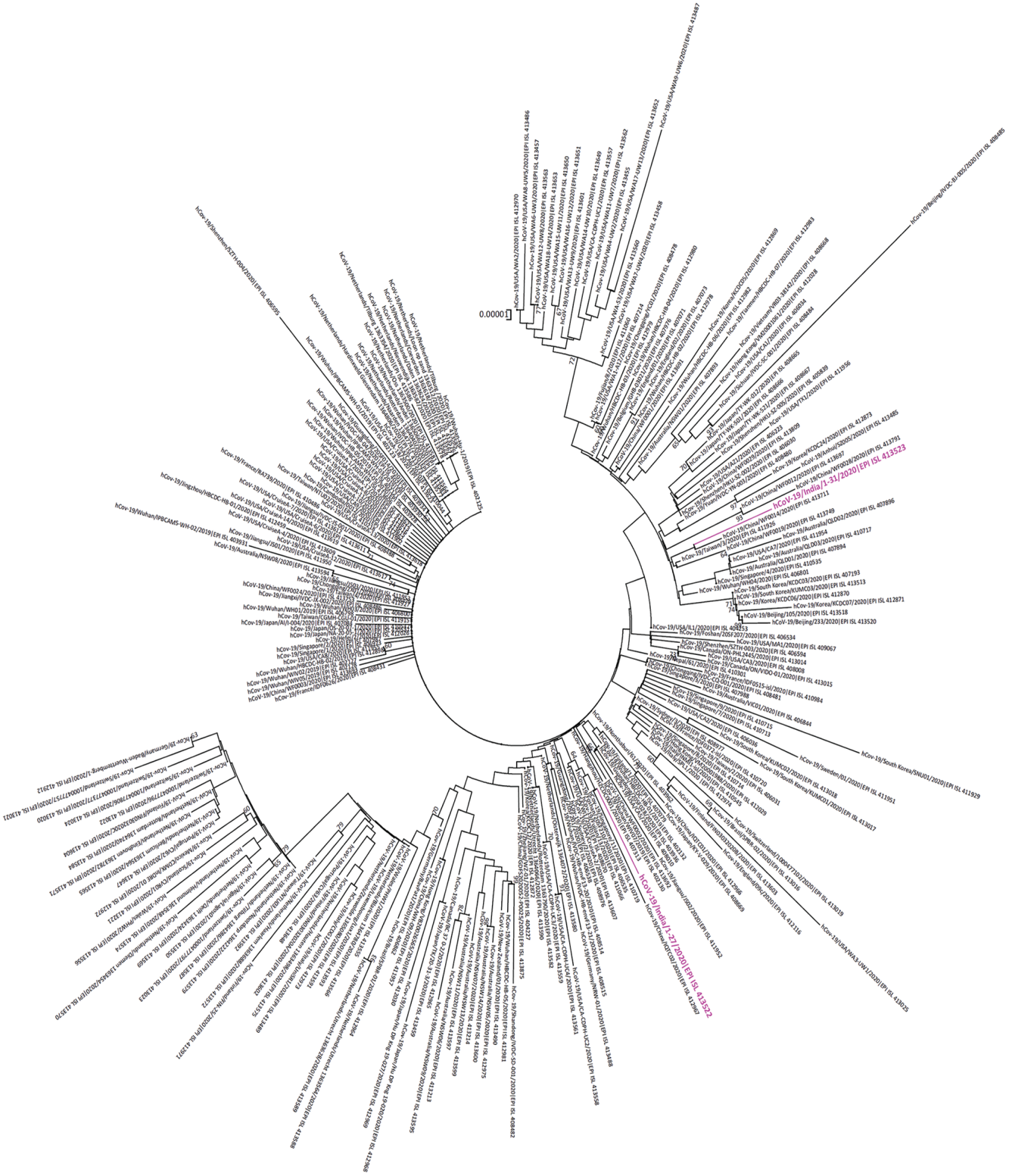

Fig. 1. Phylogenetic tree of the complete genomes of severe acute respiratory syndrome coronavirus 2 viruses. Indian viruses are shown in magenta font colour.

Table II. Linear B-cell epitopes predicted on the spike protein of the Indian severe acute respiratory syndrome coronavirus 2

Peptide	Epitope probability	Vaxigen score	Interferon (IFN)-γ response[#]
243-HRSYLTPGDSSSGWTA-258	0.92	Antigen (0.602)	Negative (1)
327-FPNITNLCPFGEVFNA-342	0.82	Antigen (0.606)	Negative (−0.132)
404-EVIQIAPGQTGKIADY-419	0.86	Antigen (1.231)	Negative (1)
413-TGKIADYNYKLPDDFT-428	0.84	Antigen (0.9642)	Negative (−0.334)
1204-YEQYIKWPWYIWLGFI-1219	0.89	Antigen (0.951)	Negative (1)

Epitopes were predicted using a combination of the Bepipred server and the ABCpred prediction server. The antigenicity was predicted using the VaxiJen v2.0 tool. IFN-γ response was predicted using the INFepitope server. [#]Values in bracket show prediction score given by the software

Table III. Spike protein peptides capable of binding to major histocompatibility complex (MHC) class I predicted using NetCTL server

Peptide	Vaxijen	Interferon (IFN)-γ response	CTLPred Score (ANN/SVM)	MHC restriction
89-GVYFASTEK-97	0.711	Positive (1)	0.58/0.986	HLA-A*1101, HLA-A3, HLA-A*3101, HLA-A68.1, HLA-B*2705
166-FEYVSQPFL-174	0.632	Positive (0.087)	0.65/0.184	HLA-A2, HLA-A*0201, HLA-A*0205, HLA-A2.1, HLA-B*2702, HLA-B*2705, HLA-B*3701, HLA-B40, HLA-B*4403, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B60, HLA-B61, HLA-Cw*0301, H2-Kb, H2-Kk,
256-WTAGAAAYY-264	0.630	Positive (0.576)	0.82/0.544	HLA-A1, HLA-B*2702, HLA-B*3501, HLA-B*4403, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B*5801, HLA-B62, HLA-Cw*0702
348-VYAWNRKRI-356	0.500	Positive (0.499)	0.93/0.497	HLA-A24, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*51, HLA-Cw*0401, H2-Db, H2-Kd, H2-Kk
503-YQPYRVVVL-511	0.596	Positive (0.292)	0.40/0.596	HLA-A*0201, HLA-A*0205, HLA-A24, HLA-B14, HLA-B*2702, HLA-B*2705, HLA-B*3902, HLA-B*5201, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B60, HLA-B62, HLA-B7, HLA-B8, HLA-Cw*0401, HLA-Cw*0602, H2-Dd, H2-Kb, H2-Ld
510-VLSFELLHA-518	1.077	Positive (0.268)	0.86/0.276	HLA-A*0201, HLA-A*0205, HLA-A3, HLA-B*5301, HLA-B*51, HLA-B62
825-TLADAGFIK-833	0.578	Positive (0.014)	0.75/0.992	HLA-A1, HLA-A*1101, HLA-A3, HLA-A*3101, HLA-A68.1, HLA-A20, HLA-B*2705
1058-VVFLHVTYV-1066	1.512	Positive (1)	0.77/0.779	HLA-A2, HLA-A*0201, HLA-A*0205, HLA-A68.1, HLA-A2.1, HLA-B14, HLA-B*5101, HLA-B*5102, HLA-B*5103, HLA-B*5201, HLA-B*5301, HLA-B*5401, HLA-B*51
1210-WPWYIWLGF-1218	1.495	Positive (0.221)	0.68/0.0695	HLA-B*2702, HLA-B*2705, HLA-B*3501, HLA-B*3801, HLA-B*5101, HLA-B*5102, HLA-B*5201, HLA-B*5301, HLA-B*5401, HLA-B*51, HLA-B*5801, HLA-B62, HLA-B*0702, HLA-Cw*0401, HLA-Cw*0702. H2-Ld

Threshold of >0.7 nM was used for increased specificity of the prediction. The peptides were reconfirmed using CTLPred server using default parameters. The peptides that were classified as epitopes were further checked for their antigenicity score using the VaxiJen v2.0 tool

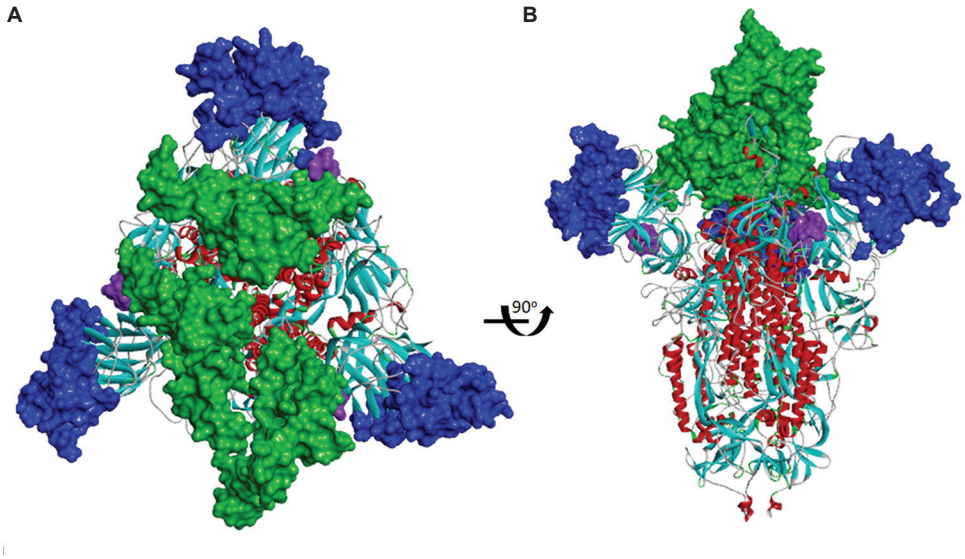

Fig. 2. Predicted conformational B-cell epitopes mapped on the pre-fusion structure of the modelled Indian severe acute respiratory syndrome coronavirus 2 spike protein using the pre-fusion structure of severe acute respiratory syndrome-coronavirus-2 (6VSB.PDB) (colour key: blue - epitopes 67-261; green - epitopes 341-507 based on the predicted epitopes as shown in Supplementary Table II). (**A**) Top view (**B**) Side view.

The two Indian SARS-CoV-2 sequences were found to be non-identical (0.04% nt divergence), and the result of phylogenetic analysis indicated that there were two different introductions into the country. A recent study using 52 published GenBank sequences showed evidence of substantial genetic heterogeneity and estimated the time to the most recent common ancestor to be December 5, 2019 (95% confidence interval: November 6 - December 13, 2019)[22]. Continuous monitoring and analysis of the sequences from the affected countries would be vital to understand the genetic evolution and rates of substitution of the SARS-CoV-2.

The comparison of the amino acid sequences of the non-structural (nsp1-nsp16) and structural polyproteins was undertaken with reference to the Wuhan-Hu1 strain for molecular characterization. Some human *Betacoronaviruses*, including HCoV-HKU1 (lineage A), have a polybasic cleavage site as well as predicted O-linked glycans near the S1/S2 cleavage site of the spike protein. As published recently, the polybasic cleavage site that has not been previously observed in related lineage B *Betacoronaviruses* and is a unique feature of SARS-CoV-2 was noted in the Indian SARS-CoV-2. The mutation Arg408Ile in the spike protein of one of the Indian sequences is noted to be in the RBD and Ala930Val, is located in the S2 domain. However, both are away from the ACE2 receptor-binding interface[19,23]. Mutations in the spike protein sequences of SARS-CoV-2 observed currently are localized over the S1 and S2 domains and, so far have not been found in the ACE2-binding interface.

From the alignment of the spike protein sequences of SARS CoV-1 and SARS-CoV-2 (Wuhan-Hu1 and India), it can be observed that the three nucleotide-deletion in the case 1 SARS-CoV-2 from India, is located close to the insert 1 region of the SARS CoV-1 (Supplementary Fig. 3). Notably, case 1 and case 2 were in close contact while travelling to India, but due to the absence of the complete genome of case 2, the genetic relatedness and source of infection could not be pinpointed.

Among the SARS-CoV structural proteins, the spike protein has been found to elicit neutralizing antibodies[24]. In this study, it was observed that of the five B-cell linear epitopes, which were predicted, four epitopes were present in the S1 domain and one in the S2 domain. Prediction of conformational B-cell epitopes revealed that one of these (residue positions 341-505) in the spike protein incorporates two of the predicted linear epitopes (327-342 and 404-419) having good antigenicity along with a favourable IFN-γ response that enables differentiation and proliferation of the B-cells[25]. Notably, an equivalent epitope (347-499) is predicted for the model generated using the SARS-CoV-1 S protein as a template. In both cases, this epitope lies within the RBD[6]. Although the epitope has two putative N-linked glycosylation sites within it at positions 330 and 332, the probability of these sites being actually glycosylated is very low. A major immuno-

dominant epitope has been reported from SARS-CoV between residues 441 and 700[26]. Hence, the predicted B-cell conformational epitope identified in the present study may play an important role in initiating a B-cell response. Among the five linear epitopes predicted in this study, epitopes 327-342 and 1204-1219 are conserved between SARS-CoV-2 and SARS-CoV-1. Epitopes 243-258, 404-419 and 413-428 are found to have variations.

The spike protein of SARS-CoV has also been reported to be immunogenic and elicit high IFN-γ-specific T-cell response[26]. The prediction results in this study revealed that nine possible CTL epitopes possessing good antigenicity and inducing IFN-γ response were present in the S protein. A recent report[27] also predicted T-cell epitopes in the S protein based on a similar ANN/SVM method and antigenicity score. Although the IFN-γ response was not considered by these authors, it was noted that two of the predictions were found to be common. Among the T-cell epitopes predicted in the present study, four epitopes 89-97 and 256-264 in the S1 domain and 825-833 and 1058-1066 in the S2 domain were found to have good CTL prediction scores with a broad HLA allele coverage of A and B supertypes. These HLA supertypes being predominant in the Indian population, the predicted epitopes may be considered suitable for future experiments towards vaccine design.

To conclude, the prompt intervention by the Government of India and the health authorities of the State of Kerala, ensured that the said cases did not become secondary foci of transmission. Further, the timely identification of SARS-CoV-2 in these suspected cases by the ICMR-NIV, Pune, has helped in the isolation of the patients, containment and enhanced surveillances for the virus and its restricted movement. The availability of the genomic sequences of the identified cases will contribute to the public repositories and help towards the development of diagnostics, vaccines and antivirals. The sequence data would also help in tracking the virus from its origin and evolution with its transmission in time.

Availability of data: Sequences are deposited in GISAID database, with accession numbers EPI ISL 413522 and EPI ISL 413523.

***Acknowledgment*:** Authors acknowledge the encouragement and support extended by Prof. (Dr) Balram Bhargava, Secretary to the Government of India, Department of Health Research, Ministry of Health and Family Welfare, and Director-General, Indian Council of Medical Research (ICMR), New Delhi, and Drs Raman Gangakhedkar and Nivedita Gupta, Division of Epidemiology & Communicable Diseases, ICMR, New Delhi. Authors thank the staff of ICMR-NIV, Pune, including Dr Gajanan Sapkal, Diagnostic Virology Group, staff of National Influenza Center: Shrimati V. Vipat, S. Jadhav, Drs S. Bharadwaj, R. Ghug, Ms U. Saha, Servshri H. Kengle, A. Awhale, V. Malik, Ms A. Jagtap, Shri A. Gondhalikar, Ms S. Digraskar, Ms P. Malsane, Shri V. Awatade, Ms S. Bhorekar, Dr S. Salve, Ms P. Shinde, Dr B. Nimhas, Shri T. Raut, Maximum Containment Facility, Dr Sreelekshmy Mohandas, Shrimati Savita Patil, Shri Hitesh Dighe, Shrimati Ashwini Waghmare, Shri Shrikant Baradkar, Ms Kaumudi Kalale, Epidemiology Section: Drs B.V. Tandale, Y.K. Gurav, Shilpa Tomar, A. Devshetwar, Bioinformatics Section: Shri Atul Walimbe, Shrimati Bhagyashree Kasbe, Shri Chandan Saini and Ms Deepika Chowdhary, Director Office, for her support. Authors also acknowledge the contribution of Kerala State Officials for case monitoring, sample collection, packaging and shipment with the support of Dr Meenakshi, Additional DHS (Public Health), Department of Health Services, Dr Amar Fettle, State Nodal Officer, Dr R. Nikhilesh Menon, Assistant Surgeon and Assistant Nodal Officer, Ernakulam, Department of Health Services, and Dr L.R. Chithra, Assistant Surgeon, Department of Health Services, Government of Kerala.

***Financial support & sponsorship*:** None.

***Conflicts of Interest*:** None.

References

1. Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol Mol Biol Rev* 2005; *69* : 635-64.
2. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019; *17* : 181-92.
3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, *et al*. A novel coronavirus from patients with pneumonia in China, 2019. *New Engl J Med* 2020; *382* : 727-33.
4. Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *J Med Virol* 2020; *92* : 401-2.
5. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, *et al*. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* 2020; *395* : 507-13.
6. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* 2020; *395* : 565-74.
7. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, *et al*. The species Severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology Nature Publishing Group* 2020; *5* : 536-44.
8. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect Genet Evol* 2020; *79* : 104212.

9. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* 2020. doi: 10.1038/s41564-020-0688-y.
10. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* 2005; 3*09* : 1864-8.
11. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J Virol* 2020; *94*. pii: e00127-20.
12. World Health Organization. Coronavirus disease (COVID-2019) situation reports. WHO; 2020. Available from: *https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports*, accessed on February 29, 2020.
13. World Health Organization. Coronavirus disease (COVID-19) technical guidance: Laboratory testing for 2019-nCoV in humans. Available from: *https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance,* accessed on February 29, 2020.
14. Ministry of Health & Family Welfare, Government of India; 2020. Available from: *https://mohfw.gov.in/node/4904*, accessed on February 18, 2020.
15. Yadav PD, Albariño CG, Nyayanit DA, Guerrero L, Jenks MH, Sarkale P, *et al*. Equine Encephalosis Virus in India, 2008. *Emerg Infect Dis* 2018; *24* : 898-901.
16. Yadav PD, Whitmer SLM, Sarkale P, Ng TFF, Goldsmith CS, Nyayanit DA, *et al*. Characterization of novel reoviruses [Wad Medani virus (Orbivirus) and Kundal (Coltivirus)] collected from hyalomma antolicum ticks in India during during surveillance for Crimean Congo Hemorrhagic fever. *J Virol* 2019. doi:10.1128/JVI.00106-19.
17. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 2017; *1* : 33-46.
18. Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016; *33* : 1870-4.
19. Andersen K, Rambaut A, Lipkin I, Holmes EC, Garry R. The proximal origin of SARS-CoV-2. Available from: *http://virological.org/t/the-proximal-origin-of-sars-cov-2/398*, accessed on February 24, 2020.
20. Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect Dis* 2020. pii: S1473-3099(20)30113-4.
21. To KK, Tsang OT, Chik-Yan Yip C, Chan KH, Wu TC, Chan JMC, *et al*. Consistent detection of 2019 novel coronavirus in saliva. *Clin Infect Dis* 2020. pii: ciaa149.
22. Volz E, Baguelin M, Bhatia S, Boonyasiri A, Cori A, Cucunubá Z, *et al*. Report 5: Phylogenetic analysis of SARS-CoV-2. Available from: *https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-phylogenetics-15-02-2020.pdf*, accessed on February 24, 2020.
23. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun* 2019; *10* : 2342.
24. Buchholz UJ, Bukreyev A, Yang L, Lamirande EW, Murphy BR, Subbarao K, *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* : 9804-9.
25. O'Neil D, Swanton C, Jones A, Medd PG, Rayment N, Chain B. IFN-γ down-regulates MHC expression and antigen processing in a human B cell line. *J Immunol* 1999; *162* : 791-8.
26. Janice Oh HL, Ken-En Gan S, Bertoletti A, Tan YJ. Understanding the T cell immune response in SARS coronavirus infection. *Emerg Microbes Infect* 2012; *1* : E23.
27. 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.

For correspondence: Dr Sarah S. Cherian, Department of Bioinformatics, ICMR-National Institute of Virology, 20-A, Dr Ambedkar Road, Pune 411 001, Maharashtra, India
e-mail: cheriansarah@yahoo.co.in