Genomic characterization and phylogenetic analysis of SARS-CoV-2 during the early phase of the pandemic in Asia

Jale Moradi[1*], Mohsen Moghoofei[1*], Mohammad Doroudian[2], Ramin Abiri[1]

[1]Department of Microbiology, Faculty of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran

[2]Department of Cell and Molecular Biology, Faculty of Biological Sciences, Karazmi University, Tehran, Iran.

*Corresponding Authors:

Jale Moradi, Ph.D. Department of Microbiology, Faculty of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran. Email: Jale.moradi@kums.ac.ir. Mohsen Moghofei, Ph.D. Department of Microbiology, Faculty of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran. Email. Mohsen.moghoofei@kums.ac.ir

Running title: Phylogenetic analysis of SARS-CoV-2 in Asia

Abstract

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) as the current coronavirus pandemic is an infectious disease that was initially confirmed in China (December 2019). In the current study, we assessed the genome variation of the SARS-CoV-2 viruses circulated in Asia in the first months of the pandemic. We randomly analyzed 131 complete sequences of SARS-CoV-2 from December 2019 to April 2020. The results showed that there were fifteen major mutations in Asia which most of them were co-evolved. These prevalent co-mutations resulted in clade G, GH, GR, S and O. Furthermore, sequences within 26144G>T point mutation had low variability without any co-mutation which formed clade V. Our results indicate that most of the circulated viruses in Asia in the early time of the pandemic had collected in five co-mutation groups.

Keywords: SARS-CoV-2, Phylogenetic analysis, Asia, Variation, Mutation

Introduction

Coronaviruses are enveloped positive-strand RNA viruses which belong to the *Orthocoronavirinae* subfamily and *Coronaviridae* family. *Orthocoronavirinae* has itself four genera including *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* and *Gammacoronavirus* [1]. Seven Coronavirus species have caused infection in humans, although four of them including HCoV-OC43, HCoV-229E, HCoV-NL63 and HCoV-HKU1 are not clinically important in immunocompetent individuals [2]. Initially, *Betacoronavirus* (HCoV-OC43) and *Alphacoronavirus* (HCoV-229E) were identified. These are cause of common colds and considered of modest clinical importance [3, 4]. *Alphacoronavirus* (HcoV-NL63) and *Betacoronavirus* (HcoV-HKU1) cause bronchiolitis in children and community-acquired pneumonia, respectively [5, 6]. There are three additional Coronaviruses which are highly pathogenic and have caused epidemics in human populations, including (i) *severe acute respiratory syndrome coronavirus* (SARS-CoV) (*Betacoronavirus*, subgenus *Sarbecovirus)*, identified in China in 2002 and spread to the 29 countries with the mortality rate of ~10% and abruptly ended in 2003, (ii) *Middle East respiratory syndrome coronavirus* (MERS-CoV) (*Betacoronavirus*, subgenus *Merbecovirus)*, emerged in Saudi Arabia in 2012 with ~34% mortality rate which have detected in 27 countries and (iii) *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) (*Betacoronavirus*, subgenus *Sarbecovirus)* is a novel coronavirus that initially detected in Wuhan city (China) in late 2019 [7-11]. The World health organization (WHO) reported that 223 countries, areas or territories have been confirmed SARS-CoV-2 infection with 89707115 cases and 1940352 deaths as of January 13, 2021.

With regards to Coronaviruses genome, The 5' end encodes a poly-protein which cleaved to 16 non-structural proteins (Nsp1 to Nsp16). The 3' end provides critical proteins for the virus such as envelope glycoproteins spike (S), membrane (M), envelope (E), and nucleocapsid (N) [12].

Genotyping analysis is an essential tool to determine the mutations in the SARS-CoV-2 genome. Any variations in the vaccine candidate proteins (e.g. structural proteins) need to be analyzed prior to vaccine design. Moreover, genotyping data can be used to predict the efficacy of global vaccines in countries with high rates of various mutations [13-15]. In this study, we analyzed the complete sequences of SARS-CoV-2 to reveal genetic distance and mutation rate among Asian countries.

Materials and Methods

The submitted complete sequences of SARS-CoV-2 from Asian countries and reference sequences were obtained from GISAID and GeneBank, respectively. Alignment was performed with MAFFT (v7.455) [16]. The alignment results were visualized and trimmed for quality and length compatibility using Unipro UGENE software [17]. Phylogenetic tree constructed with maximum likelihood method using RAxML-NG v. 0.9.0 [18]. Transfer bootstrap expectation (TBE) with 1000 replicates used for branch support [19]. ETE3, the python framework was used to visualize and analyze the resulting tree [20]. Nucleotide substitutions were retrieved from the output of alignment and annotated to find protein changes.

Results

Data collection, preparation and alignment

Complete sequences of SARS-CoV-2 in GISAID were released 4572 to April 7, 2020. Among the collected samples, 641 sequences belonged to Asia that 604 sequences were categorized based on distinct regions. We analyzed 604 Asian complete sequences to remove low quality sequences with unknown base N content and selected sequences which were related to the patients except the regions with the lack of human samples. Totally, 131 sequences were selected in 12 Asian countries which all of them were related to the patients except for Vietnam sequences which belonged to the cell culture (Vero cells) (***Table 1***). The sequences with their related GISAID accession numbers is shown in Supplementary ***Table 1***. The reference sequence for SARS-CoV-2 was obtained from GenBank with the accession number of NC_045512.2 which was the source for sequence retrieval and analysis. All sequences were trimmed to 29698 bp. Five hundred and thirty-eight variations were identified which were related to the 197 variable sites (Supplementary ***Table 2***).

Table 1: Frequency of SARS-Cov-2 Sequences from Asian countriesinvolved in this study

Country	Total	Select	Abbr**
Georgia	13	12	GE
Hong Kong	63	3	HK
India	16	14	IN
Japan	102	24	JA
Kuwait	4	2	KU
Malaysia	7	1	MA
Nepal	1	1	NE
Pakistan	2	1	PA
Saudi Arabia	3	3	SA
Singapore	37	6	SI
South Korea	13	11	SK
Taiwan	22	8	TA
Thailand	2	2	TH
Vietnam*	8	5	VI
Cambodia	1	1	CM
China	310	37	CN
	604	131	

China Regions	Total	Select	Abbr**
Anhui	2	1	CNAn
Beijing	5	3	CNBe
Chongqing	3	0	
Fujian	2	2	CNFu
Guangdong	74	3	CNGu
Guangxi	6	0	
Guangzhou	1	0	
Hangzhou	36	5	CNHa
Henan	1	0	
Hubei	32	4	CNHu
Jiangsu	4	3	CNJs
Jiangxi	26	5	CNJx
NanChang	1	1	CNNa
Shandong	9	1	CNSd
Shanghai	96	7	CNSh
Sichuan	1	0	
Yunnan	2	0	
Zhejiang	9	2	CNZh
	310	37	

* All sequences related to the Vero cells
** Abbreviation

Genome variation and annotation

The variations were distributed in the SARS-CoV-2 genomes, except the ORF6 region. Almost half of the variations were associated with ORF1ab polyprotein (52%) and minimum frequencies were related to the ORF7b and E protein sequences (0.5%). In relation to the type of the mutations, the frequencies of missense, synonymous, nonsense, and non-coding SNPs were 62%, 33%, 0.5%, and 4%, respectively (***Figure 1***). We analyzed genome variation sites based on Asian countries and we found most of the variations were correlated to the Chinese sequences. Cambodia, Nepal, Thailand, and Malaysia had less than 2% of variations in their sequences (***Figure 2***). Fifteen out of 197 variable sites were the most prevalent variations in the sequences from Asia (***Table 2***). Mutations of the *ORF1ab* gene were related to three amino acid changes in this polyprotein. The most distributed variations among Asian countries (≥ 5 countries) were 241C>T, 1397G>A, 3037C>T, 8782C>T, 11083G>T, 14408C>T, 23403A>G, 26144G> and 28144G>T (***Figure 3***). The other most common variations were distributed in four countries. Further analysis revealed the occurrences of co-mutations in Asian sequences and showed five types of co-mutations in these genomes (***Table 3***). The most prevalent co-mutations were seen in 1397G>A, 28688T>C, 29742G>T and 11083G>T variation sites.

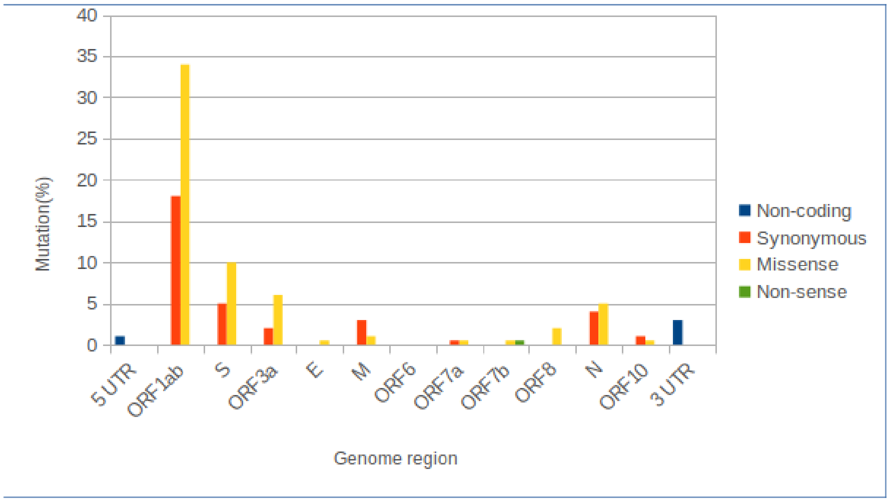

Figure 1: Genome distribution of SARS-CoV-2 mutations

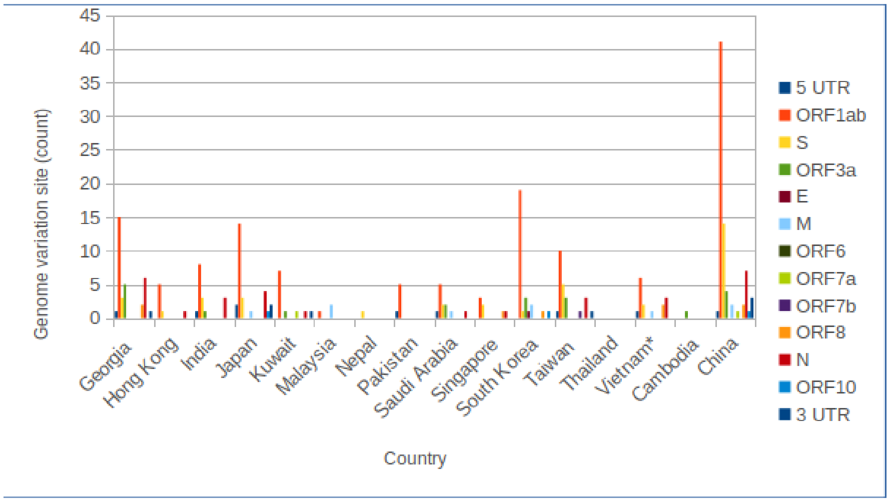

Figure 2: Frequency of SARS-Cov-2 variation sitesin Asia

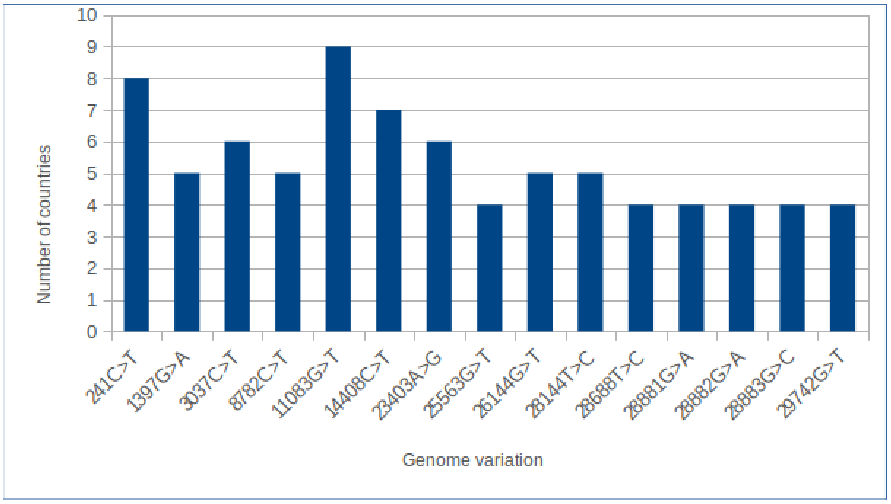

Figure 3: Distribution of most common variations

Table 2: High frequency mutations in SARS-CoV-2 sequences of Asia.

Genome Change	Gene	Protein Change	Mutation type	Clade
241C>T	5´UTR		Non-coding	
1397G>A	ORF1ab	V378I	Missense	
3037C>T			Synonymous	
8782C>T			Synonymous	
11083G>T		L3606F	Missense	
14408C>T		P4715L	Missense	
23403A>G	S	D614G	Missense	Clade G
25563G>T	ORF3a	Q57H	Missense	
26144G>T		G251V	Missense	Clade V
28144T>C	ORF8	L84S	Missense	Clade S
28688T>C	N		Synonymous	
28881G>A		R203K	Missense	
28882G>A			Synonymous	
28883G>C		G204R	Missense	
29742G>T	3´UTR		Non-coding	

Table 3: Co-mutations in SARS-Cov-2 sequences from Asia.

Co-occurring Mutations	Clade*
241C>T, 3037C>T, 14408C>T, 23403A>G	G
28881G>A, 28882G>A, 28883G>C, 23403A>G	GR
25563G>T, 23403A>G	GH
8782C>T, 28144T>C	S
1397G>A, 28688T>C, 29742G>T, 11083G>T	O

Phylogenetic analysis

Phylogenetic analysis of complete sequences of SARS-CoV-2 from Asia showed that there were four clades in this region. Three of these clades were included in the globally major clades which introduced by GISAID, named clade G (D614G variant of S glycoprotein), clade V (G251V variant of ORF3a protein product), and clade S (L84S variant of ORF8 protein product) (Figure 4). In this study, Clade S with 22 sequences was located at the base of the tree which is the most related to the reference sequences and initially collected Chinese samples. Sequences from China (11 out of 37), South Korea (8 out of 11), Vietnam (1 out of 5), India (1 out of 14) and Georgia (1 out of 12) were collected in this clade, which is common in the USA and China. These sequences were collected in January and February 2020, indicating that the early frequent transmission of the virus from China to South Korea. In the current study, clade G was consisting 34 sequences from India (13 out 14), Georgia (7 out of 12), Saudi Arabia (2 out 3), Taiwan (4 out of 8), Vietnam (3 out of 5) and Japan (5 out of 24) in which all the sequences were collected in March and April 2020. These results showed that transmission to India has occurred in this time. Clade V consists of 9 sequences from South Korea (2 out of 11), Taiwan (1 out of 8), Singapore (4 out of 6) and Hong Kong (2 out of 3). The clade O included the sequences which carried 11083G>T mutation. This mutation was co-evolved with 1397G>A, 28688T>C, and 29742G>T. This clade was consisting 33 sequences from Japan (17 out of 24), Georgia (4 out of 12), China (5 out of 37), Taiwan (2 out of 5), Kuwait (2 out of 2), Pakistan (1 out of 1), Hong Kong (1 out of 3) and Saudi Arabia (1 out of 3). These results showed that the most frequent variants which are circulated in Asia belong to the clade O.

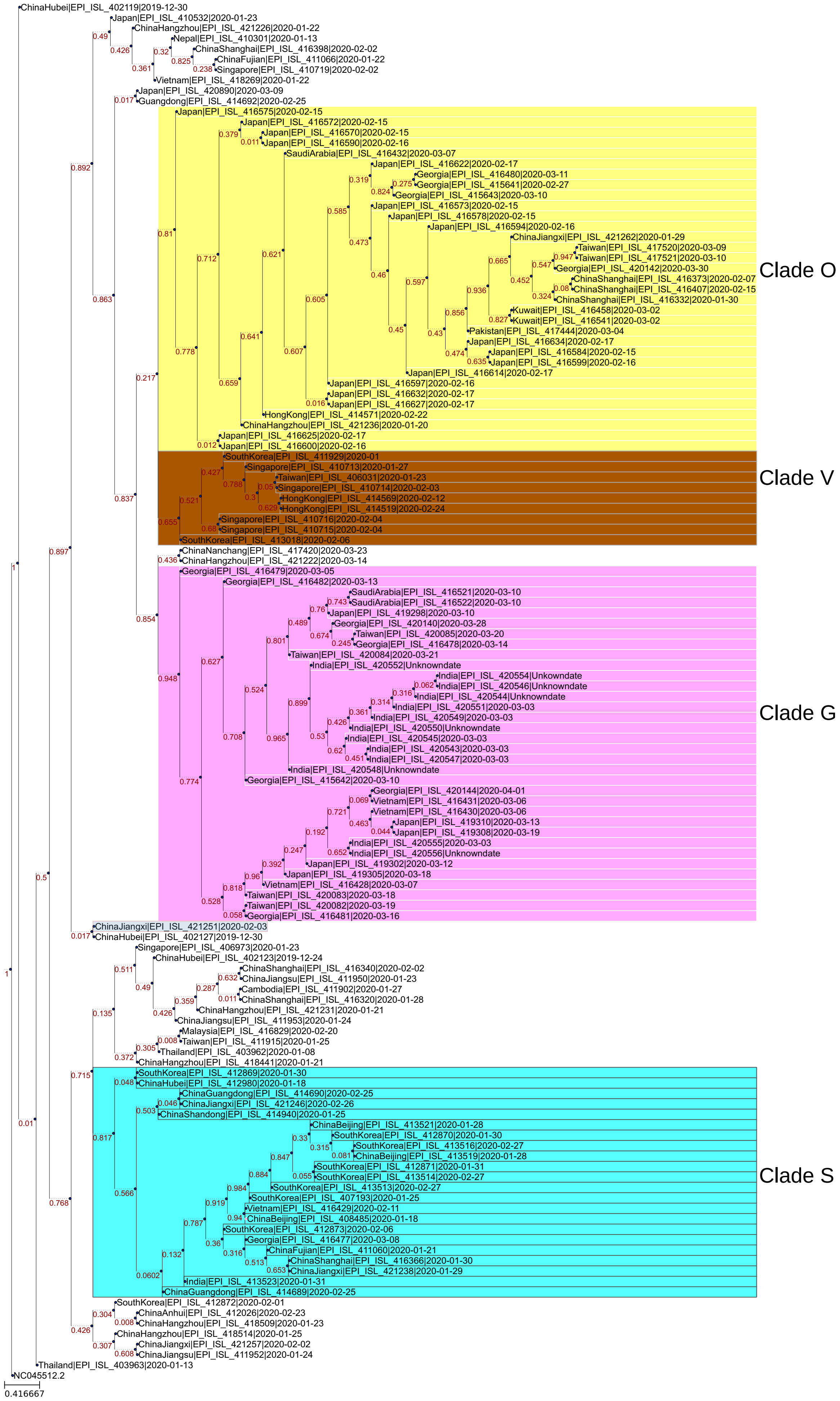
ChinaHubei|EPI_ISL_402119|2019-12-30
Japan|EPI_ISL_410532|2020-01-23
ChinaHangzhou|EPI_ISL_421226|2020-01-22
Nepal|EPI_ISL_410301|2020-01-13
ChinaShanghai|EPI_ISL_416398|2020-02-02
ChinaFujian|EPI_ISL_411066|2020-01-22
Singapore|EPI_ISL_410719|2020-02-02
Vietnam|EPI_ISL_418269|2020-01-22
Japan|EPI_ISL_420890|2020-03-09
Guangdong|EPI_ISL_414692|2020-02-25
Japan|EPI_ISL_416575|2020-02-15
Japan|EPI_ISL_416572|2020-02-15
Japan|EPI_ISL_416570|2020-02-15
Japan|EPI_ISL_416590|2020-02-16
SaudiArabia|EPI_ISL_416432|2020-03-07
Japan|EPI_ISL_416622|2020-02-17
Georgia|EPI_ISL_416480|2020-03-11
Georgia|EPI_ISL_415641|2020-02-27
Georgia|EPI_ISL_415643|2020-03-10
Japan|EPI_ISL_416573|2020-02-15
Japan|EPI_ISL_416578|2020-02-15
Japan|EPI_ISL_416594|2020-02-16
ChinaJiangxi|EPI_ISL_421262|2020-01-29
Taiwan|EPI_ISL_417520|2020-03-09
Taiwan|EPI_ISL_417521|2020-03-10
Georgia|EPI_ISL_420142|2020-03-30
ChinaShanghai|EPI_ISL_416373|2020-02-07
ChinaShanghai|EPI_ISL_416407|2020-02-15
ChinaShanghai|EPI_ISL_416332|2020-01-30
Kuwait|EPI_ISL_416458|2020-03-02
Kuwait|EPI_ISL_416541|2020-03-02
Pakistan|EPI_ISL_417444|2020-03-04
Japan|EPI_ISL_416634|2020-02-17
Japan|EPI_ISL_416584|2020-02-15
Japan|EPI_ISL_416599|2020-02-16
Japan|EPI_ISL_416614|2020-02-17
Japan|EPI_ISL_416597|2020-02-16
Japan|EPI_ISL_416632|2020-02-17
Japan|EPI_ISL_416627|2020-02-17
HongKong|EPI_ISL_414571|2020-02-22
ChinaHangzhou|EPI_ISL_421236|2020-01-20
Japan|EPI_ISL_416625|2020-02-17
Japan|EPI_ISL_416600|2020-02-16
Clade O
SouthKorea|EPI_ISL_411929|2020-01
Singapore|EPI_ISL_410713|2020-01-27
Taiwan|EPI_ISL_406031|2020-01-23
Singapore|EPI_ISL_410714|2020-02-03
HongKong|EPI_ISL_414569|2020-02-12
HongKong|EPI_ISL_414519|2020-02-24
Singapore|EPI_ISL_410716|2020-02-04
Singapore|EPI_ISL_410715|2020-02-04
SouthKorea|EPI_ISL_413018|2020-02-06
Clade V
ChinaNanchang|EPI_ISL_417420|2020-03-23
ChinaHangzhou|EPI_ISL_421222|2020-03-14
Georgia|EPI_ISL_416479|2020-03-05
Georgia|EPI_ISL_416482|2020-03-13
SaudiArabia|EPI_ISL_416521|2020-03-10
SaudiArabia|EPI_ISL_416522|2020-03-10
Japan|EPI_ISL_419298|2020-03-10
Georgia|EPI_ISL_420140|2020-03-28
Taiwan|EPI_ISL_420085|2020-03-20
Georgia|EPI_ISL_416478|2020-03-14
Taiwan|EPI_ISL_420084|2020-03-21
India|EPI_ISL_420552|Unknowndate
India|EPI_ISL_420554|Unkowndate
India|EPI_ISL_420546|Unknowndate
India|EPI_ISL_420544|Unknowndate
India|EPI_ISL_420551|2020-03-03
India|EPI_ISL_420549|2020-03-03
India|EPI_ISL_420550|Unknowndate
India|EPI_ISL_420545|2020-03-03
India|EPI_ISL_420543|2020-03-03
India|EPI_ISL_420547|2020-03-03
India|EPI_ISL_420548|Unknowndate
Georgia|EPI_ISL_415642|2020-03-10
Georgia|EPI_ISL_420144|2020-04-01
Vietnam|EPI_ISL_416431|2020-03-06
Vietnam|EPI_ISL_416430|2020-03-06
Japan|EPI_ISL_419310|2020-03-13
Japan|EPI_ISL_419308|2020-03-19
India|EPI_ISL_420555|2020-03-03
India|EPI_ISL_420556|Unknowndate
Japan|EPI_ISL_419302|2020-03-12
Japan|EPI_ISL_419305|2020-03-18
Vietnam|EPI_ISL_416428|2020-03-07
Taiwan|EPI_ISL_420083|2020-03-18
Taiwan|EPI_ISL_420082|2020-03-19
Georgia|EPI_ISL_416481|2020-03-16
Clade G
ChinaJiangxi|EPI_ISL_421251|2020-02-03
ChinaHubei|EPI_ISL_402127|2019-12-30
Singapore|EPI_ISL_406973|2020-01-23
ChinaHubei|EPI_ISL_402123|2019-12-24
ChinaShanghai|EPI_ISL_416340|2020-02-02
ChinaJiangsu|EPI_ISL_411950|2020-01-23
Cambodia|EPI_ISL_411902|2020-01-27
ChinaShanghai|EPI_ISL_416320|2020-01-28
ChinaHangzhou|EPI_ISL_421231|2020-01-21
ChinaJiangsu|EPI_ISL_411953|2020-01-24
Malaysia|EPI_ISL_416829|2020-02-20
Taiwan|EPI_ISL_411915|2020-01-25
Thailand|EPI_ISL_403962|2020-01-08
ChinaHangzhou|EPI_ISL_418441|2020-01-21
SouthKorea|EPI_ISL_412869|2020-01-30
ChinaHubei|EPI_ISL_412980|2020-01-18
ChinaGuangdong|EPI_ISL_414690|2020-02-25
ChinaJiangxi|EPI_ISL_421246|2020-02-26
ChinaShandong|EPI_ISL_414940|2020-01-25
ChinaBeijing|EPI_ISL_413521|2020-01-28
SouthKorea|EPI_ISL_412870|2020-01-30
SouthKorea|EPI_ISL_413516|2020-02-27
ChinaBeijing|EPI_ISL_413519|2020-01-28
SouthKorea|EPI_ISL_412871|2020-01-31
SouthKorea|EPI_ISL_413514|2020-02-27
SouthKorea|EPI_ISL_413513|2020-02-27
SouthKorea|EPI_ISL_407193|2020-01-25
Vietnam|EPI_ISL_416429|2020-02-11
ChinaBeijing|EPI_ISL_408485|2020-01-18
SouthKorea|EPI_ISL_412873|2020-02-06
Georgia|EPI_ISL_416477|2020-03-08
ChinaFujian|EPI_ISL_411060|2020-01-21
ChinaShanghai|EPI_ISL_416366|2020-01-30
ChinaJiangxi|EPI_ISL_421238|2020-01-29
India|EPI_ISL_413523|2020-01-31
ChinaGuangdong|EPI_ISL_414689|2020-02-25
Clade S
SouthKorea|EPI_ISL_412872|2020-02-01
ChinaAnhui|EPI_ISL_412026|2020-02-23
ChinaHangzhou|EPI_ISL_418509|2020-01-23
ChinaHangzhou|EPI_ISL_418514|2020-01-25
ChinaJiangxi|EPI_ISL_421257|2020-02-02
ChinaJiangsu|EPI_ISL_411952|2020-01-24
Thailand|EPI_ISL_403963|2020-01-13
NC045512.2
0.416667

Figure 4: Phylogenetic analysis of SARS-CoV-2 sequences

Discussion

In the current study, we performed comprehensive phylogenetic analyses of the SARS-CoV-2 to investigate the Asian epidemic spread of this virus. Although, the aim of this study was not to determine the functional consequences of the detected mutations, they might implicate in the functionality of the virus.

Our results showed the maximum of the variations were related to ORF1ab polyprotein (52%) while the minimum was related to ORF7b and E protein sequences (0.5%). These findings were in accordance with previous published works where almost 65% of variations were observed in an investigation by Koyama et al. [21][22, 23].

The most common mutation types found in this study were missense (62%) followed by synonymous, nonsense, and non-coding SNPs were 62%, 33%, 0.5%, and 4%, respectively (***Figure 1***) by which three amino acid changes was mostly observed in ORF1ab polyprotein. The most distributed variations among Asian countries (≥ 5 countries) were 241C>T, 1397G>A, 3037C>T, 8782C>T, 11083G>T, 14408C>T, 23403A>G, 26144G> and 28144G>T (***Figure 3***). Our results reported five types of co-mutations.

It has been previously reported that the most frequent mutation was the C>T transition (55.1%), followed by A>G transition (14.8%) in Europe, the Americas and Africa resulting in 205,482 amino acid changing (58.2% of the total) [24]. The pivotal role of these mutations has been previously shown to provide a stronger transmission capacity of SARS-CoV-2 whereas there is no indication for others to increase transmissibility of SARS-CoV-2. [25, 26]. For example, Chen et al. indicated the frequency of some mutations, particularly (V367F, S477N, N439K, V483A) lead to stronger transmission capacity [25].

There are four important clades in Asia including clade G (D614G variant of S glycoprotein), clade V (G251V variant of ORF3a protein product), clade S (L84S variant of ORF8 protein product) and clade O (***Figure 4***). Continuous monitoring of mutations could be critical in following the movement of SARS-CoV-2 between individuals and across geographical regions. For instance, the analysis of the clades, in this study, throughout the year revealed the original L clade was first reported in Asia (China) in December 2019, subsequently the G clade was observed in Europe in January 2020. In March 2020, G and G-derived clades have then been detected in North America and Asia and are known as the most rapidly growing viral subpopulation globallyb[24, 27, 28]. Clade L, as the original viral strain, is 7% of all sequences, followed by clades S and V which have the same frequencies worldwide [24]. Previous studies have shown that clade G is the most prevalent clade in European countries which is also true in Asia as our analysis showed[24]. The G clade and its derived clades, GH and GR, are the most widespread and more frequent clades observed in SARS-CoV-2 genomes (74% of all reported sequences) [24]. The most frequent clade was the D614G variant which is in accordance with our study. Due to the location of this variant, in a B-cell epitope, it could be speculated that D614G variant influences vaccine effectiveness. [29]. The collected data from new variants might help to develop novel antiviral drug candidates, also the adaptation of current ones to tackle the new structural features of SARS-CoV-2 [13]. Another role of D614G variant is on speed of viral replication. The reason is that all strains with this mutation have another mutation in their proteins that is involved in viral replication (RdRp P323L) [21]. RdRp is a target of favipiravir and remdesivir, two well-known drug candidates against COVID-19, therefore these mutations may lead to emergence of drug-resistant strains [30].

In our study, the sequences which are clustered in the clade S (L84S) were collected in January and February 2020 indicating the early frequent transmission of the virus was from China to South Korea. The first mutation-derived from clade L (as the reference genome) was clade S and appeared at the beginning of 2020 [24] which was mostly observed in the Americas [21, 31]. Similar to our results, L84S clade was the second major clade which was detected among passengers from Wuhan during the early days of the coronavirus [21]. Third clade in the current study was clade V. Clade V has low variability and it is not included in the co-mutation list which is compatible with the early collection date of the included sequences. This clade was not a prevalent type in the current study, although the previous studies have demonstrated that clade V is the most prevalent clade in Europe and Asia. A previous study demonstrated that this clade appeared mid-January 2020 (simultaneously with the original clade G) [24]. Also, we had the clade O, carrying a variation of 11083G>T that most of the Asian sequences belonged to. This mutation was co-evolved with 1397G>A, 28688T>C, and 29742G>T. In the previous studies, this clade was described as a group of sequences that did not follow any criteria [24]. Although, in this study we found that this group has common co-mutations.

In summary, we report that there are fifteen frequent variations in Asia. Most of the sequences have been collected in clade G, S, and O. All of these three clades include sequences with co-mutations. Clade V has the least variability in comparison to the other clades, has not any co-mutations and is not a prevalent clade in our analysis. Taken together, all of these evidence indicate the circulated Asian viruses in the early pandemic had evolved and had differences with the early ancestors.

Conclusion

This study provides an overview of the most prevalent mutations and evolutionary relationships among the sequences from Asian countries in the first months of the outbreak. The results presented the geographical variations that can be used to find transmission roots. Some of these variants are expected to implicate in viral and host factors, which can impact on transmission and disease severity. Although the lack of complete genome sequence from some countries and imbalanced frequency of sequences in the available countries in that period of time would affect the outcome.

Acknowledgments

We gratefully acknowledge the authors and the originating and submitting laboratories of the sequences from GISAID database.

Conflict of interests

The authors declare that there are no conflicts of interests.

Author Contributions

JM contributed in data collection, data analysis and drafting the manuscript, MM discussed the results and MD and RA contributed to finalizing the writing of the manuscript.

References

1. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB: **Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)**. *Nucleic acids research* 2018, **46**(D1):D708-D717.
2. Zeng Z-Q, Chen D-H, Tan W-P, Qiu S-Y, Xu D, Liang H-X, Chen M-X, Li X, Lin Z-S, Liu W-K: **Epidemiology and clinical characteristics of human coronaviruses OC43, 229E, NL63, and HKU1: a study of hospitalized children with acute respiratory tract infection in Guangzhou, China**. *European Journal of Clinical Microbiology & Infectious Diseases* 2018, **37**(2):363-369.
3. Vabret A, Mourez T, Gouarin S, Petitjean J, Freymuth F: **An outbreak of coronavirus OC43 respiratory infection in Normandy, France**. *Clinical infectious diseases* 2003, **36**(8):985-989.
4. Pene F, Merlat A, Vabret A, Rozenberg F, Buzyn A, Dreyfus F, Cariou A, Freymuth F, Lebon P: **Coronavirus 229E-related pneumonia in immunocompromised patients**. *Clinical infectious diseases* 2003, **37**(7):929-932.
5. Pyrc K, Berkhout B, Van Der Hoek L: **The novel human coronaviruses NL63 and HKU1**. *Journal of virology* 2007, **81**(7):3051-3057.
6. Lau SK, Woo PC, Yip CC, Tse H, Tsoi H-w, Cheng VC, Lee P, Tang BS, Cheung CH, Lee RA: **Coronavirus HKU1 and other coronavirus infections in Hong Kong**. *Journal of clinical microbiology* 2006, **44**(6):2063-2071.
7. Zhong N, Zheng B, Li Y, Poon L, Xie Z, Chan K, Li P, Tan S, Chang Q, Xie J: **Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003**. *The Lancet* 2003, **362**(9393):1353-1358.
8. Zaki AM, Van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA: **Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia**. *New England Journal of Medicine* 2012, **367**(19):1814-1820.
9. of the International CSG: **The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2**. *Nature Microbiology* 2020, **5**(4):536.
10. Peiris JS, Yuen KY, Osterhaus AD, Stöhr K: **The severe acute respiratory syndrome**. *New England Journal of Medicine* 2003, **349**(25):2431-2441.
11. Organization WH: **Middle East respiratory syndrome coronavirus (MERS-CoV)**. In.; 2019.
12. Cui J, Li F, Shi Z-L: **Origin and evolution of pathogenic coronaviruses**. *Nature Reviews Microbiology* 2019, **17**(3):181-192.
13. Amanat F, Krammer F: **SARS-CoV-2 vaccines: status report**. *Immunity* 2020.
14. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, Zhou Y, Du L: **Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine**. *Cellular & molecular immunology* 2020, **17**(6):613-620.
15. Yin C: **Genotyping coronavirus SARS-CoV-2: methods and implications**. *Genomics* 2020.
16. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Molecular biology and evolution* 2013, **30**(4):772-780.
17. Okonechnikov K, Golosova O, Fursov M, Team U: **Unipro UGENE: a unified bioinformatics toolkit**. *Bioinformatics* 2012, **28**(8):1166-1167.

18. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A: **RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference**. *Bioinformatics* 2019, **35**(21):4453-4455.
19. Lemoine F, Entfellner J-BD, Wilkinson E, Correia D, Felipe MD, De Oliveira T, Gascuel O: **Renewing Felsenstein's phylogenetic bootstrap in the era of big data**. *Nature* 2018, **556**(7702):452-456.
20. Huerta-Cepas J, Serra F, Bork P: **ETE 3: reconstruction, analysis, and visualization of phylogenomic data**. *Molecular biology and evolution* 2016, **33**(6):1635-1638.
21. Koyama T, Platt D, Parida L: **Variant analysis of SARS-CoV-2 genomes**. *Bulletin of the World Health Organization* 2020, **98**(7):495.
22. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CC, Boshier FA: **Emergence of genomic diversity and recurrent mutations in SARS-CoV-2**. *Infection, Genetics and Evolution* 2020:104351.
23. Khailany RA, Safdar M, Ozaslan M: **Genomic characterization of a novel SARS-CoV-2**. *Gene reports* 2020:100682.
24. Mercatelli D, Giorgi FM: **Geographic and Genomic Distribution of SARS-CoV-2 Mutations**. 2020.
25. Chen J, Wang R, Wang M, Wei G-W: **Mutations strengthened SARS-CoV-2 infectivity**. *arXiv preprint arXiv:200514669* 2020.
26. van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F: **No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2**. *Nature Communications* 2020, **11**(1):5986.
27. Hufsky F, Lamkiewicz K, Almeida A, Aouacheria A, Arighi C, Bateman A, Baumbach J, Beerenwinkel N, Brandt C, Cacciabue M: **Computational Strategies to Combat COVID-19: Useful Tools to Accelerate SARS-CoV-2 and Coronavirus Research**. 2020.
28. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi FM: **coronapp: a Web Application to Annotate and Monitor SARS-CoV-2 Mutations**. *bioRxiv* 2020.
29. Koyama T, Weeraratne D, Snowdon JL, Parida L: **Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment**. *Pathogens* 2020, **9**(5):324.
30. Li C, Wang L, Ren L: **Antiviral mechanisms of candidate chemical medicines and traditional Chinese medicines for SARS-CoV-2 infection**. *Virus research* 2020:198073.
31. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF: **The proximal origin of SARS-CoV-2**. *Nature medicine* 2020, **26**(4):450-452.
32. Guan Q, Sadykov M, Mfarrej S, Hala S, Naeem R, Nugmanova R, Al-Omari A, Salih S, Al Mutair A, Carr MJ: **A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic**. *International Journal of Infectious Diseases* 2020, **100**:216-223.