Genomic variations in SARS-CoV-2 strains at the target sequences of nucleic acid amplification tests

Canhui Cao[1], Ruidi Yu[1], Shaoqing Zeng[1], Dan Liu[1], Wenjian Gong[1], Ruyuan Li[1], Siyuan Wang[1], Yuan Yuan[1], Jianhua Chi[1], Jiahao Liu[1], Yang Yu[1], Xiaofei Jiao[1], Guangyao Cai[1], Ning Jin[1], Fei Ye[2], Qinglei Gao[1]

[1]National Medical Center for Major Public Health Events, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China
[2]Department of Neurosurgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Submitted: 22 January 2021
Accepted: 7 February 2021

Arch Med Sci
DOI: https://doi.org/10.5114/aoms/133120

Corresponding authors:
Qinglei Gao
National Medical Center for Major Public Health Events
Tongji Hospital
Tongji Medical College
Huazhong University of Science and Technology
Wuhan, China
Phone: +86-27-83663351
Fax: +86-27-83662681
E-mail: qingleigao@hotmail.com

Fei Ye
Department of Neurosurgery
Tongji Hospital
Tongji Medical College
Huazhong University of Science and Technology
Wuhan, China
Phone: +86-27-83663351
Fax: +86-27-83662681
E-mail: yeyuanbei@hotmail.com

Wei Zhang
National Medical Center for Major Public Health Events
Tongji Hospital
Tongji Medical College of Huazhong University of Science and Technology
Wuhan, China
Phone: +86 27 83663351
Fax: +86 27 83662681
E-mail: zhangwei_tjh@qq.com

Abstract

Introduction: Nucleic acid amplification is the main method used to detect infections of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). However, the false-negative rate of nucleic acid tests cannot be ignored.
Material and methods: Herein, we demonstrated genomic variations at the target sequences for the tests and the geographical distribution of the variations across countries by analyzing the whole-genome sequencing data of SARS-CoV-2 strains from the 2019 Novel Coronavirus Resource (2019nCoVR) database.
Results: Among the 21 pairs of primer sequences in regions ORF1ab, S, E, and N, the total length of primer and probe target sequences was 938 bp, with 131 (13.97%) variant loci in 2415 (38.96%) isolates. Primer targets in the N region contained the most variations that were distributed among the most isolates, and the E region contained the fewest. Single nucleotide polymorphisms were the most frequent variation, with C to T transitions being detected in the most variant loci. G to A transitions and G to C transversions were the most common and had the highest isolate density. Genomic variations at the three mutation sites N: 28881, N: 28882, and N: 28883 were the most commonly detected, including in 608 SARS-CoV-2 strains from 33 countries, especially in the United Kingdom, Portugal, and Belgium.
Conclusions: Our work comprehensively analyzed genomic variations at the target sequences of the nucleic acid amplification tests, offering evidence to optimize primer and probe target sequence selection, thereby improving the performance of the SARS-CoV-2 diagnostic test.

Key words: nucleic acid amplification techniques, mutation, COVID-19, polymerase chain reaction, DNA primers, WGS, whole genome sequencing, genetic variations.

Introduction

The coronavirus disease 2019 (COVID-19) outbreak has resulted in a Public Health Emergency of International Concern since 30 January 2020, and has been classified as an ongoing pandemic since 11 March 2020 [1, 2]. Possible prevention, curative treatment and accurate diagnosis are in urgent need [3–5]. Currently, the diagnosis of COVID-19 relies on positive pathogenic testing, epidemiological exposure history,

computed tomography (CT) imaging features, and serological characteristics [6–8], as well as symptoms that are atypical [9–12]. Positive pathogenicity is commonly confirmed by nucleic acid testing via reverse transcription polymerase chain reaction (RT-PCR), virus isolation and sequencing, or specific antibody detection. Currently, nucleic acid testing is the primary method of diagnosing COVID-19 [10, 12, 13]. Hence, the sensitivity and specificity of nucleic acid testing are crucial. Although it is highly specific, the sensitivity is not satisfactory [14]. Various factors may interfere with test sensitivity [15–18]. However, few studies mention false negatives due to variations in primer target sequences.

RT-PCR kits have been designed to detect SARS-CoV-2 genetically [19]. Among the SARS-related viral genomes, there are three regions with conserved sequences: 1) the RdRP gene (RNA-dependent RNA polymerase gene) in the open reading frame ORF1ab region, 2) the E gene, and 3) the N gene [20]. Both the RdRP and E genes had high analytical sensitivity for detection, whereas the N gene provided poorer analytical sensitivity [10, 20]. The sensitivity of nucleic acid detection kits relies on the binding of primers and probes to the virus genome [10, 21]. High variation frequency, especially the variation near the 3′ end, might influence the primers or probes binding to the virus genome [22]. Thus, variations in the target sequence and how these variations affect the accuracy of the tests need to be elucidated.

Although most primer and probe sets were designed to focus on specific and conserved sequences, there were still some variations in the target sequences of the nucleic acid test assays of SARS-CoV-2 strains. Several studies have compared the sensitivity and efficiency of different SARS-CoV-2 detection assays [23, 24]. However, few people pay attention to the variations and geographic distribution that may affect the PCR performance of detection kits used on a large scale in the population. Herein, we comprehensively analyzed genomic variations in the target sequences of the nucleic acid amplification tests and their geographic distribution. Our results are expected to provide evidence for optimizing the selection of detection kits used on a large scale in the population, thereby improving the diagnostic accuracy of the tests. The study was approved by the Medical Ethical Committee of Tongji Hospital (TJIRB20200406).

Material and methods

2019 Novel Coronavirus Resource

We used the 2019 Novel Coronavirus Resource [25, 26], constructed by the Chinese Academy of Sciences. It integrates genomic and proteomic sequences as well as their metadata from the Global Initiative on Sharing All Influenza Data, National Center for Biotechnology Information, China National GeneBank, National Microbiology Data Center, and China National Center for Bioinformation/National Genomics Data Center. The comprehensive appendix data were available at HARVARD Dataverse (https://dataverse.harvard.edu/) using the Dataset Persistent ID: doi:10.7910/DVN/KZ4KBM, entitled Genomic Variations in SARS-CoV-2 Strains at the Target Sequences of Nucleic Acid Amplification Tests.

Mapping of genomic variations

The WHO offered molecular assay protocols for SARS-CoV-2 that have been shared (WHO, 2020), including seven institutes. The sequences of primers and probes were mapped against the SARS-CoV-2 isolate Wuhan-Hu-1, complete genome (NCBI Reference Sequence: NC_045512.2) with bowtie2 [27]. The positions of the primers and probes used in the nucleic acid testing kits were identified. Then, the sequences of the primers and probes were compared with the variant loci offered by 2019nCoVR, with the variant loci and corresponding isolates downloaded from the database.

Variation dynamic curve

Virus isolates with variations at each site divided by country at each time point were counted. The curve shows the trends of isolates from different countries over time at some variant loci [25, 26].

Variation density and isolate density calculation

The length of primers and probes and the number of variant loci covered by the primers and probes in each region were calculated first. The variation density of a region was the number of variant loci divided by the length of primers and probes. Also, the number of isolates presenting variations in the primer and probe target sequence was calculated. Isolate density was the number of isolates divided by the length of the target sequence.

The variance of time and regional variations

First, the frequency of population occurrence at each variant loci over time or country was identified by the Ensembl Variant Effect Predictor (VEP) [28]. Time variance was the variance based on the frequency of population occurrence at each time point, to assess the dispersion of changes at that

site. Taking the country as a unit, we calculated regional variance based on the frequency of population occurrence in each country, to evaluate the dispersion of variations at that location.

Statistical analysis

The majority of statistical analyses were performed using the SPSS software package version 22. The χ^2 test was used to compare the differences in counting data between groups. The variance was used to evaluate the degree of dispersion as described above. A two-sided *p*-value less than 0.005 was considered statistically significant.

Results

Variation landscape of primer and probe target sequences in the SARS-CoV-2 strains

Twenty-one pairs of primers and probes obtained from seven institutes worldwide were analyzed. The primers and probes were mainly focused on the replicase complex (ORF1ab), spike (S), envelope (E), and nucleocapsid (N) regions. The total length of the primer and probe target sequence was 938 bp, with 131 (variation density = 13.97%) variant loci. Also, 2415 (38.96%, isolate density = 2.57 per locus) of the 6198 SARS-CoV-2 strains from the 2019 Novel Coronavirus Resource (2019nCoVR) database had variations in the target sequence (Tables I and II).

There were 9 pairs of primers and probes (No. 1 to 9) in the ORF1ab region. This region was the region with the most primer and probe pairs. They covered 484-18909, including 42 variant loci and 271 SARS-CoV-2 strains, with variations at these sites. The most common variant loci in the ORF1ab region were 514 (87 virus strains, 32.10% of the 271 strains), followed by 13402, and 490 (Figure 1 A). As for the S region, there were only three pairs of primers covered from 24354 to 24900. There were 12 variant loci and 34 SARS-CoV-2 strains, with 24368 (14 virus strains, 41.18% of the 34 strains) being the most common variant loci (Figure 1 B). Interestingly, there were two pairs of primers and probes provided by two different institutes located in the E region, and the two pairs were the same. It covered 26269 to 26381. Barely five variant loci and eight SARS-CoV-2 strains were discovered. The most common variant loci were located at 26340, with four virus strains that had variation at this site, occupying 50% of the eight virus strains (Figure 1 C). Although the N region only had seven pairs of primers and probes covered from 28287 to 29282, it had the most variations (71 loci) and involved virus strains (2102, 5.8 per locus, 33.91% of the 6198 SARS-CoV-2 virus strains from the database) among the four regions. The most common variant loci in the N region were 28881 with 609 virus strains (28.97% of the 2102 virus strains), 28882 with 608 virus strains (28.92%), and 28883 with 608 virus strains (28.92%). These three variant loci had the most related isolates among the four regions (Figure 1 D).

The total length of the primer and probe target sequence of the ORF1ab region was 415 bp, which was the longest among the four regions. There were 42 mutation sites in these target sequences of ORF1ab (variation density = 0.10). The N region, the most mutative region, had the most variant loci (71 loci) and the highest variation density (0.19), with a 365 bp target sequence length. The primer or probe target sequences in the S region

Table I. Polymerase chain reaction kits for SARS-CoV-2

No.	Region		Primer or probe sequence	Target sequence	Location	Institute
1	ORF1ab	F	TTCGGATGCTCGAACTGCACC	TTCGGATGCTCGAACTGCACC	484-504	NIID Japan
		R	CTTTACCAGCACGTGCTAGAAGG	CCTTCTAGCACGTGCTGGTAAAG	874-896	
2	ORF1ab	F	CTCGAACTGCACCTCATGG	CTCGAACTGCACCTCATGG	492-510	NIID Japan
		R	CAGAAGTTGTTATCGACATAGC	GCTATGTCGATAACAACTTCTG	816-837	
3	ORF1ab	F	ACCTCATGGTCATGTTATGG	ACCTCATGGTCATGTTATGG	502-521	NIID Japan
		R	GACATAGCGAGTGTATGCC	GGCATACACTCGCTATGTC	805-823	
4	ORF1ab	F	ATGAGCTTAGTCCTGTTG	ATGAGCTTAGTCCTGTTG	12690-12707	NIID Japan
		R	CTCCCTTTGTTGTGTTGT	ACAACACAACAAAGGGAG	12780-12797	
		P	AGATGTCTTGTGCTGCCGGTA	AGATGTCTTGTGCTGCCGGTA	12717-12737	
5	ORF1ab	F	CCCTGTGGGTTTTACACTTAA	CCCTGTGGGTTTTACACTTAA	13342-13362	China CDC
		R	ACGATTGTGCATCAGCTGA	TCAGCTGATGCACAATCGT	13442-13460	
		P	CCGTCTGCGGTATGTGGAAAGGTTATGG	CCGTCTGCGGTATGTGGAAAGGTTATGG	13377-13404	
6	ORF1ab	F	GGTAACTGGTATGATTTCG	GGTAACTGGTATGATTTCG	14080-14098	IP France
		R	CTGGTCAAGGTTAATATAGG	CCTATATTAACCTTGACCAG	14167-14186	
		P	TCATACAAACCACGCCAGG	TCATACAAACCACGCCAGG	14105-14123	

Table I. Cont.

No.	Region		Primer or probe sequence	Target sequence	Location	Institute
7	ORF1ab	F	GTGARATGGTCATGTGTGGCGG	GTGAAATGGTCATGTGTGGCGG	15431-15452	Charité, Germany
		R	CARATGTTAAASACACTATTAGCATA	TATGCTAATAGTGTTTTTAACATTTG	15505-15530	
		P	CCAGGTGGWACRTCATCMGGTGATGC	CCAGGTGGAACCTCATCAGGAGATGC	15469-15494	
		P	CAGGTGGAACCTCATCAGGAGATGC	CAGGTGGAACCTCATCAGGAGATGC	15470-15494	
8	ORF1ab	F	GTGARATGGTCATGTGTGGCGG	GTGAAATGGTCATGTGTGGCGG	15431-15452	Charité, Germany
		R	CARATGTTAAASACACTATTAGCATA	TATGCTAATAGTGTTTTTAACATTTG	15505-15530	
		P	CAGGTGGAACCTCATCAGGAGATGC	CAGGTGGAACCTCATCAGGAGATGC	15470-15494	
9	ORF1ab	F	TGGGGYTTTACRGGTAACCT	TGGGGTTTTACAGGTAACCT	18778-18797	HKU
		R	AACRCGCTTAACAAAGCACTC	GAGTGCTTTGTTAAGCGTGTT	18889-18909	
		P	TAGTTGTGATGCWATCATGACTAG	TAGTTGTGATGCAATCATGACTAG	18849-18872	
10	S	F	TTGGCAAAATTCAAGACTCACTTT	TTGGCAAAATTCAAGACTCACTTT	24354-24377	NIID Japan
		R	TGTGGTTCATAAAAATTCCTTTGTG	CACAAAGGAATTTTTATGAACCACA	24876-24900	
11	S	F	TCAAGACTCACTTTCTTCCAC	TCAAGACTCACTTTCTTCCAC	24364-24384	NIID Japan
		R	ATTTGAAACAAAGACACCTTCAC	GTGAAGGTGTCTTTGTTTCAAAT	24834-24856	
12	S	F	AAGACTCACTTTCTTCCACAG	AAGACTCACTTTCTTCCACAG	24366-24386	NIID Japan
		R	CAAAGACACCTTCACGAGG	CCTCGTGAAGGTGTCTTTG	24830-24848	
13	E	F	ACAGGTACGTTAATAGTTAATAGCGT	ACAGGTACGTTAATAGTTAATAGCGT	26269-26294	Charité, Germany
		R	ATATTGCAGCAGTACGCACACA	TGTGTGCGTACTGCTGCAATAT	26360-26381	
		P	ACACTAGCCATCCTTACTGCGCTTCG	TACACTAGCCATCCTTACTGCGCTTCG	26331-26357	
14	E	F	ACAGGTACGTTAATAGTTAATAGCGT	ACAGGTACGTTAATAGTTAATAGCGT	26269-26294	IP France
		R	ATATTGCAGCAGTACGCACACA	TGTGTGCGTACTGCTGCAATAT	26360-26381	
		P	ACACTAGCCATCCTTACTGCGCTTCG	ACACTAGCCATCCTTACTGCGCTTCG	26332-26357	
15	N	F	GACCCCAAAATCAGCGAAAT	GACCCCAAAATCAGCGAAAT	28287-28306	US CDC
		R	TCTGGTTACTGCCAGTTGAATCTG	CAGATTCAACTGGCAGTAACCAGA	28335-28358	
		P	ACCCCGCATTACGTTTGGTGGACC	ACCCCGCATTACGTTTGGTGGACC	28309-28332	
16	N	F	CGTTTGGTGGACCCTCAGAT	CGTTTGGTGGACCCTCAGAT	28320-28339	NIH Thailand
		R	CCCCACTGCGTTCTCCATT	AATGGAGAACGCAGTGGGG	28358-28376	
		P	CAACTGGCAGTAACCA	CAACTGGCAGTAACCA	28341-28356	
17	N	F	GGGAGCCTTGAATACACCAAAA	GGGAGCCTTGAATACACCAAAA	28681-28702	US CDC
		R	TGTAGCACGATTGCAGCATTG	CAATGCTGCAATCGTGCTACA	28732-28752	
		P	AYCACATTGGCACCCGCAATCCTG	ATCACATTGGCACCCGCAATCCTG	28704-28727	
18	N	F	GGGGAACTTCTCCTGCTAGAAT	GGGGAACTTCTCCTGCTAGAAT	28881-28902	China CDC
		R	CAGACATTTTGCTCTCAAGCTG	CAGCTTGAGAGCAAAATGTCTG	28958-28979	
		P	TTGCTGCTGCTTGACAGATT	TTGCTGCTGCTTGACAGATT	28934-28953	
19	N	F	AAATTTTGGGGACCAGGAAC	AAATTTTGGGGACCAGGAAC	29125-29144	NIID Japan
		R	TGGCAGCTGTGTAGGTCAAC	GTTGACCTACACAGGTGCCA	29263-29282	
		P	ATGTCGCGCATTGGCATGGA	ATGTCGCGCATTGGCATGGA	29222-29241	
20	N	F	TAATCAGACAAGGAACTGATTA	TAATCAGACAAGGAACTGATTA	29145-29166	HKU
		R	CGAAGGTGTGACTTCCATG	CATGGAAGTCACACCTTCG	29236-29254	
		P	CAAATTGTGCAATTTGCGG	CAAATTGCACAATTTGCCC	29180-29198	
21	N	F	TTACAAACATTGGCCGCAAA	TTACAAACATTGGCCGCAAA	29164-29183	US CDC
		R	GCGCGACATTCCGAAGAA	TTCTTCGGAATGTCGCGC	29213-29230	
		P	ACAATTTGCCCCCAGCGCTTCAG	ACAATTTGCCCCCAGCGCTTCAG	29188-29210	

Variation sites are highlighted in gray. China CDC – the Chinese Center for Disease Control and Prevention, IP France – Institute Pasteur, Paris, France, US CDC – Centers for Disease Control and Prevention, the United States, NIID Japan – National Institute of Infectious Diseases, Japan, HKU – School of Public Health, Hong Kong University, NIH Thailand – the National Institute of Health, Thailand.

Table II. Polymerase chain reaction kit related variations of SARS-CoV-2

No.	Variation sites		Type	Isolates	No. of variation sites in this kit	No. of related isolates in this kit	Notes
1	F	3	SNP	37	6	61	
	R	3	SNP	24			
2	F	3	Deletion, SNP	5	8	22	
	R	5	SNP	17			
3	F	7	SNP, Deletion	93	8	94	
	R	1	SNP	1			
4	F	5	SNP	10	6	12	
	R	0	SNP	0			
	P	1	SNP	2			
5	F	1	SNP	1	5	55	
	R	1	SNP	1			
	P	3	SNP	53			
6	F	0	SNP	0	3	5	
	R	1	SNP	2			
	P	2	SNP	3			
7	F	2	SNP	2	6	8	
	R	0	SNP	0			
	P	2	SNP	3			
	P	2	SNP	3			
8	F	2	SNP	2	4	5	
	R	0	SNP	0			
	P	2	SNP	3			
9	F	1	SNP	11	6	20	
	R	4	SNP	8			
	P	1	SNP	1			
10	F	5	SNP	19	6	20	
	R	1	SNP	1			
11	F	10	SNP	31	11	33	Primer related to the most variation sites
	R	1	SNP	2			
12	F	9	SNP	30	10	32.00	
	R	1	SNP	2			
13	F	1	SNP	1	5	8	
	R	1	SNP	1			
	P	3	SNP	6			
14	F	1	SNP	1	5	8	
	R	1	SNP	1			
	P	3	SNP	6			
15	F	5	SNP	7	13	74	
	R	3	SNP	25			
	P	5	SNP	42			
16	F	3	SNP	10	10	32	
	R	6	SNP	8			
	P	1	SNP	14			
17	F	2	SNP	90	14	110	
	R	4	SNP	7			
	P	8	SNP	13			

Table II. Cont.

No.	Variation sites		Type	Isolates	No. of variation sites in this kit	No. of related isolates in this kit	Notes
18	F	8	SNP	1838	15	1847.00	Primer related to the most isolates
	R	7	SNP	9			Kits related to the most variation sites and isolates
	P	0	SNP	0			
19	F	2	SNP	14	11	25	
	R	3	SNP	3			
	P	6	SNP	8			
20	F	2	SNP	6	11	29	
	R	7	SNP	20			
	P	2	SNP	3			
21	F	3	SNP	7	6	13	
	R	0	SNP	0			
	P	3	SNP	6			

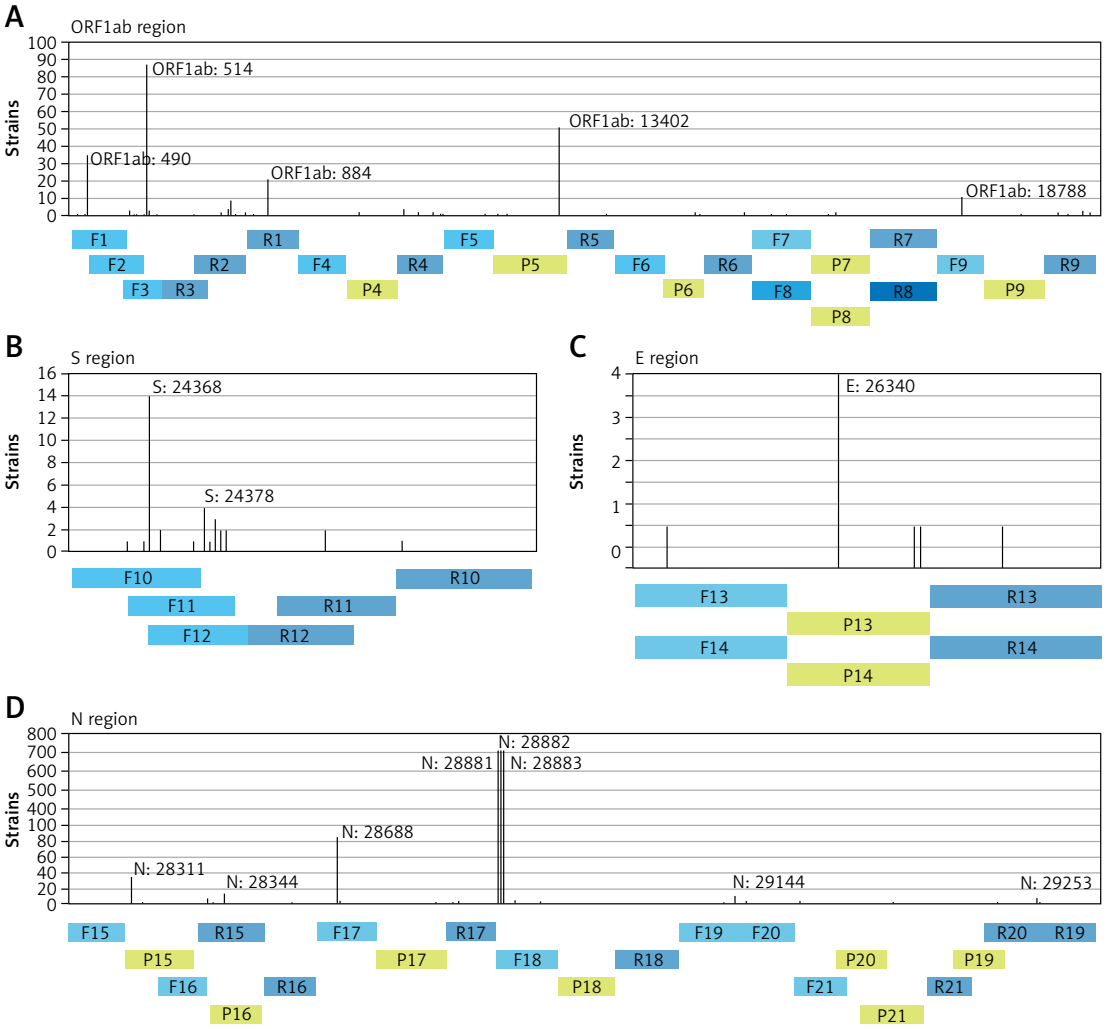

Figure 1. Variation landscape of primer and probe target sequences in SARS-CoV-2 strains. Variant loci and numbers of mutant strains in ORF1ab (**A**), S (**B**), E (**C**), and N region (**D**)

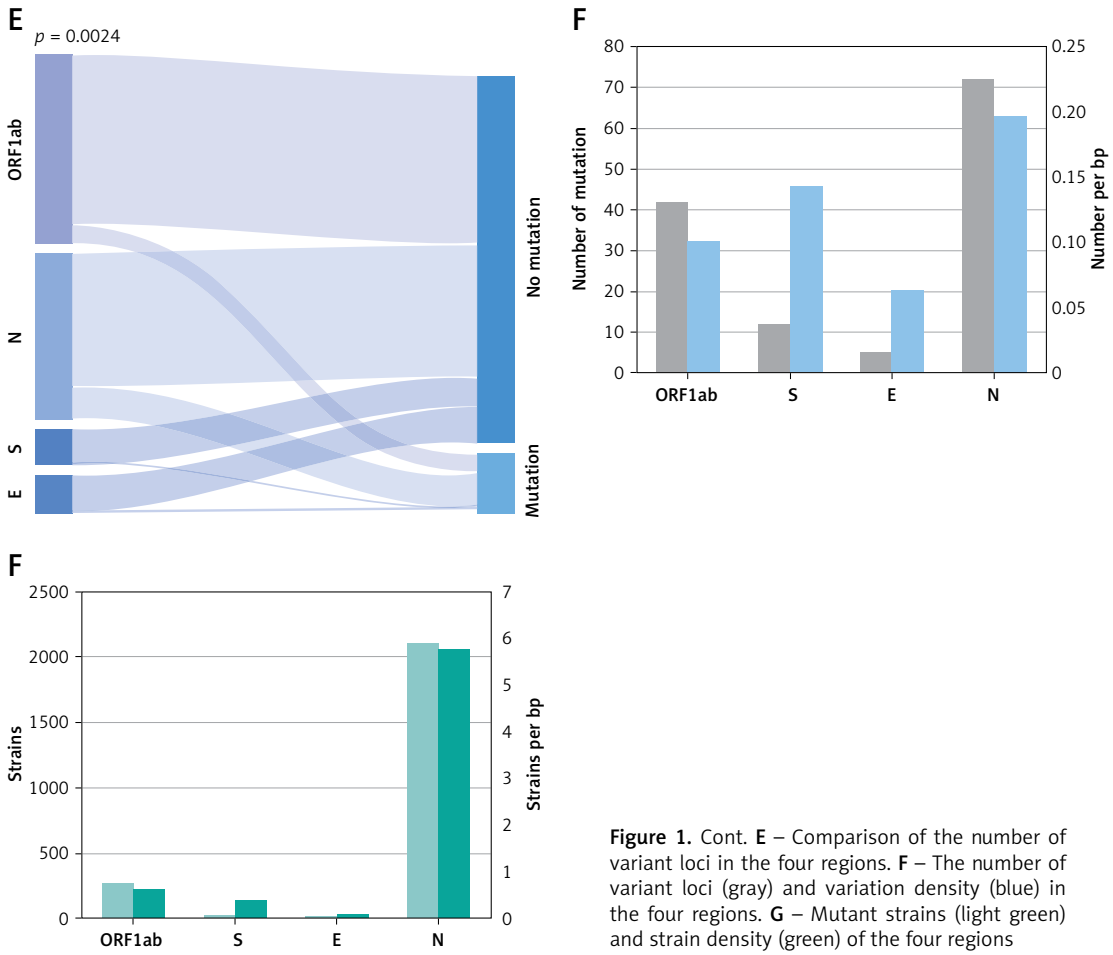

Figure 1. Cont. **E** – Comparison of the number of variant loci in the four regions. **F** – The number of variant loci (gray) and variation density (blue) in the four regions. **G** – Mutant strains (light green) and strain density (green) of the four regions

were only 84 bp in length. However, only 12 variant loci were identified with high variation density (0.14). The target sequence of the E region was the shortest (74 bp) and the least mutative, with the fewest variant loci (5 loci) and the lowest variation density (0.07, $p = 0.024$) (Figures 1 E, F).

Among these four regions, there were 2415 (isolate density = 2.57 per locus) virus strains with variant loci among primer or probe target sequences. In ORF1ab, 271 (isolate density = 0.65 per locus) isolates had variant loci involved in the primers or probes, and 34 isolates (isolate density = 0.40 per locus) in the S region. Significantly, the fewest isolates and lowest isolate density were demonstrated in the E region, with 8 strains and 0.11 isolates per locus. Conversely, the N region had the most isolates and the highest density, with 2102 isolates and 5.78 isolates per locus (Figure 1 G).

Variation analysis of SARS-CoV-2 strains in the target sequence of nucleic acid amplification tests

To further illustrate the variations, we investigated the specific variation types. Among the variant loci covered by paired primers and probes, the most common variations were single nucleotide polymorphisms (SNPs), accounting for 96% of all variations observed. Only 3% of the variations were deletions, and 1% of the variations were both SNP and deletion (Figure 2 A). Regarding the involved SNPs, 58% were transitions and 35% were transversions. A total of 37% of SNP variant loci were C to T transitions, the most common type of SNP, followed by G to T transversions and T to C transitions (Figure 2 B).

In the 6198 SARS-CoV-2 virus strains, the G to A transition involved the most isolates (1218) with the highest isolate density (121.8 per locus) in all types of SNPs, followed by G to C transversion and T to C transition (Figure 2 C). The C to T transition was the most common SNP type, accounting for 42% of the SNPs in the ORF1ab region, 38% in the S region, 60% in the E region, and 39% in the N region. The ORF1ab region contained 12 SNPs. There were eight SNP types in the S region. In addition to the C to T transition and A to G transition (15%), the other six types uniformly accounted for 7–8% separately. The region with the fewest SNP types was the E region, and there were only three types of SNPs. In the N region, there were 16 types

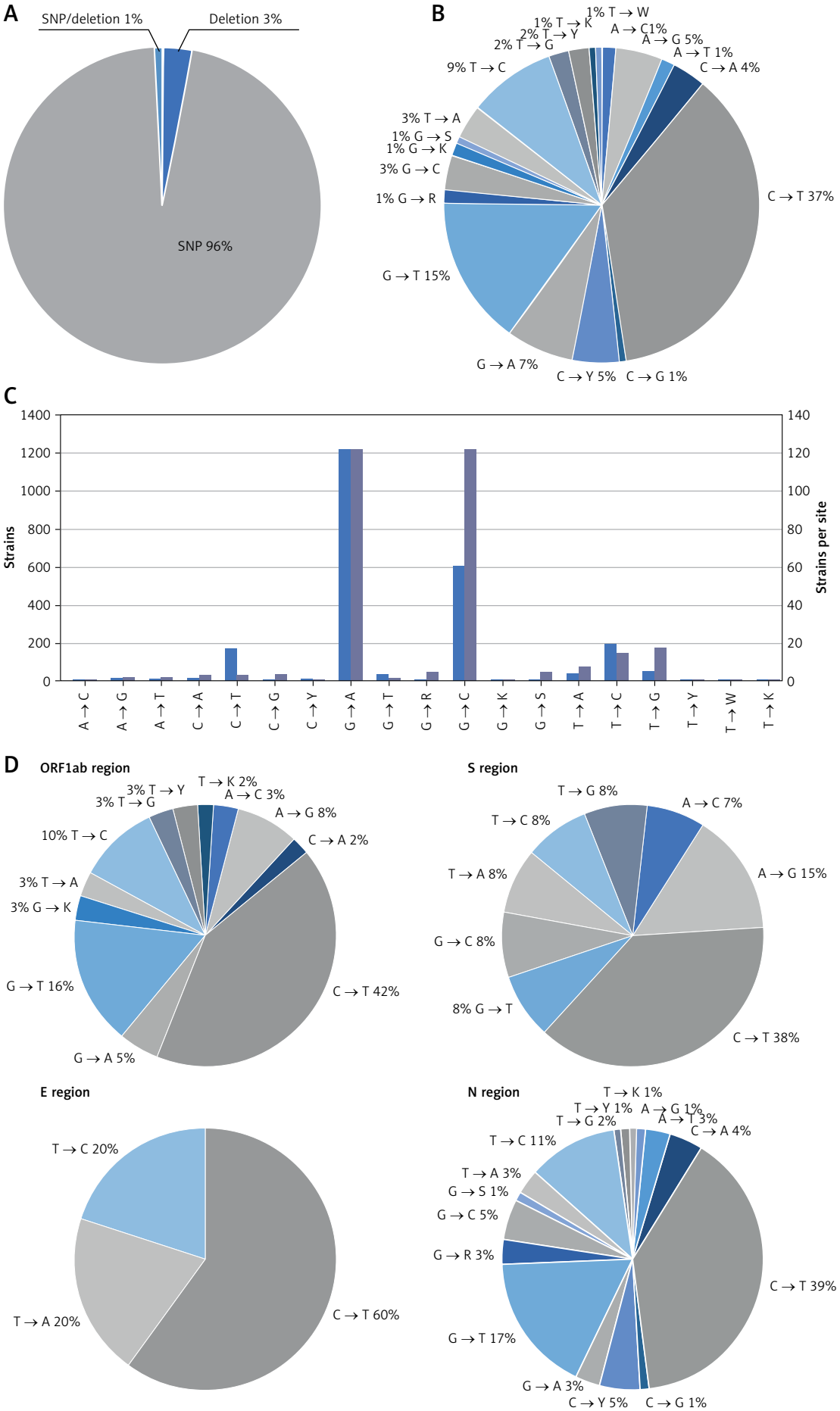

Figure 2. Variation analysis of SARS-CoV-2 strains in the target sequence of the nucleic acid amplification tests. **A** – Pie chart of variation types. **B** – Pie chart of SNP types. **C** – Strain number and strain density of each SNP type. **D** – Pie chart of the SNP types in ORF1ab, S, E, and N regions

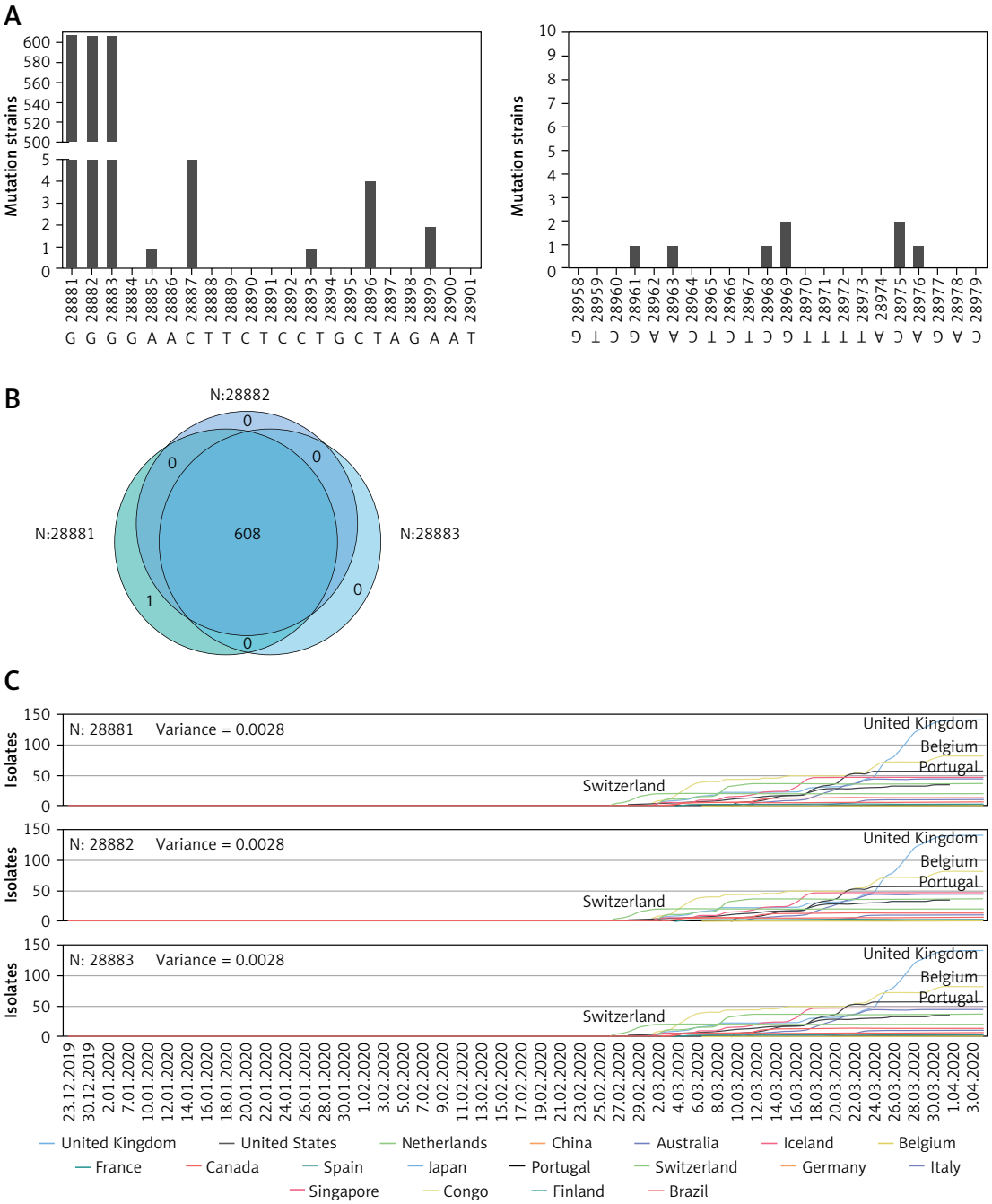

Figure 3. Variation analysis of N:28881, N: 28882, and N:28883 loci. **A** – Variation landscape of primers F-18 and R-18. **B** – Venn of the mutant strains containing variation in N:28881, N: 28882, or N:28883. **C** – Variation curve of N:28881, N: 28882, and N:28883 over time and by country

of SNPs, the most abundant. Besides the C to T transition, the most common type of SNPs were G to T transversion and T to C transition (Figure 2 D).

Variation analysis of N:28881, N: 28882, and N:28883 loci

To investigate the variations at the most common variant loci, we analyzed the occurrence of related primers, isolates, and variations over time and within different geographical regions. The variant loci involving the most isolates in the sequence targeted by primers or probes were N: 28881 (609), N: 28882 (608), and N:28883 (608). The primers covering these three variant loci were, F: GGGGAACTTCTCCTGCTAGAAT, and R: CAGACATTTTGCTCTCAAGCTG, recommended by China CDC. Besides 28881-28883, some variant loci had many corresponding mutant virus strains, such as 28887, 28896, and 28969 (Figure 3 A). The virus strains tended to carry 608 overlapping strains at 28881, 28882, and 28883. Only one

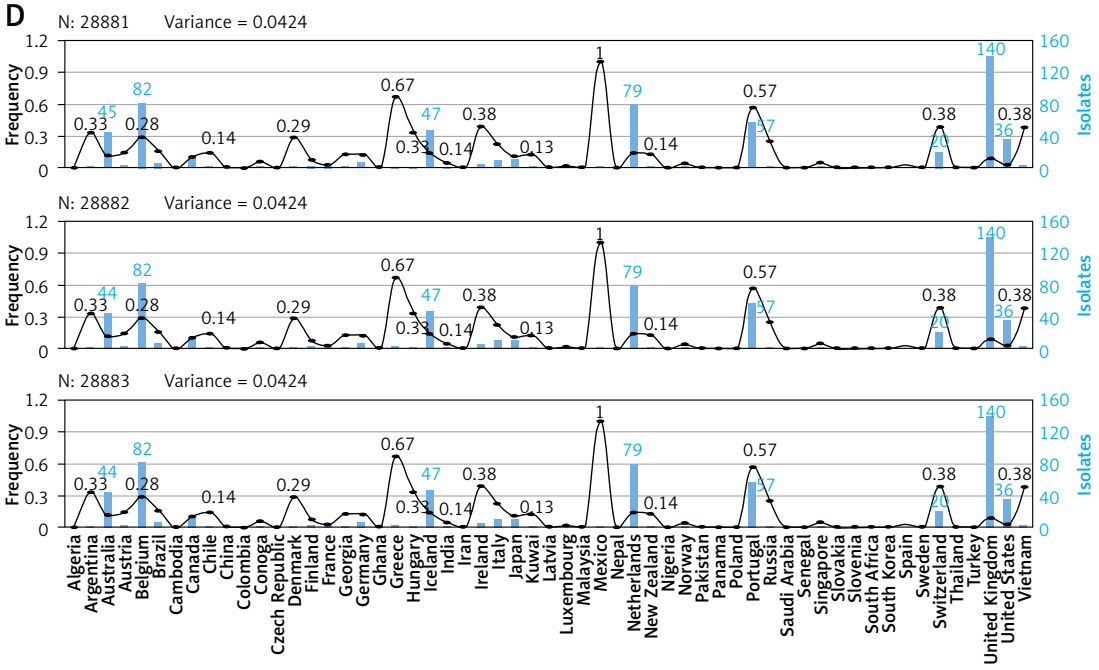

Figure 3. Cont. **D** – The number of isolates and the variation frequency of N:28881, N: 28882, and N:28883 by country. Regional variances are shown at the top; blue histogram indicates isolate number; black line chart indicates variation frequency

strain is uniquely displayed at 28881. It comes from Australia (Figure 3 B). The variant strains at these three loci showed the same mutation curve and mutation frequency in different countries (Figures 3 B, C), which indicates that except for this strain from Australia, the linkage pattern of mutation pairs by country is almost the same. At the end of February, major mutant isolates were obtained from Switzerland. At the end of March and the beginning of April, the United Kingdom (UK) provided a large portion of the mutant isolates, followed by Belgium and Portugal. The time variances of the three sites were the same (Figure 3 C). In detail, the unique strain, at 28881, came from Australia, with 45 isolates at this site. In other countries, except Australia, the trend was the same at 28881, 28882, and 28883. Most isolates came from the UK, with 140 mutant isolates at these three sites separately, followed by Belgium, Netherlands, Portugal, and Iceland. The variation frequencies of the three sites were also the same, showing different geographical preferences with isolate distribution. The highest variation frequency was from Mexico (1.00), whose mutant isolates were extremely limited. Greece and Portugal also had high variation frequencies at the three sites. The regional variances of the three sites equaled 0.0424 (Figure 3 D).

Isolate analysis of common variant loci

Geographical distributions of several other common variant loci were also demonstrated, such as ORF1ab:490, ORF1ab:514, ORF1ab:13402, N:28311 and N:28688 (Figure 4). At ORF1ab:490, the most mutant isolates were from Australia (21 isolates) and the United States (11 isolates). The highest variation frequencies at ORF1ab:490 were in New Zealand and Georgia (0.125, regional variance = 0.0007) (Figure 4 A). The forward primer of the first pair of primers and probes from Japan National Institute of Infectious Diseases targeted this variant locus (Figure 1 A and Table I). At ORF1ab:514, most mutant isolates were from the Netherlands (66 isolates, 75.86%) with a high variation frequency (0.12). The highest variation frequencies at ORF1ab:514 were observed in Slovakia (Figure 4 B). This variant locus was targeted by the forward primer of the No. 3 pair of primers and probes from Japan National Institute of Infectious Diseases (Figure 1 A and Table I). Prominently, at ORF1ab:13402, 96.08% (49 isolates) of mutant isolates were from Belgium, with the second highest variation frequency. The highest variation frequency at ORF1ab:13402 was from Denmark (0.29, regional variance = 0.0020) (Figure 4 C). This variant locus was targeted by the No. 5 pair of primers and probes provided by China CDC (Figure 1 D and Table I). The most isolates at N:28311 were from Australia (20, 57.14%). However, the variation frequency of Australia at N:28311 was not high. The highest variation frequencies at N:28311 occurred in Saudi Arabia, South Korea, and Malaysia (regional variance = 0.0053) (Figure 4 D). The probe of the No. 15 pair of primers and probes provided by US CDC targeted N:28311

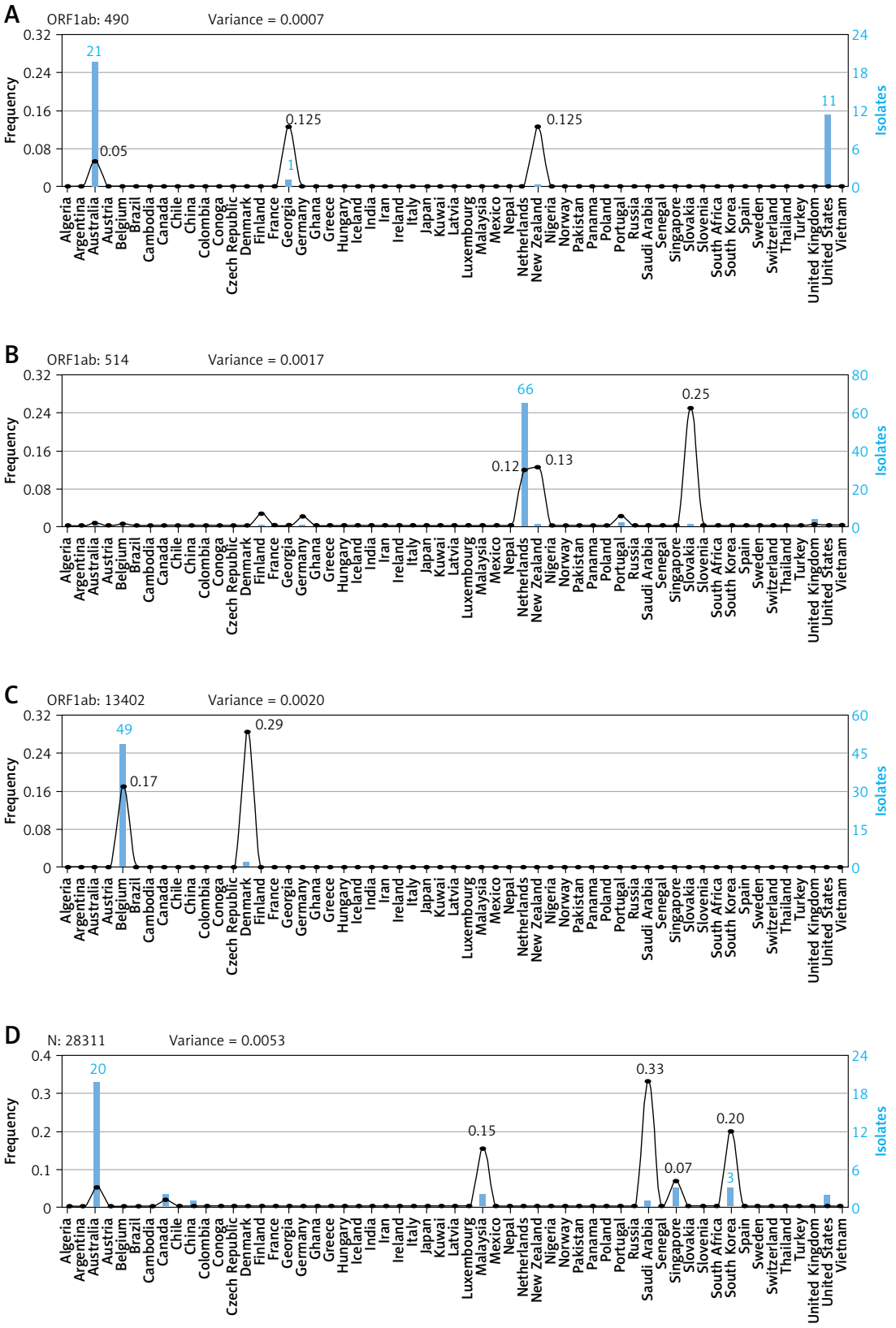

Figure 4. Isolates analysis of common variant loci. The number of isolates and the variation frequency of ORF1ab: 490 (**A**), ORF1ab: 514 (**B**), ORF1ab: 13402 (**C**), N: 28311 (**D**), and N: 28688 (**E**). Regional variances are shown at the top; blue histogram indicates isolate number; black line chart indicates variation

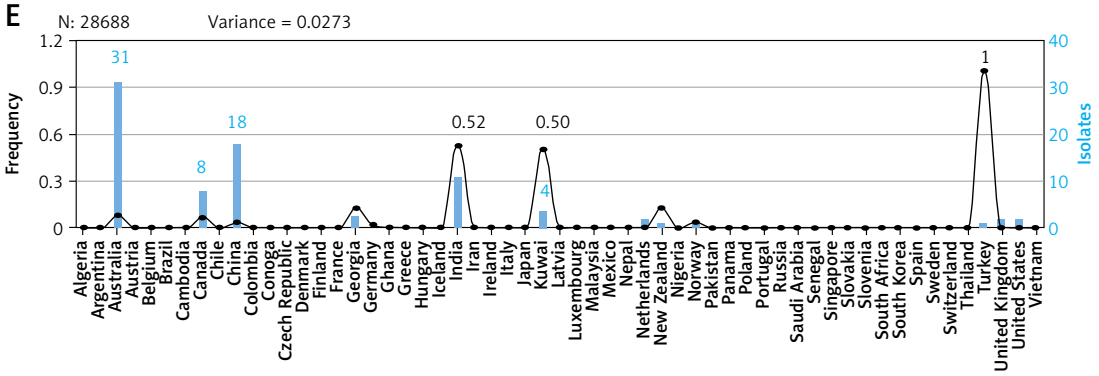

Figure 4. Cont. N: 28688 (**E**). Regional variances are shown at the top; blue histogram indicates isolate number; black line chart indicates variation

(Figure 1 D and Table I). The distribution of mutant isolates at N:28688 was decentralized. There were 31 (36.47%) isolates from Australia, 18 (21.18%) from China, and 11 (12.94%) from India. High variation frequencies at N:28688 were observed in Turkey, India, and Kuwait (regional variance = 0.0273) (Figure 4 E). N:28688 was targeted by the forward primer of the No. 17 pair of primers and probes, which was recommended by US CDC (Figure 1 D and Table I, Figure 5).

Discussion

In the present study, our fundamental purpose is to reveal the variations in the target sequence of primers and probes in COVID-19 detection kits used on a large scale in the population. We compared the primer and probe sequences used to detect SARS-CoV-2 infection to those exported from the 2019nCoVR database to gain important insights into virus variations in primer and probe target regions.

By integrating 21 pairs of primer sequences in regions ORF1ab, S, E, and N from 6198 SARS-CoV-2 strains, we found 31 (13.97%) variant loci, involving 2415 (38.96%) isolates, especially in the N region, which had the highest variation density and the highest isolate density. The most variant loci and the most isolates were also located in the N region (Figures 1 F, G). The high variation frequency and included isolate density implied that the N region might not be suitable for primer design, which might elevate the false-negative rate. Therefore, merely nucleic acid testing alone is not enough to confirm a negative case; it should be combined with other tests such as antibody detection [10]. In contrast, the lowest variation density and the lowest isolate density were observed in the E region. Additionally, it had the fewest variant loci and the fewest isolates. Thus, the E region might be the most conserved region, suitable for primer design [22]. This is consistent with the findings of Udugama *et al.* (2020) and Corman *et al.* (2020, Jan) that the RdRP and E genes had high analytical sensitivity for detection, whereas the N gene provided poorer analytical sensitivity [10, 20, 29]. However, only two identical pairs of primers and probes were from the E region, which might cause an incomplete understanding of this region in the present study.

The variations involved in the primers and probes were mainly SNPs, which were abundant in the SARS-CoV-2 genome [30–32]. Mutations and sequence mismatch affected the PCR output, and mismatches between diagnostic PCR assays and the coronavirus SARS-CoV-2 genome were present [33–37]. N:28881, N:28882, and N:28883 were the most influential variant loci, involving 608 or 609 viral isolates. It might be better to avoid designing primers or probes whose target sequences contain these three sites. As only the No. 18 primer pair targeted this sequence in the 21 pairs, verification with different nucleic acid detection kits could elevate sensitivity [38]. As viruses evolve during outbreaks, SNPs in primer or probe binding regions could alter the sensitivity of PCR assays [24]. We analyzed the variations in the target sequence of primers and probes in COVID-19 detection kits used on a large scale in the population and their geographic distribution, thereby providing a reference for optimizing the selection of detection kits used on a large scale in the population, improving the diagnostic accuracy of the tests. Sapoval *et al.* calculated that each probe/primer sequence that was available on the WHO website contained 2.529 iSNV and/or 2.477 SNPs, suggesting the potential for a drop in the sensitivity of the affected probes and primers [39].

The 17th pair of primers and probes used to be recommended by the US CDC, involving 110 isolates, the second most among 21 pairs of primers and probes. N: 28688 is the most common mutation site among the above sequences, and at this site, the United States has the third most isolates since May 12th, 2020. This pair of primers and probes was removed from the US CDC recommended protocol updated in July [40]. At other common muta-

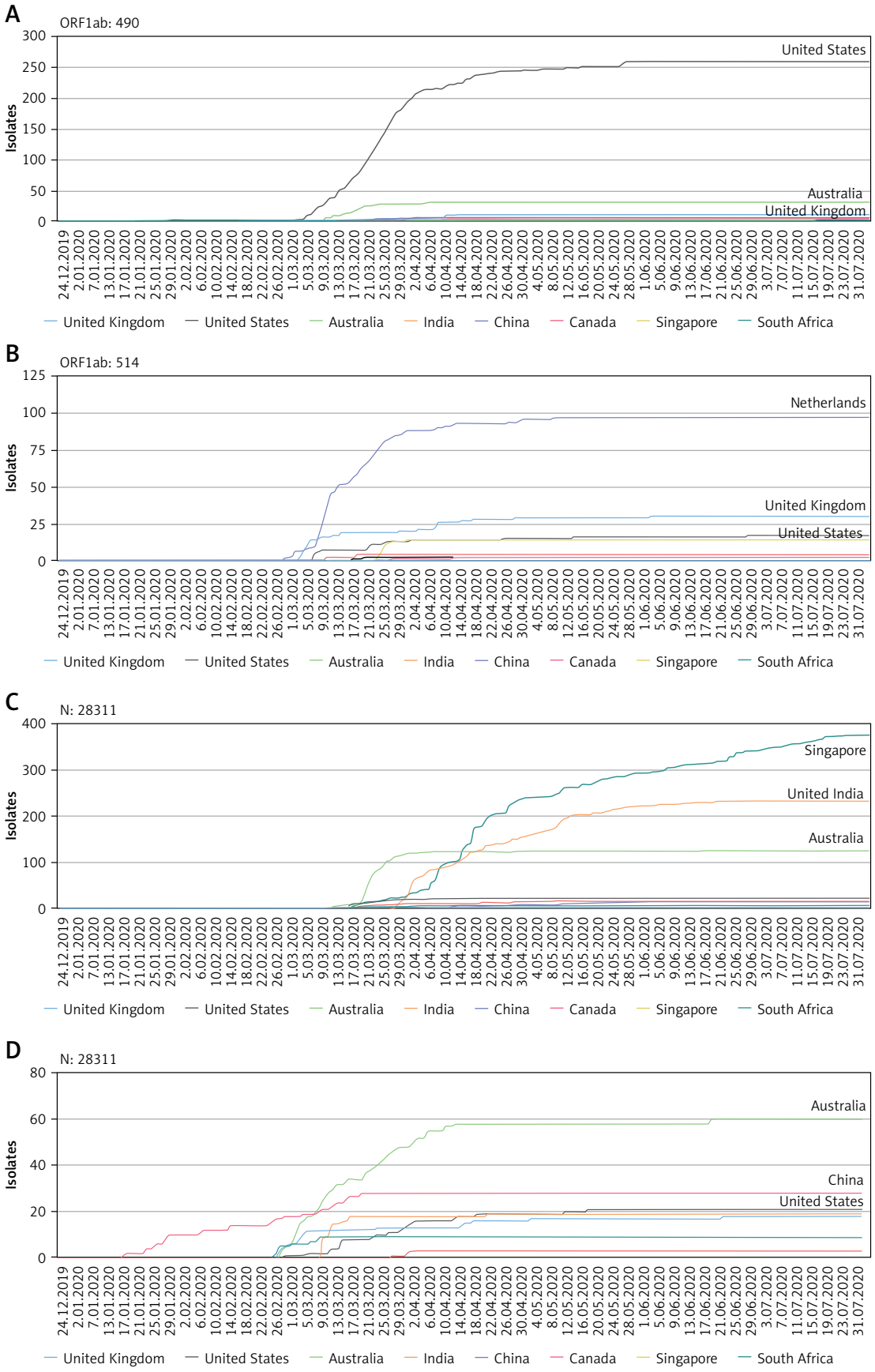

Figure 5. Isolates distribution of common variant loci. The number of isolates and the statistic date of ORF1ab: 490 (**A**), ORF1ab: 514 (**B**), N: 28311 (**C**), and N: 28688 (**D**). The countries and regions with the most isolates are marked

tion sites such as N:28881- N:28883, ORF1ab:490, ORF1ab: 514, ORF1ab: 13402 and N: 28311, the region where mutant strains often appear does not overlap with the region where this kit is used (Figure 5). In the selection and application of detection kits in the future, it would be better not to use the 18th pair of primers and probes provided by China CDC in the United Kingdom, Belgium, the Netherlands and Portugal (Figure 3). Japan National Institute of Infectious Diseases recommended the No. 1 and No. 3 pairs of primers and probes. The first pair should be avoided in the US and the third one should be avoided in the Netherlands, the UK and the US (Figures 4 A, 4 B, 5 A, 5 B). It would be better not to use the 5th pair in Belgium (Figure 4 C). Singapore, India and Australia ought to avoid using the 15th pair of primers and probes (Figures 4 D, 5 C).

There were some limitations to this study. Initially, not all kits used worldwide were available from the WHO website. Although the virus strain data are updated frequently, the strains we used merely presented the situation before our research ended (April 3rd, 2020). Moreover, it has been reported that variations in primer binding sites could affect PCR efficiency [41, 42]. Thus, we should further validate the relationship between variations and PCR efficiency in the future.

In conclusion, by integrating 21 pairs of primer target sequences in regions ORF1ab, S, E, and N from 6198 SARS-CoV-2 strains, the total length of primer and probe target sequences was 938 bp, with 131 (13.97%) variant loci, involving 2415 (38.96%) isolates. The E region had the highest priority in target sequence selection, and the N region the lowest priority. SNP was the most frequent variation with C to T transitions involving the most variant loci. Genomic variations at the three mutation sites N: 28881, N: 28882, and N: 28883 were the most commonly detected variations, including 608 SARS-CoV-2 strains from 33 countries, especially in the UK, Portugal, and Belgium. Herein, we comprehensively analyzed the genomic variations at the target sequences of the nucleic acid amplification tests, offering evidence for the optimization of primer and probe target sequence selection, providing implications for the identification of infected patients via the tests.

Acknowledgments

The authors thank all health-care workers and people involved in fighting against COVID-19. We would like to acknowledge the platform 2019nCoVR provided by the Chinese Academy of Sciences. We thank the seven institutes for the disclosure of nucleic acid test kit contents. We appreciate the WHO for providing SARS-CoV-2 relevant data.

Canhui Cao and Ruidi Yu contributed equally. Fei Ye, Wei Zhang, and Qinglei Gao contributed equally.

Conflict of interest

The authors declare no conflict of interest.

References

1. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020].
2. Hui DS, Azhar EI, Madani TA, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health – The latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis 2020; 91: 264-6.
3. Azeez SA, Alhashim ZG, Al Otaibi WM, et al. State-of-the-art tools to identify druggable protein ligand of SARS-CoV-2. Arch Med Sci 2020; 16: 497-507.
4. Borgio JF, Alsuwat HS, Al Otaibi WM, et al. State-of-the-art tools unveil potent drug targets amongst clinically approved drugs to inhibit helicase in SARS-CoV-2. Arch Med Sci 2020; 16: 508-18.
5. Nabavi S, Habtemariam S, Clementi E, et al. Lessons learned from SARS-CoV and MERS-CoV: FDA-approved Abelson tyrosine-protein kinase 2 inhibitors may help us combat SARS-CoV-2. Arch Med Sci 2020; 16: 519-21.
6. New Coronavirus Pneumonia Diagnosis and Treatment Guidelines (Trial Version 7) [http://www.nhc.gov.cn/yzygj/s7653p/202003/46c9294a7dfe4cef80dc7f5912eb1989.shtml]
7. Cheng MP, Papenburg J, Desjardins M, et al. Diagnostic testing for severe acute respiratory syndrome-related coronavirus-2. Ann Intern Med 2020; 172: 726-34.
8. Liu C, Huang Q, Wang P, et al. COVID-19 disease: novel clinical manifestations and therapeutic exploration. Arch Med Sci 2020; doi:10.5114/aoms.2020.98401.
9. Bordi L, Nicastri E, Scorzolini L, et al. Differential diagnosis of illness in patients under investigation for the novel coronavirus (SARS-CoV-2), Italy, February 2020. Euro Surveill 2020; 25: 2000170.
10. Udugama B, Kadhiresan P, Kozlowski HN, et al. Diagnosing COVID-19: the disease and tools for detection. ACS Nano 2020; 14: 3822-35.
11. Guan WJ, Liang WH, Zhao Y, et al. Comorbidity and its impact on 1590 patients with Covid-19 in China: a nationwide analysis. Eur Respir J 2020; 55: 2000547.
12. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 2020; 382: 1708-20.
13. Beeching NJ, Fletcher TE, Beadsworth MBJ. Covid-19: testing times. BMJ 2020; 369: m1403.
14. Xie X, Zhong Z. Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. Radiology 2020; 296: E41-5.
15. Li Y, Yao L, Li J, et al. Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. J Med Virol 2020; 92: 903-8.
16. Wenling W, Yanli X, Ruqin G, et al. Detection of SARS-CoV-2 in different types of clinical specimens. JAMA 2020; 323: 1843-4.
17. Pan Y, Long L, Zhang D, et al. Potential false-negative nucleic acid testing results for severe acute respirato-

ry syndrome coronavirus 2 from thermal inactivation of samples with low viral loads. Clin Chem 2020; 66: 794-801.
18. Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. N Engl J Med 2020; 382: 1177-9.
19. Chu DKW, Pan Y, Cheng SMS, et al. Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia. Clin Chem 2020; 66: 549-55.
20. Corman VB, Brünink S, Zambon M. Diagnostic detection of Wuhan coronavirus 2019 by real-time RT-PCR; 2020; Geneva. World Health Organization.
21. Freeman WM, Walker SJ, Vrana KE. Quantitative RT-PCR: Pitfalls and Potential. BioTechniques 1999; 26: 112-25.
22. Dieffenbach CW, Lowe TM, Dveksler GS. General concepts for PCR primer design. PCR Methods Appl 1993; 3: S30-7.
23. Nalla AK, Casto AM, Huang MLW, et al. Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit. J Clin Microbiol 2020; 58: e00557-20.
24. Vogels CBF, Brito AF, Wyllie AL, et al. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT-qPCR primer-probe sets. Nat Microbiol 2020; 5: 1299-305.
25. Zhao WM, Song SH, Chen ML, et al. The 2019 novel coronavirus resource. Yi Chuan 2020; 42: 212-21.
26. 2019 nCoVR [https://bigd.big.ac.cn/ncov].
27. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9: 357-9.
28. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. Genome Biol 2016; 17: 122.
29. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill 2020; 25: 2000045.
30. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics 2020; 112: 3588-96.
31. Chang TJ, Yang DM, Wang ML, et al. Genomic analysis and comparative multiple sequence of SARS-CoV2. J Chin Med Assoc 2020; 83: 537-43.
32. Yao H, Lu X, Chen Q, et al. Patient-derived mutations impact pathogenicity of SARS-CoV-2. medRxiv 2020.04.14.20060160.
33. Khan KA, Cheung P. Presence of mismatches between diagnostic PCR assays and coronavirus SARS-CoV-2 genome. R Soc Open Sci 2020; 7: 200636.
34. Lefever S, Pattyn F, Hellemans J, Vandesompele J, et al. Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays. Clin Chem 2013; 59: 1470-80.
35. Rejali NA, Moric E, Wittwer CT, et al. The effect of single mismatches on primer extension. Clin Chem 2018; 64: 801-9.
36. Stadhouders R, Pas SD, Anber J, et al. The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. J Mol Diagn 2010; 12: 109-17.
37. Whiley DM, Sloots TP. Sequence variation in primer targets affects the accuracy of viral quantitative PCR. J Clin Virol 2005; 34: 104-7.
38. Liu X, Feng J, Zhang Q, et al. Analytical comparisons of SARS-COV-2 detection by qRT-PCR and ddPCR with multiple primer/probe sets. Emerg Microbes Infect 2020; 9: 1175-9.
39. Sapoval N, Mahmoud M, Jochum MD, et al. Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission. bioRxiv 2020; 2020.07.02.184481.
40. CDC's Diagnostic Test for COVID-19 Only and Supplies [https://www.cdc.gov/coronavirus/2019-ncov/lab/virus-requests.html].
41. Kalle E, Kubista M, Rensing C. Multi-template polymerase chain reaction. Biomol Detect Quantif 2014; 2: 11-29.
42. Kwok S, Kellogg DE, McKinney N, et al. Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. Nucleic Acids Res 1990; 18: 999-1005.