Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Indian Journal of Medical Microbiology 39 (2021) 73–80

Contents lists available at ScienceDirect

Indian Journal of Medical Microbiology

journal homepage: www.journals.elsevier.com/indian-journal-of-medical-microbiology

Original Research Article

TSP-based PCR for rapid identification of L and S type strains of SARS-CoV-2

Biswajyoti Borkakoty [a,*], Nargis K. Bali [b]

[a] Indian Council of Medical Research-Regional Medical Research Centre for NE Region, Bokel, Dibrugarh, Assam, 786010, India
[b] Department of Clinical Microbiology, Sher-I Kashmir Institute of Medical Sciences, Soura, Srinagar, Jammu & Kashmir, India

ARTICLE INFO

Keywords:
TSP-PCR
SARS-CoV-2
L and S type
RT-PCR

ABSTRACT

Background: In the initial few months of the COVID-19 pandemic, two distinct strains of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) were identified (L and S strain) based on a tightly linked SNP between two widely separated nucleotides at location 8782 (ORF1ab T8517C) and position 28,144 (ORF8: C251T, codon S84L).

Materials and methods: A Type Specific Primer based one step RT-PCR (TSP-PCR) test to distinguish the L and S type strains of SARS-CoV-2 without the need for viral genome sequencing, was developed. The study also analyzed 18,221 whole genome sequences (WGS) available up to April 2020 to know the prevalence of L and S type of strains. Phylogenetic and recombination analysis of SARS-CoV-2 genome with nearest animal and human coronaviruses were analyzed using MEGA X and SimPlot version 3.5.1 software respectively.

Results: The rapid TSP-PCR distinguished the L and S type strains of SARS-CoV-2 by amplifying a specific 326 bp and 256 bp fragment of the L and S type strain respectively. The test was used to analyzed 120 random SARS-CoV-2 positive samples from Assam, India among which 118 were found to be of L-type strains only. On analysis of 18,221 WGS, it was found that L type was the predominant strain with an overall prevalence ~90%. However, pockets of high prevalence of S-type strains (>35%) were still in circulation in Washington region in April 2020. The study did not detect any significant recombination events between closely related coronavirus and SARS-CoV-2.

Conclusion: TSP-based PCR for identification of circulating strains of SARS-CoV-2, will add in rapid identification of strains of COVID-19 pandemic to understand the spread of the virus, its transmissibility and adaptation into human population. Though, the S-type strains have decreased drastically across the globe since April 2020, the role of TSP-PCR in geographical niches where such strains are still prevalent may help in rapidly distinguishing the strains and study its evolution.

1. Introduction

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2 or HCoV-19) responsible for the coronavirus disease 2019 (COVID-19), first identified in December 2019 in Wuhan, China has caused the deadliest pandemic in last hundred years since the Spanish flu of 1918. It is the seventh coronavirus to infect human beings after the four common cold viruses (HCoV-229E, HCoV-OC43, HCoV-NL-63 and HCoV-HKU1) causing mild disease and the three severe disease-causing coronavirus outbreaks (SARS-CoV and MERS-CoV and now SARS-CoV-2) in 2003–03 and 2012 and 2019–20 respectively [1]. SARS-CoV-2 is a *Betacoronavirus* under subgenus *Sarbecovirus* (linage B) with a 29.9 kb + ssRNA genome having the nearest similarity of 96.2% with a bat coronavirus (RaTG13 strain) isolated from a horse-shoe bat (*Rhinolophus affinis*) from Yunnan province of China in 2013 [2,3]. It has 16 non-structural proteins (Nsp-1 to 16) coded by the ORF1ab and four major structural proteins namely: Spike protein (S) that plays a crucial role in binding to host ACE2 receptor for entry, envelope protein (E), membrane protein (M), nucleocapsid protein (N) and other accessory proteins [4].

It is estimated that the SARS-CoV-2 has recently jumped from animal host to human during the later months of 2019 (time to most common recent ancestor (tMCRA) Oct 6- December 11, 2019) [5]. There has been rapid diversification of the virus forming clades and subclades or haplotype clusters over the past six months but overall divergence is still low (<0.1%). Host factors will be playing a role in adaptation of the virus and the rapid global spread of the virus provides it with ample opportunity for natural selection for diversification into clades/clusters/linages. The selective pressure will favor one clade over the

* Corresponding author. ICMR-Regional Medical Research Centre, Dibrugarh, Assam, 786010, India.
E-mail address: biswaborkakoty@gmail.com (B. Borkakoty).

https://doi.org/10.1016/j.ijmmb.2021.01.003
Received 12 August 2020; Accepted 17 December 2020
Available online 15 January 2021

other including difference in virulence across different geographic regions of the world. A recent study performing phylogenetic network analysis found three major phylogenetic clusters named A, B and C with many sub clusters [6]. Another recent study, based on presence of three mutations including one in the spike protein D614G (a A-to-G base change at position 23,403 with reference strain resulting in change of amino acid from Aspartic acid to Glycine) labeled as 'G' Clade has emerged as the dominant and highly transmissible strain replacing the other initial clades of the virus [7]. Various whole genome phylogenetic network studies analyzing the evolution of the virus is recording up to 11 major phylogenetic clades based on temporal evolution [8]. Tang et al. in February 2020, analyzing 103 complete SARS-CoV-2 genomes, have found a tightly linked SNPs between two widely separated nucleotides at location 8782 (ORF1ab T8517C) and position 28,144 (ORF8: C251T, codon S84L). They categorized the SARS-CoV-2 virus into two major circulating strains 'S' type for serine and 'L' type for leucine at position 84 in ORF8, respectively [9]. Further they documented that L-type strains are the dominant strain circulating (~70%) and S-type the minor type and the ancestral type [9]. The significance with context to its difference in transmissibility, aggressiveness or virulence are yet to be established.

In our present study, we designed a Type-Specific Primer based PCR (TSP-PCR) by one-step RT-PCR based on discrimination of either 'T' or 'C' at position 28,144 of SARS-CoV-2, which can rapidly identify within 3 h the type of strains (L or S type) circulating in the region. This low-cost and rapid diagnosis of the strains types without viral genome sequencing the whole genome or partial genome has applicability in monitoring the strains circulating in different regions of the world.

2. Methods and materials

2.1. Design of conventional TSP-based PCR

Reference strain of SARS-CoV-2 (Ref seq: NC_045512.2) was downloaded from National Centre for Biotechnology Information (NCBI) nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2). A 366 bp long ORF8 region (27894–28259) along with partial N gene region was selected from the whole genome sequence of 29,903 bp for designing the primers for the TSP-PCR. Forward Primer for 'L' and 'S' type differed at the 3′ prime end at position 28,144 coding for Thymine (codon 84 leucine) for L-type primers and Cytosine (codon 84 serine) for 'S' type primer. The primers were designed manually and Tm value was kept at about 56 °C–60 °C for both the forward and reverse primers of each set. Initially, six different sets of primers were designed including primers with mutation of interest at exactly 3′ prime end, penultimate position and on 3rd nucleotide from the 3′ prime end. Standardization protocols were run and best primer set that can distinguish S-and L-type strains were selected (supplementary Table-S1 shows all the primers sequences used and supplementary figure S1 shows the S-type primer gel run with three different sets of primer). Primers were checked for self-priming, primer hetro-dimer, hairpins and also similarity search against human genome in nucleotide BLAST in NCBI database. The selected TSP-based PCR primer sequences are shown in Table-1. Primers were synthesized commercially (Eurofins, Bangalore, India). For the positive control for 'L' and 'S' type strain, gene fragment synthesis of 400 bp long each covering the critical mutation at position 28,144 and also covering both forward and reverse primer was synthesized commercially (Eurofins, Bangalore, India). The DNA gene fragment were reconstituted with 10 mM Tris buffer pH 7.4. Final working stock of 5 ng/μl were aliquoted and stored in −20 °C. Thermal profile and primer concentrations were standardized using a one-step RT-PCR kit (Qiagen OneStep RT-PCR kit, Hilden, Germany) in a gradient thermal cycler system (Veriti 96 well, ABI, ThermoFisher Scientific, USA). Nasopharyngeal/oropharyngeal swab sample in VTM were extracted using nucleic acid a commercial spin-column based extraction kit (QIAamp Viral RNA Mini Kit, Qiagen, Hilden, Germany), and 120 SARS-CoV-2 positive samples with C_T-value below 30 cycles detected by qRT-PCR (TaqMan assay) were selected for typing of 'L' and 'S' strain for identification by TSP-PCR. Further, another 120 SARS-CoV-2 negative samples were also tested with the in-house designed TSP-PCR to assess the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the test. Moreover, to check for cross-reactivity with other respiratory pathogens, the in-house-designed TSP-PCR assay was performed with available strains of HCoV-OC43 (8 samples), HCoV-NL63 (2 samples), HCoV-229E (2 samples), Human rhinovirus (32 samples), respiratory syncytial virus (10 samples), Influenza A/H1N1/pdm09 (30 samples) and Influenza B (20 samples) in the author's laboratory.

Table 1
Type specific primer (TSP) sequence designed to distinguish the L and S type strains of SARS-CoV-2.

Primer Name	Primer Sequence 5′—3′ direction	Position in relation to (Ref Seq)	Amplicon size
1. Set-1 ORF8-L-FP-326	TCGGTAATTATACAGTTTCCTGTTT	28120–28144	326 bp
2. Set-1 ORF8-L-RP-326	GAGTGAGAGCGGTGAACCAA	28445–28426	
3. Set-1 ORF8-S-FP-256	TCGGTAATTATACAGTTTCCTGTTC	28120–28144	256 bp
4. Set-1 ORF8-S-RP-256	CCCACTGCGTTCTCCATTCT	28375–28356	

In brief, RT-PCR was performed in a final concentration of 25 μl volume, with 5X Qiagen OneStep RT-PCR buffer 5 μl, dNTP mix 1 μl, forward & reverse primer 1 μl each at final concentration of 0.4 μM, Qiagen OneStep RT-PCR enzyme mix 1 μl, nuclease free water 11 μl and sample template 5 μl. A 40 cycle PCR was performed with the following thermal cycling condition: Reverse transcription at 55 °C for 30 min, initial PCR activation at 95 °C for 15 min, 40 cycle of denaturation at 95 °C for 30s, annealing at 52 °C for 30s, extension at 72 °C for 45s and final extension at 72 °C for 10 min. PCR amplicons were loaded in a 2% agarose gel with ethidium bromide (1:10,000 concentration), 100 bp DNA ladder, positive control (gene fragment for L and S) and negative control (NFW) and electrophoresed in a horizontal electrophoresis system with TAE buffer at 100 mA for 60 min. PCR products were visualized in a gel documentation system (myECL imager, ThermoFisher Scientific, Waltham, USA). PCR amplified a specific 326 bp amplicon for the L type positive control/samples and 256 bp long fragment for the S type positive control. One selected L-type PCR product was DNA sequenced in a Genetic analyzer (ABI Genetic analyzer 3500, ThermoFisher, Scientific, Waltham, USA) available in the institute. The sample was submitted to GenBank (Accession number MT429168).

A whole genome phylogenetic analysis of 33 sequences (32 full genome and one partial sequence MT429168/India/Assam/JMC4) was performed to distinguish the L-type and S-type strains phylogenetically. Further, evolutionary history of the L and S-type were also studied to know if the S-type of strains forms a unique clade and to know the interclade pairwise-distance with L-strains. In brief, 32 complete genome sequences representative of 'L' and 'S' type strains along with strains labeled as clade G and reference strains (Ref seq: NC_045512.2) were downloaded from GISAID (Global Initiative on Sharing All Influenza Data) and NCBI GenBank. Phylogenetic and molecular evolutionary analyses were conducted using MEGA version X [10]. The sequences were aligned by MUSCLE multiple alignment in the MEGA X software and maximum likelihood fits of 24 different nucleotide substitution model were checked for using the best fit model. The model with the lowest

Akaike Information Criterion, corrected score was found to be General time reversal with Gamma distribution and invariant sites (GTR + G + i), which was selected for performing the phylogenetic analyses using maximum likelihood method with 500 bootstrap replicates.

To further know the prevalence of 'L' and 'S' type strains circulating in the world, 18,221 whole genome sequence of SARS-CoV-2 with high coverage was downloaded from the GISAID database based on the date of collection of samples and segregated into month-wise data from January 1, 2020 to April 30, 2020. The month-wise data were uploaded into MEGA X version software and screened for thymine or cytosine at position 28,144 based on the reference sequence (Ref seq: NC_045512.2). Further, specifically 1938 whole genome sequences of SARS-CoV-2 with high coverage from Washington DC, USA (as this geographical region was reporting higher prevalence of S-type strains) were also analyzed month wise to document the prevalence of S and L type strains.

Further a total of 25 full genomes of Coronavirus of all genera (that served as reference sequences) were downloaded from NCBI nucleotide database (https://www.ncbi.nlm.nih.gov/nuccore/?term=). Full-genome multiple alignment was performed by MUSCLE program available under MEGA version X software. Evolutionary history of the complete genome sequence was performed to construct a phylogenetic tree based on maximum likelihood method based on the General Time Reversible model with gamma distribution and evolutionary invariable sites (GTR + G + i) as per model testing in MEGA X software. The final phylogenetic analysis was performed in the same software using 500 bootstrap replicates. Also to know if any recombination events occurred in the SARS-CoV-2 genome with the closely related coronavirus genomes, we performed a similarity analysis of SARS-CoV-2 (Ref seq: NC_045512.2) with the full genome sequence of five closely related strains from bats, pangolins and humans (hCoV-19/bat/Yunnan/RaTG13/2013|EPI_ISL_402131, Bat-SL-CoVZC45 MG772933.1, hCoV-19/pangolin/Guandong/1/2019|EPI_ISL_410721; hCoV-19/pangolin/Guangxi/P2V/2017| EPI_ISL_410542 and SARS-CoV NC_004718.3) using SimPlot version 3.5.1 software.

3. Result

The TSP-PCR conducted on random 120 (qRT-PCR SARS-CoV-2 positive) samples with C_T-value below 30 cycles collected from the state of Assam (between 30th March to June 5, 2020), India, showed that all 118 out 120 samples were of L-type strains, while two samples were not amplified. The TSP-PCR didn't detect any S-type strains and only the positive control of S-type was amplified during the PCR run (see Fig. 1A). The L-type strains specifically were amplified by the L-type primers and amplified a specific 326 bp long segment covering the critical position 28,144 of SARS-CoV-2. Further, to prove that the specific region of concern was amplified during TSP-PCR, a L-type sample was DNA sequenced and was found to be indeed L-type sequence with thymine (Uracil in the virus genome) at position 28,144 instead of cytosine. The sequence was submitted to GenBank (Accession id: MT429168) (see Fig. 1B). Moreover, 120 negatives samples in qRT-PCR were also found to be negative in the in-house-designed TSP-PCR test. The sensitivity of the assay compared to qRT-PCR for detecting SARS-CoV-2 with C_T-value within 30 cycles was 98.3%, while specificity was 100%, and PPV and NPV were 100% and 98.3% respectively. Also, the TSP-PCR did not amplify other know positive samples for Influenza A & B, HCoV-OC43, HCoV-NL63, HCoV-229E, Human rhinovirus, and Respiratory syncytial virus.

Further, A total of 18,221 whole genome sequences with high coverage was downloaded from the GISAID database which covered all the sequences deposited from January to April 2020. Table- 2 displays the month wise data on the distribution of L and S type strains globally. It shows that L-type was the predominant type strain circulating in the world and the S-type strain frequency decreased gradually over few months of the outbreak. In January 2020, the S-type strain had a global prevalence of ~36%, then it decreased to ~26% in February, ~12% in March and 3.9% in April 2020 (see table-2). Further, separately 1938 complete genome sequence of SARS-CoV-2 was analyzed from strains collected from Washington DC, USA as this region had shown a niche for the S-type strains in USA. It was seen that 51.6% of strains from January to April 2020, in Washington DC, USA were of the S-type strains (see

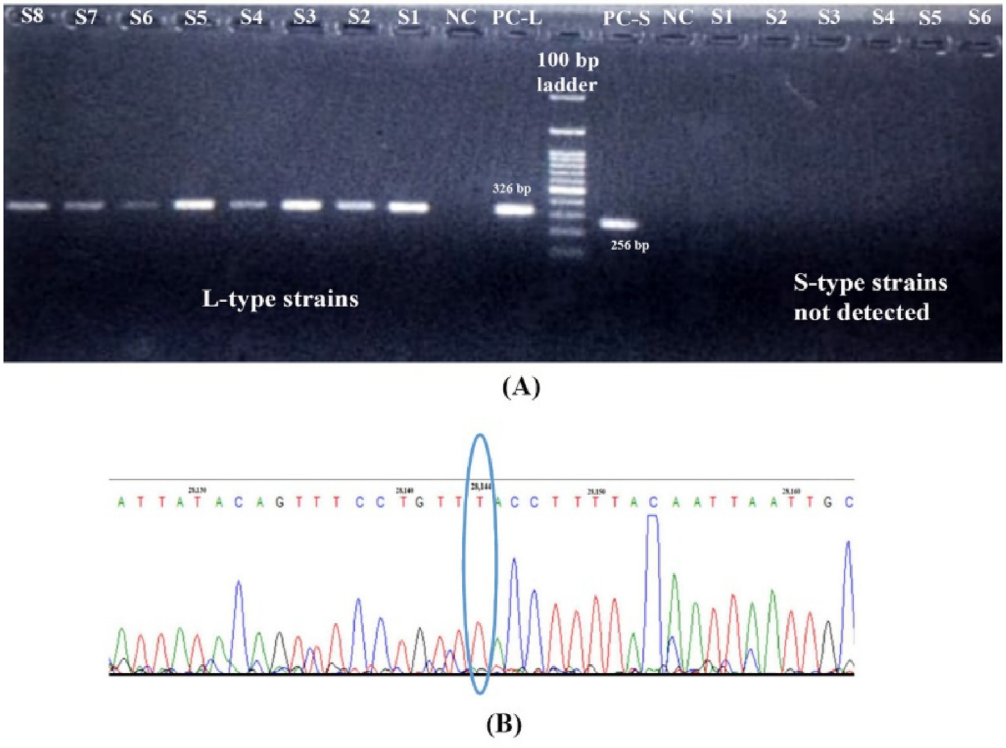

Figure-1. (A & B): The Type Specific Primer based conventional PCR (RT-PCR) distinguishing the L-Type and the S-type strains of SARS-CoV-2. Figure-1(A) shows the agarose-gel electrophoresis of the specific L-type strains with an amplicon size of 326 bp long and S-type positive control shows a length of 256 bp amplicon. Figure-1 (B) shows the raw sequence chromatogram of the sequence of interest of the DNA sequenced L-type strain (GenBank Accession number MT429168). Nucleotide position 28,144 is 'Thymine' that places the strain to L-type.

Table 3). Though the frequency of S-type strains dropped from 100% in January 2020 to 35.7% in April 2020, it remained high in Washington DC, USA compared to other geographical regions of the world. Further the authors of this study also screened 2311 whole genome sequences with high coverage submitted in GISAID from India up to September 2020 and found that only 77 out of 2311 sequences (3.3%) were of S-type while all the rest were L-type (not shown in tables).

Moreover, the complete genome phylogenetic analysis of 33 sequences (32 complete genome and one partial) reveals the type-S strains forming a distinct clade (S-clade/19B), while the D614G (clade-G) forming another distinct clade. Further, the ancestral L-type strains diverged into clade G and other clades (see figure- 2). The overall, pairwise distance calculated using the maximum composite likelihood model showed a maximum p-distance of 0.0006 (0.06% dissimilarity) among the 32 full genome sequence studied in comparison to the reference strain. The tree in Fig. 2 is rooted to the S-type clade (ancestral clade). The reference strain (NC_045512.2, collected in Wuhan in December 2020) shows the closest sequence similarity to sequence isolated in Jiangshu province, China, in January 2020 (p-distance: 0.000000) followed by strains collected in January 2020 from Guangdong province, China (p-distance: 0.00001) and Italy (p-distance: 0.00006) collectively label as 19A clade. The inter group or inter clade p-distance of S type, D614G (clade G or A2a clade) and rest L type with reference to the 19A clade L-type strains are as follows: S clade type-S has a p-distance of 0.00018 or 0.018% dissimilarity with the 19A clade. The D614G (G clade) showed an average 0.00025 (0.02%) p-distance, while the later L-type that had formed a separate clade (B6/L-type) has an average of 0.00023 or (0.02%) pairwise distance from the L-type 19A clade (see Fig. 2).

Further, it seems the evolution of mutation at site D614G in the spike protein may have occurred independently in both L-type and S-type strains. It was observed that D614G mutation was seen in S-type strains collected in February 2020 from Wuhan, China, but it still formed a single clade with the rest of the S-type strains. While the D614G that evolved from the original L-type strains has formed a distinct clade G as seen in the phylogenetic analysis of 32 whole genome sequencing data (see figure-2).

Phylogenetic analysis of the 25 complete genome sequence of SARS-CoV-2 with nearest strains and reference strains of *Betacoronavirus, alphacoronavirus, gammacoronavirus* and *deltacoronavirus* is shown in figure-3. The SARS-CoV-2 strains had the closest nucleotide similarity (96.3%) with RaTG13 strain isolated (labeled green in figure-3) from the faecal swab of a bat (*Rhinolophus affinis*) from Yunnan province of China in 2013 (Strain Yunnan/RaTG13; GISAID accession ID: EPL_ISL_402131). It was observed that the RaTG13 strain was the most closely related strain to SARS-CoV-2, followed by coronavirus from Malayan pangolins (~91% similarity not shown in figure-3). Other more closely related strains were also from bats (strain ID: Bat-SL-CoVZ45 and Bat-SL-CoVZXC21) which had a sequence similarity of ~89%. Distance matrix analysis using the basic pairwise distance (p-distance) showed a dissimilarity of 19.7% and 43.8% of SARS-CoV-2 from SARS-CoV (labeled in yellow in figure-3) and MERS-CoV (labeled in blue in figure-3) respectively.

Table 2
Global distribution of L and S type strains of SARS-CoV-2 between January 2020 to April 2020 (Total genome 18,221).

Geographic region: Whole World			
Month of collection of samples as per GISAID	SARS-CoV-2 whole genome with high coverage	Total S-type (serine at codon 84 in ORF8) N (%)	Total L-type (Leucine at codon 84 in ORF8) N (%)
January 2020	313	113 (36.1%)	200 (63.9%)
February 2020	567	148 (26.1%)	418 (73.7%)
March 2020	11,602	1374 (11.8%)	10,225 (88.1%)
April 2020	5739	214 (3.9%)	5523 (96.2%)
Total	**18,221**	**1849 (10.2%)**	**16,366 (89.8%)**

Table 3
Distribution of L and S type strains month wise (January to April 2020) in Washington DC, USA (total genome 1938 complete genome).

Geographic region: USA/Washington			
Month of collection of samples as per GISAID	SARS-CoV-2 whole genome with high coverage	Total S-type (serine at codon 84 in ORF8) N (%)	Total L-type (Leucine at codon 84 in ORF8) N (%)
January 2020	03	03 (100%)	0 (0%)
February 2020	54	52 (96.3%)	02 (3.7%)
March 2020	1223	727 (59.4%)	496 (40.6%)
April 2020	658	235 (35.7%)	421 (64%)
Total	**1938**	**1017 (51.6%)**	**919 (47.4%)**

The recombination analysis using SimPlot version 3.5.1 software showed that throughout the full genome of SARS-CoV-2, RaTG13 strain was the closest virus strain with the reference sequence of SARS-CoV-2 (Fig. 4 B). A major dissimilarity is seen in the spike protein region of SARS-CoV-2 compared to the other analyzed strains (Fig. 4A & B). There was no evidence of recent recombination events in SARS-CoV-2 strains with RaTG13 and other analyzed strains on Bootscan analysis performed in SimPlot software (see Fig. 5).

4. Discussion

The coronavirus disease, 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has spread rapidly globally causing 16.8 million cases and over 6,62,000 deaths globally till July 30, 2020 i.e. within 7 months of the start of the pandemic [11]. An unprecedented volume of scientific literature scrutinizing the virus and the pandemic, at a rate never seen before has flooded the scientific community. Researchers across the globe are studying the evolution of the virus and monitoring its characteristics including transmissibility and virulence. In February 2020, a group of researchers from China had initially classified the SARS-CoV-2 into two major strains, notably L and S based on a tightly linked SNPs between two widely separated nucleotides at location 8782 (ORF1ab T8517C) and position 28,144 (ORF8: C251T, codon S84L) [9]. The S type strains differed from the early L-type strains (19A clade) by 0.018% in whole genome nucleotide sequence in our analysis. Tang et al. documented that the S type strain was most likely the ancestral version of SARS-CoV-2 based on SNPs at position 8782 (orf1ab) and 28144 (ORF8), as those SNPs (similar to S type strains) were also seen in animal coronavirus including the nearest bat strain RaTG13, GD pangolin-CoV, GX pangolin-CoV, Bat SARS Sr–CoV-ZXC21 and Bat SARS Sr-CoVZC45 [9]. Though the frequency of S type of strains has drastically decreased within first four months of the outbreak from a global prevalence of ~36% in January to ~ 4% in April 2020, there are niches of the S-type strain circulating at a higher frequency ~36% in April 2020 in some geographical regions of the world such as Washington DC region. However, only 3.3% (77/2311) whole genomes submitted from India in GISAID database were of the S-type strains. The present study designed a one-step RT-PCR assay based on type specific primer targeting the crucial nucleotide at position 28144 to distinguish the strains within 3 h' time without performing a DNA sequencing of the virus. This would be helpful to molecular epidemiologist to distinguish the L and S type of strains within hours. Though the frequency of S-type strains has decreased significantly since April 2020, there may be niches and isolated geographical regions where S-type strains may yet be prevalent at a significant number. Genetic sequence data or strain typing is increasingly recognized as an important tool in infectious disease epidemiology [5]. SARS-CoV-2 RNA genome is estimated to have a mutation rate of $\sim 6 \times 10^{-4}$ nucleotide/genome/year which is similar with other RNA viruses and has accumulated only a moderate genetic diversity since its outbreak with an average pairwise difference of 9.6 SNPs between any two genomes [5]. Our analysis of 18,221 whole genome sequences (Jan to April 2020) showed that L-type strains with 89.8% prevalence was the

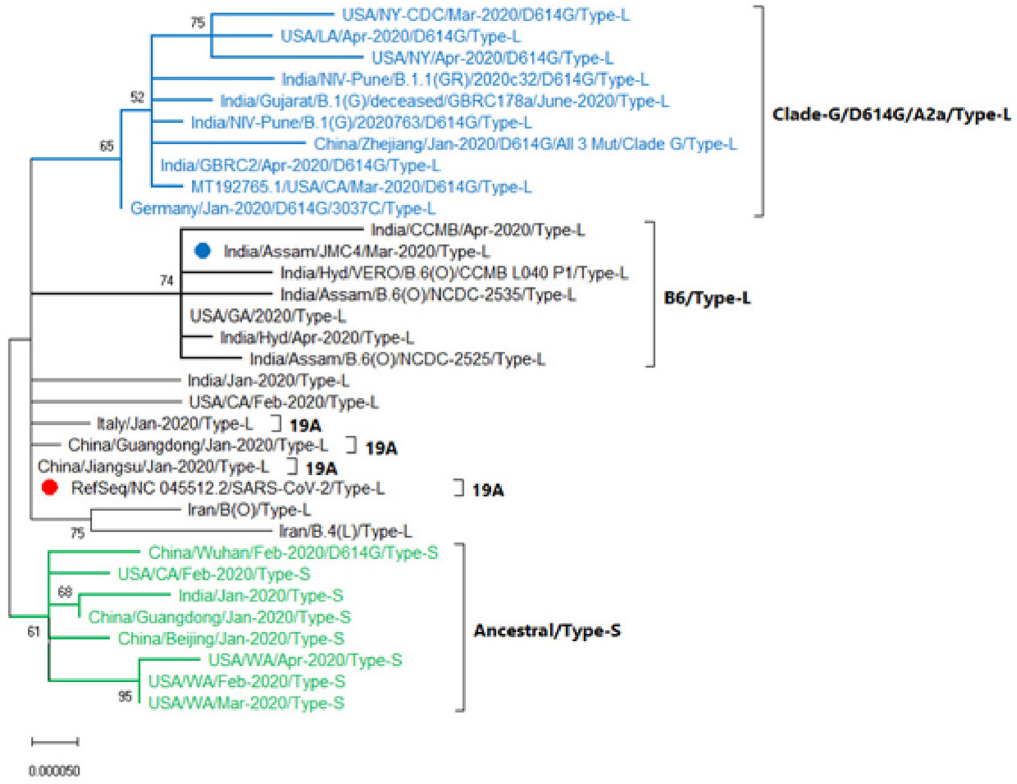

Figure-2. Phylogenetic analysis involving 33 nucleotide sequences of SARS-CoV-2 (32 full genome ~29 kb and one partial sequence.) The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model after conducting a best fit model. The tree is rooted to the ancestral S-type clade (labeled green). A discrete Gamma distribution was used to model evolutionary rate differences among sites. A 500 number of bootstrap replicates was performed to arrive at a consensus tree. Evolutionary analyses were conducted in MEGA X [10].

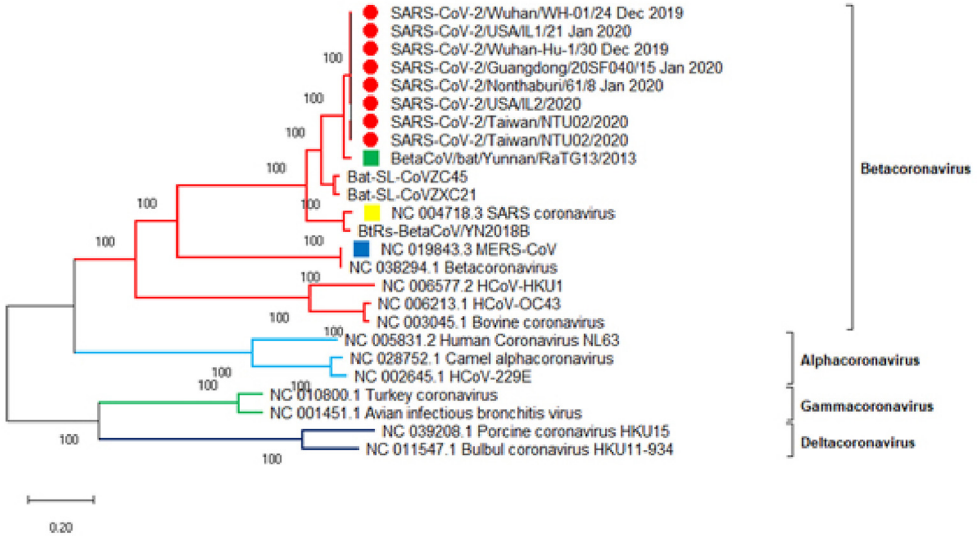

Figure-3. Full genome molecular phylogenetic analysis of the SARS-CoV-2 was performed using the Maximum Likelihood method based on General time reversible model with gamma distribution and evolutionary invariable sites as per model testing in MEGA X software. The analysis involved 25 complete genome sequences (size ~ 26–31 kb) including reference sequences of all four coronavirus genera and nearest sequences of the SARS-CoV-2. A 500 number of bootstrap replicates was performed to arrive at a consensus tree. Evolutionary analyses was conducted in MEGA X [10].

predominant type circulating across the globe and S-type with 10.2% was the minor prevalent strain up to April 2020. Further analysis using representative 32 whole genome sequence from January to May 2020, including strains of early L and S type, Clade G (D614G), B6/Type-L showed a distinct clade differentiation from the S-type strains which too formed a separate clade from the L-type strains. Further, based on presence of the unique SNPs characterizing the S-type strains in animal coronavirus including the RaTG13 strain or GD_Panglong-CoV strains, the S-type strains may be the ancestral SARS-CoV-2 that jumped species barrier [9]. Now world over, the S-type strains has been smoothly being replaced by the predominant L-type strains which have further diverged into Clade G or A2a (D614G) clade and other clades.

By the end of June 2020, the predominant strain of SARS-CoV-2 worldwide had the D614G mutation known as the G-variant or G-clade [7]. The change in amino acid from Aspartic acid (D) to Glycine (G) at position 614 of spike protein is found to have 3–9 times more infectivity in in vitro cell line study using pseudo virus [12]. Further, researchers found no worse clinical outcome in people infected with G clade of SARS-CoV-2, however they tend to transmit at slightly faster rate [12]. A recent study from India (preprint available in BioRxiV), analyzing 104 whole genome sequences reveal that SARS-CoV-2 strains from India came in three wave clusters and can be grouped into two major clades and one minor sub-clade. They found that 25% of Indian strains belong to the clade G (A2a) while 62.5% belonged to unclassified cluster which

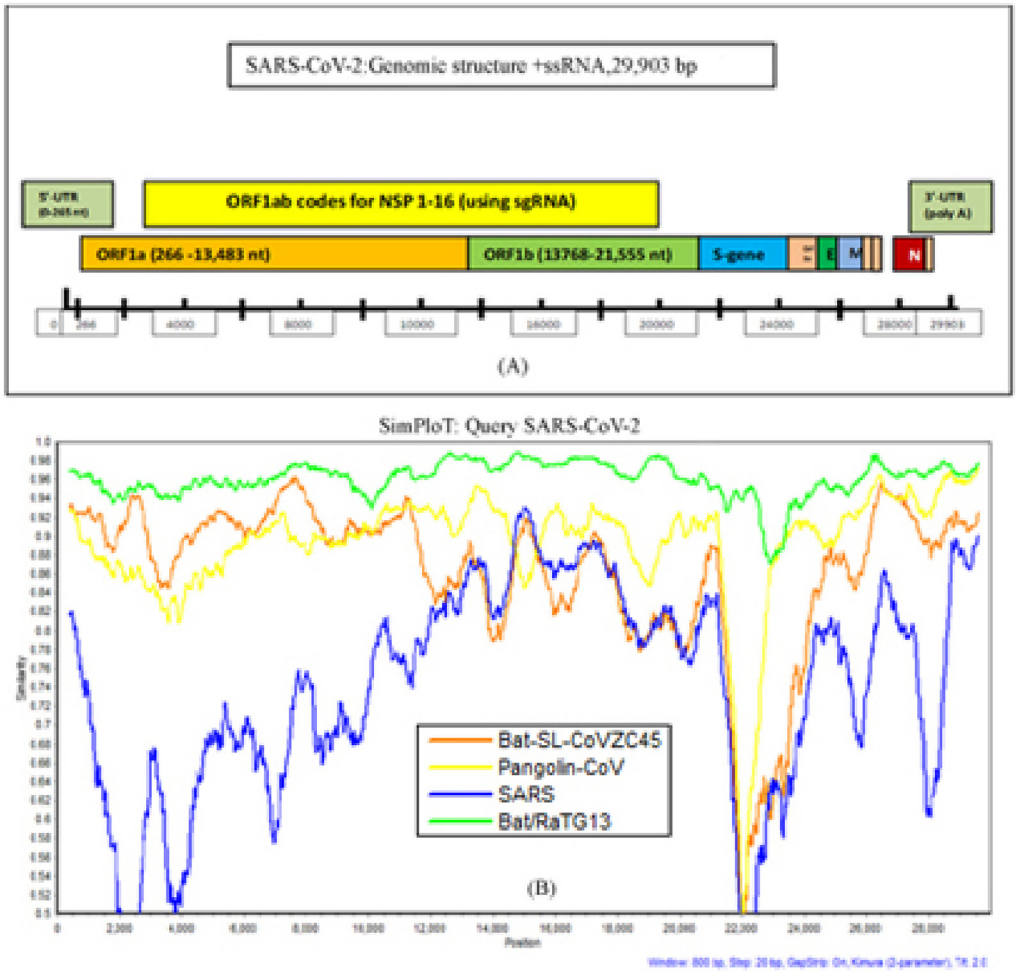

Figure-4. (A &B): Genomic organization and similarity analysis: (A): Genomic Organization of SARS-CoV-2, +ssRNA 29,903 bp genome. (B): Similarity plot (SimPlot version 3.5.1 software) analysis of five full genomes of closely related genomes of SARS-CoV-2 including strain Bat/RaTG13 (closest), SARS-CoV, bat-SL-CoVZC45 and Pangolin-CoV strain from Guangdong province, China.

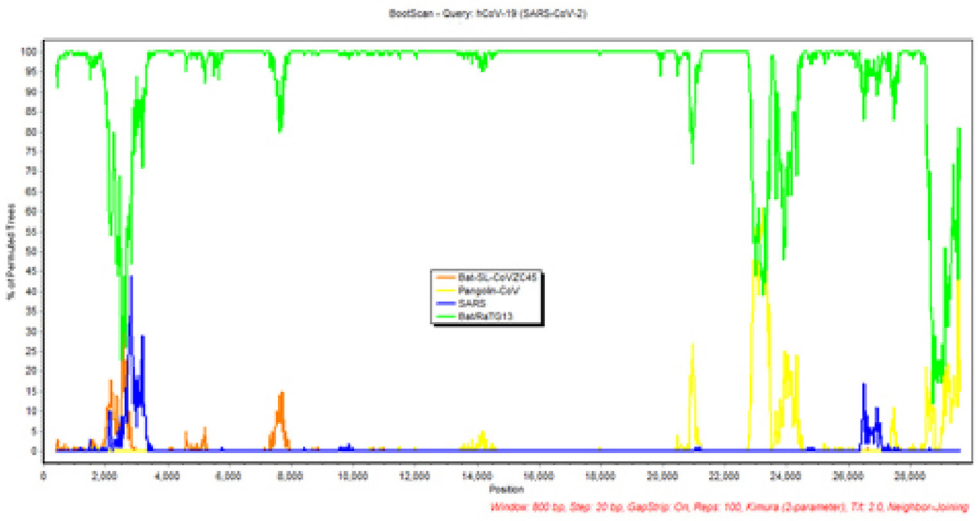

Figure-5. Bootscan Plot: Recombination event analysis was performed for full genome sequences of five closely related strains of SARS-CoV-2 including strain Bat/RaTG13, SARS-CoV, bat-SL-CoVZC45 and Pangolin-CoV strains from Guangdong province, China. Analysis was performed in SimPlot version 3.5.1 software, keeping a bootscan window of 800 bp, a 20 bp Step using Kimura 2-parameter algorithm and a 100 bootstrap replicates. No significant recombination event detected during the analysis.

they re-defined as A4 clade and a minor subclade (6.7%) classified under A3 clade [13]. Another study from central India reported that the D614G mutation (A2a clade/G clade) was the predominant type (~46%) followed by A4 and B clades [14]. A recent study from Odisha, India reported prevalence of four major circulating clades i.e. 19A (19.3%), 19B (17.8%), 20A (36.1%), 20B (29.7%), and one minor clade 20C (2.0%). Both the 20A and 20B clades predominantly contains the D614G mutation (G clade) which now seems to be predominant in the state of

Odisha, India too [15]. Yet another study identified 5775 distinct genome variants in roughly ten thousand genomes they analyzed from 68 countries. They identified six-major clades and 14 subclades with the D614G variant clade (Clade G) as the most common clade [16]. The study by Dorp et al. analyzing 7666 whole genome sequences of SARS-CoV-2, identified 198 recurrent mutations (homoplasies) that were associated with 290 amino acid changes across the genome and 80% of them produced non-synonymous changes [5]. Further, in their study they documented a strong homoplasy that lies in position 11083 (orf1a encoding Nsp6), which overlaps a putative immunogenic peptide predicted to result in both $CD4^+$ and $CD8^+$ T-cell reactivity [5,17]. Biswas et al. in their study analyzing 3636 complete genome sequences collected from 55 countries, have categorized the SARS-CoV-2 into 11 major clades based on temporal evolution and that the A2a (D614G) clade was spreading rapidly across the globe [8]. Several nomenclatures have been used for SARS-CoV-2 classification including by Nextstrain, GISAID, and Cov-lineages.org [18]. Clades based on GISAID divides the SARS-CoV-2 into 7 major clades (S, L, O, V, G, GH, GR), while clades based on Nextstrain categories SARS-CoV-2 into 19B (ancestral type), 19A, 20A, 20C, 20B and cov-lineages.org categories SARS-CoV-2 into just two major clades A and B and multiple lineages [18].

However, this temporal variation over time will adapt slowly into distinct clades or genotypes as of now the overall variations in nucleotide sequence of SARS-CoV-2 globally is still less than 0.1%. In our recombination event analysis in SimPlot version 3.5.1 software using the whole genome sequences of nearest animal and human coronaviruses, no significant recombination events with SARS-CoV-2 was detected. Our test performed on limited number of representative whole genome sequences cannot rule out the chance of recombination events with other closely related coronaviruses. Moreover, it is still difficult to rule out recombination event either natural or laboratory manipulated events unless all relevant strains are analyzed. On the basis of current sequence database, all HCoVs, SARS-CoV and MERS-CoV are presumed to have originated from either bats or rodents [4]. A paper published in Nature Medicine in 2015, had predicted a potential risk of SARS-like-CoV re-emergence from viruses circulating in the bat populations. The authors of the paper had developed a chimeric virus expressing the spike protein of a SARS-like virus (SHCO14–CoV) circulating in Chinese horseshoe bat population, with the backbone of mouse adapted SARS-CoV. This chimeric virus (SHCO14-MA15), can use multiple orthologs of SARS receptor for the human ACE-2 and replicate efficiently in primary human airway cells and additionally was found to cause disease in mouse lungs in *in vivo* experiments [19]. However, the genome sequence of the chimeric virus is not available in public database for scrutiny and analysis for comparison with SARS-CoV-2. More scientific data could swing the balance of evidence to favor laboratory manipulation & recombination event over a natural slow drift from animal coronavirus as the origin of SARS-CoV-2. Obtaining related viral sequences from animal sources would be the most definitive way of revealing viral origins and understanding cross-species spread which will further aid in preventing impending outbreaks.

Conclusion: The present study developed a TSP-PCR based one step RT-PCR test to distinguish the early S-type and L-type strains of SARS-CoV-2. There is a future role of TSP-PCR in rapidly classifying clades and sub-clades of SARS-CoV-2 including rapidly identifying the clade G (D614G) or other strains without the use nucleic acid sequencing data. The authors analyzing over 18,000 whole genome sequences across the globe till April 2020 concluded that L-type strains were the predominant strains (~90%) while S-type strains were diminishing quickly and were the minor strains at ~10% frequencies with niches of higher frequencies in specific geographical region. Although, a section of the scientific community does not recognize the existence of distinct types of SARS-CoV-2 (L and S type) solely on the basis of a two tightly linked SNPs [20], however evolutionary phylogenetic analysis puts the S-type into a separate clade from the rest of the L-types. Our study could not detect any significant recombination events in the SARS-CoV-2 with the closest animal and human coronaviruses. The COVID-19 pandemic is still evolving and the SARS-CoV-2 genome will evolve over time and a constant monitoring of the strains will be useful for understanding the origin, transmissibility, virulence, design of vaccines, drug targets and diagnostics assays.

Source of support

Intramural fund of ICMR-RMRC, Dibrugarh and Regional VRDL, Dibrugarh, Department of Health Research, New-Delhi.

Contributions

First author (BB) devised, designed and performed the laboratory test whereas the second author (NKB) performed online data analysis of 18,200 complete genome, performed phylogenetic analysis and drafted the manuscript. Both the authors contributed in writing the manuscript.

Ethics

The study was done under the Department of Health Research, Government of India funded scheme "Establishment of a network of Laboratories for managing Epidemics and Natural Calamities (VRDL)" at RMRC, Dibrugarh. The institutional ethics committee of Regional Medical Research Centre for NE region, Dibrugarh Assam, approved the scheme which allows the microbiological diagnosis free of cost for early diagnosis and investigations of outbreaks & epidemics, development of rapid kits for viral infections in the region, to enumerate different viral strains/genotypes, including their molecular epidemiology, and monitor genetic variations. The study also allows the use of left-over samples for anonymous testing for pathogens.

Declaration of competing interest

None. On behalf of both the authors, the corresponding & the lead author declare that there is no conflict of interest related to the submitted manuscript.

Acknowledgement for financial support & specific scientific contribution only

The authors acknowledge the funding from the Regional VRDL, Dibrugarh, Department of Health Research (DHR), Government of India.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmmb.2021.01.003.

References

[1] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26:450–2. https://doi.org/10.1038/s41591-020-0820-9. Available at: . [Accessed 8 May 2020].
[2] Zhou P, Yang X, Wang X, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3. https://doi.org/10.1038/s41586-020-2012-7. Available at: . [Accessed 5 May 2020].
[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. N Engl J Med 2020;382(8):727–33. https://doi.org/10.1056/NEJMoa2001017. Available at, . [Accessed 5 May 2020].
[4] Chen Y, Liu Q, Guo D. Emerging coronaviruses: genome structure, replication, and pathogenesis. J Med Virol 2020;92(4):418–23. https://doi.org/10.1002/jmv.2568. Available at: . [Accessed 3 May 2020].
[5] van Dorp L, Acman M, Richard D, Shaw LP, Ford CP, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020; 83:104351. https://doi.org/10.1016/j.meegid.2020.104351. Available at: . [Accessed 8 May 2020].
[6] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA 2020;117(17):9241–3. https://doi.org/10.1073/pnas.2004999117. Available at: . [Accessed 10 May 2020].

[7] Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv 2020. https://doi.org/10.1101/2020.04.29.069054. Accessed May 10, 2020.
[8] Biswas NK, Majumder PP. Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. Indian J Med Res 2020; 151(5):450–8. https://doi.org/10.4103/ijmr.IJMR_1125_20. Accessed June 10, 2020.
[9] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020. https://doi.org/10.1093/nsr/nwaa036. Accessed June 11, 2020.
[10] Kumar S, Stecher G, Li M, Knyaz C, Tamura K, Mega X. Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.
[11] WHO. Coronavirus disease (COVID-19) situation report – 192. Available at: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200730-covid-19-sitrep-192.pdf?sfvrsn=5e52901f_8; 2020. Accessed July 31, 2020.
[12] Yurkovetskiy L, Wang X, Pascal KE, Tomkins-Tinch C, Nyalile TP, Wang Y, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell 2020;183(3):739–51. e8.
[13] Kumar P, Pandey R, Sharma P, Dhar MS, Vivekanand A, Uppili B, et al. Integrated genomic view of SARS-CoV-2 in India. bioRxiv preprint. https://doi.org/10.1101/2020.06.04.128751; 2020. this version posted June 4, 2020 (Accessed June 11, 2020).
[14] Sharma S, Dash PK, Sharma SK, Srivastava A, Kumar JS, Karothia BS, et al. Emergence and expansion of highly infectious spike: D614G mutant SARSCoV-2 in central India. bioRxiv 2020. https://doi.org/10.1101/2020.09.15.297846. preprint, . [Accessed 26 November 2020]. this version posted June 4, 2020.
[15] Raghav S, Ghosh A, Turuk J, Kumar S, Jha A, Madhulika S, et al. Analysis of Indian SARS-CoV-2 genomes reveals prevalence of D614G mutation in spike protein predicting an increase in interaction with TMPRSS2 and virus infectivity. Front Microbiol 2020;11(2847). accessed November 26, 2020.
[16] Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ 2020;98(7):495–504. https://doi.org/10.2471/BLT.20.253591. Accessed July 30, 2020.
[17] Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe 2020;27(4):671–80. https://doi.org/10.1016/j.chom.2020.03.002. e2, . [Accessed 30 July 2020].
[18] Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European region, january to June 2020. Euro Surveill 2020;25(32):2001410. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410. Accessed August 14, 2020.
[19] Menachery VD, Yount BLJ, Kari Debbink K, Agnihothram S, Gralinski LE, Plante JA, et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. Nat Med 2015;21(12):1508–13. https://doi.org/10.1038/nm.3985. Available at: . [Accessed 2 May 2020].
[20] MacLean OA, Orton RJ, Singer JB, Robertson DL. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evol 2020;6(1). https://doi.org/10.1093/ve/veaa034. veaa034. Published 2020 Apr 30, . [Accessed 28 July 2020].