doi:10.20944/preprints202005.0100.v1

Genomic characterization and phylogenetic analysis of SARS-CoV-2 in Asia

Jale Moradi*

Microbiology Department, School of Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran

*Corresponding Author: Jale Moradi, E-mail: Jale.Moradi@kums.ac.ir

Abstract

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) as the current coronavirus pandemic is an infectious disease that initially confirmed in China in late December 2019. In this study, we analyzed 131 complete sequences of SARS-CoV-2 from Asia. Our results show that there are fifteen major mutations in Asia which most of them are co-evolved. There were five groups based on co-mutations which three of them resulted in clade G including (241C>T, 3037C>T, 14408C>T, and 23403A>G), (28881G>A, 28882G>A, 28883G>C and 23403A>G) and (25563G>T and 23403A>G). Co-mutations in (8782C>T and 28144T>C) and (1397G>A, 28688T>C, 29742G>T and 11083G>T) were clustered in clade S and a new clade outside of GISAID classification, respectively. Sequences with a mutation in 26144G>T had low variability without any co-mutation which formed clade V. In this study, we showed that Most of the circulated viruses in Asia collected in five co-mutation groups which may affect the transmissibility and vaccine designing strategies.

doi:10.20944/preprints202005.0100.v1

Introduction

Coronavirus species belong to the *Coronaviridae* family, two subfamily *Letovirinae* with one genus as *Alphaletovirus* and *Orthocoronavirinae* with four genera including *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus,* and *Gammacoronavirus* (1). To date, seven Coronavirus species have caused infection in human, which 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 which are the cause of common colds and considered as modest clinical importance (3,4). *Alphacoronavirus* HcoV-NL63 and *Betacoronavirus* HcoV-HKU1 are the cause of bronchiolitis in children and community-acquired pneumonia, respectively (5,6) . There are three additional Coronaviruses which are highly pathogenic and have caused the epidemic in human populations. *severe acute respiratory syndrome coronavirus* (SARS-Cov) a *Betacoronavirus*, subgenus *Sarbecovirus,* was identified in China in 2002 and spread to the 29 countries with the mortality rate of ∼10% and abruptly ended in 2003, *Middle East respiratory syndrome coronavirus* (MERS-CoV) a *Betacoronavirus*, subgenus *Merbecovirus* emerged in Saudi Arabia in 2012 with ∼34% mortality rate which has detected in 27 countries and *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) is a *Betacoronavirus*, subgenus *Sarbecovirus* is a novel coronavirus that initially detected in Wuhan city of Hubei province in China in the late December of 2019 (7–11). Based on the WHO report, 213 countries, area, or territories have been confirmed SARS-CoV-2 infection with 3435894 cases and 239604 death as of May 4, 2020.

Coronaviruses are enveloped positive-strand RNA viruses with around 30 kb genome length. The 5' end encodes a poly-protein and consists of two-third of the genome length which cleaved to 16 non-structural proteins (Nsp1 to Nsp16). The 3' end consists of at least four structural proteins including envelope glycoproteins spike (S), membrane (M), envelope (E), and nucleocapsid (N) (12).

Genotyping analysis is an important tool to determine the mutations in the fundamental part of the SARS-CoV-2 genome. Any variations in the vaccine candidate proteins such as structural proteins need to be analyzed before vaccine design. Also, genotyping data can be used to predict the efficacy of the global vaccine in some countries with a high rate of important mutations (13–15). Therefore, in the current study, we analyzed the complete sequences of SARS-CoV-2 to reveal genetic distance and mutation rates among Asian countries.

Materials and Methods

The submitted complete sequences of SARS-CoV-2 from Asian countries and reference sequence were retrieved from GISAID (www.gisaid.org) and GeneBank, respectively. The Alignment was performed with MAFFT (v7.455) (16). The alignment result was 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 analysis of the resulted tree (20) . Nucleotide substitutions were retrieved from the output of alignment and annotated to find protein changes.

Results and Discussion

4572 SARS-CoV-2 complete sequences were released in GISAID as April 7, 2020. 641 sequences belonged to Asia. Among 641 GISAID Asian sequences, 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 select sequences related to the patients except for the regions with the lack of human samples. 131 sequences were selected in 12 Asian countries, all of the sequences were related to the patients except for Vietnam sequences which were belong to the Vero cells (Table 1). The sequences with their related GISAID accession numbers were deposited in Supplementary Table 1. The

SARS-CoV-2 reference sequence obtained from GenBank with the accession number of NC_045512.2.

Table 1: Frequency of SARS-Cov-2 Sequences from Asian countries involved in this study.

Country	Total	Select
Georgia	13	12
Hong Kong	63	3
India	16	14
Japan	102	24
Kuwait	4	2
Malaysia	7	1
Nepal	1	1
Pakistan	2	1
Saudi Arabia	3	3
Singapore	37	6
South Korea	13	11
Taiwan	22	8
Thailand	2	2
Vietnam*	8	5
Cambodia	1	1
China	310	37
	604	131

China Regions	Total	Select
Anhui	2	1
Beijing	5	3
Chongqing	3	0
Fujian	2	2
Guangdong	74	3
Guangxi	6	0
Guangzhou	1	0
Hangzhou	36	5
Henan	1	0
Hubei	32	4
Jiangsu	4	3
Jiangxi	26	5
NanChang	1	1
Shandong	9	1
Shanghai	96	7
Sichuan	1	0
Yunnan	2	0
Zhejiang	9	2
	310	37

* All sequences related to the Vero cells

All retrieved sequences were aligned and trimmed based on NC_045512.2. All sequences were trimmed to 29698 bp and a total number of 538 variations were identified, which were related to the 197 variable sites (Supplementary Table 2). The variations were distributed through the SARS-Cov-2 genomes, except the ORF6 region. Near half of the variations were associated with ORF1ab polyprotein (52%) and minimum frequencies were related to the ORF7b and E protein sequences (0.5%). Based on 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 are related to the Chinese sequences.

doi:10.20944/preprints202005.0100.v1

Cambodia, Nepal, Thailand, and Malaysia had less than 2% of variations in their sequences.

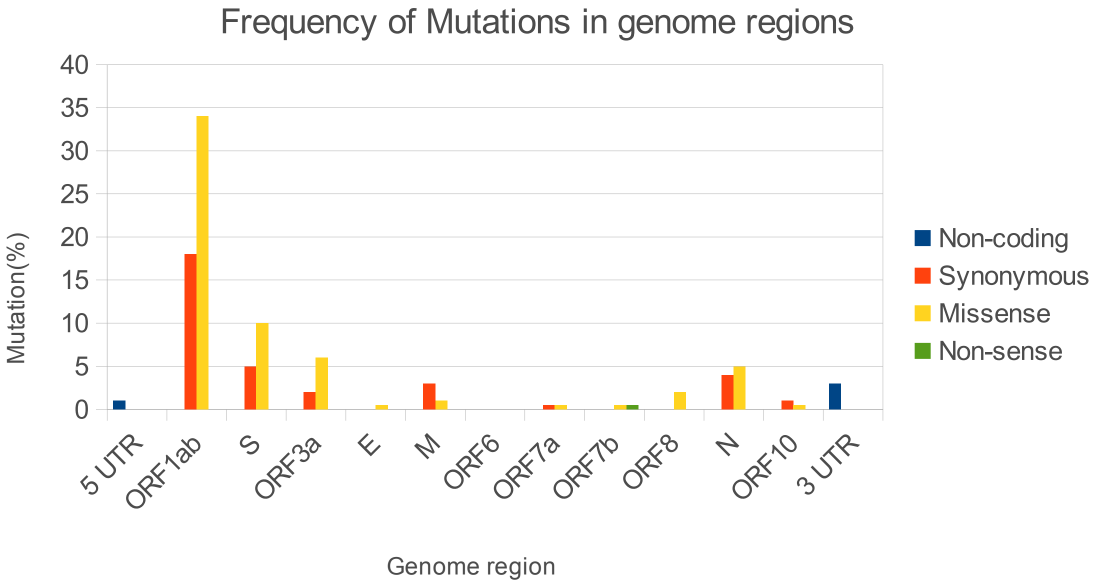

Figure 1: Distribution of SARS-CoV-2 mutations in Sequences from Asia.

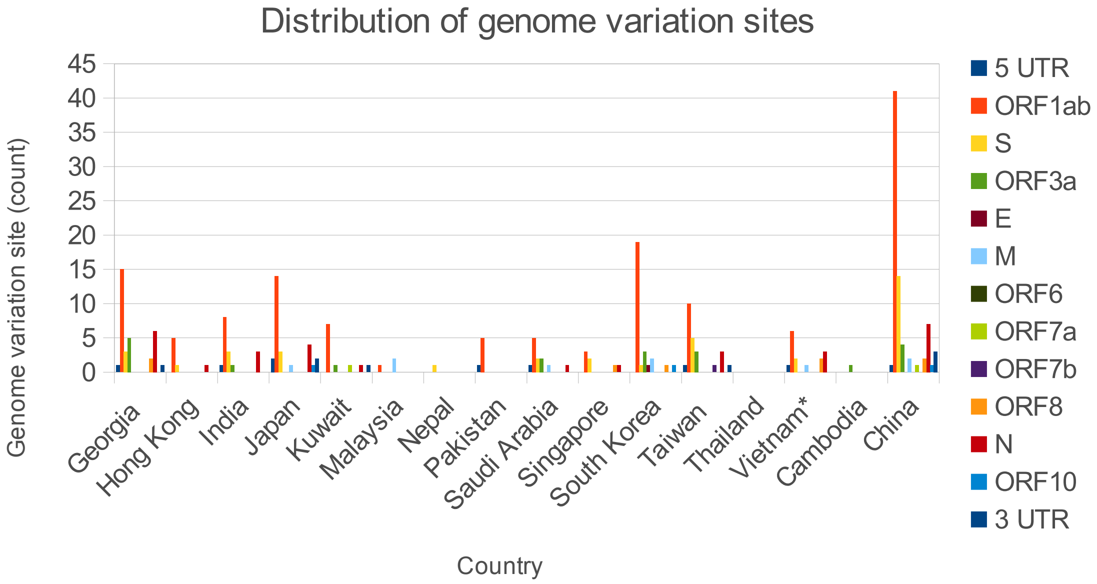

Figure 2: Frequency of SARS-Cov-2 genome variation sites in sequences from Asia.

Fifteen out of 197 variable sites were the most prevalent variations in sequences from Asia (Table 2). Most of the prevalent mutations were related to the ORF1ab gene with three amino acid changes in this polyprotein. Mutations in 241C>T, 1397G>A, 3037C>T, 8782C>T, 11083G>T, 14408C>T, 23403A>G, 26144G> and 28144G>T were the most distributed variations among Asian countries(≥ 5 countries) (Figure 3). The other most common variations were distributed in four countries.

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

Genome Change	Gene	Protein Change	Mutation type	Clade
241C>T	5UTR	Non-coding	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	G
25563G>T	ORF3a	Q57H	Missense	
26144G>T		G251V	Missense	V
28144T>C	ORF8	L84S	Missense	S
28688T>C	N		Synonymous	
28881G>A		R203K	Missense	
28882G>A			Synonymous	
28883G>C		G204R	Missense	
29742G>T	3UTR		Non-coding	

We analyzed occurrences of co-mutations in Asian sequences and found 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.

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	G
25563G>T, 23403A>G	G
8782C>T, 28144T>C	S
1397G>A, 28688T>C, 29742G>T, 11083G>T	

*Clade classification is based on underlined mutations

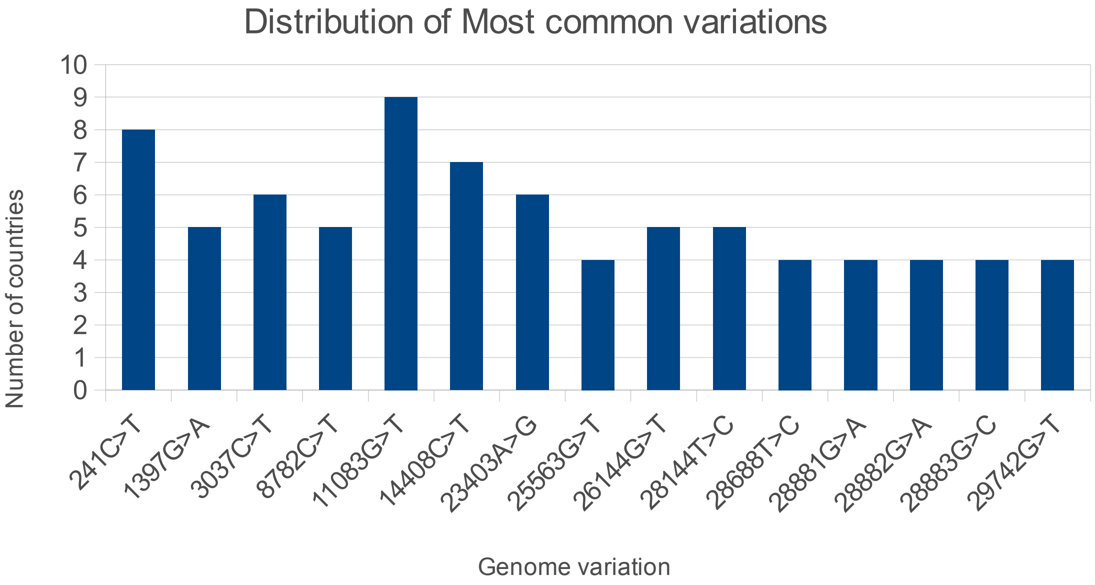

Figure 3: Distribution of SARS-Cov-2 most common mutations in Asia.

Phylogenetic analysis of complete sequences of SARS-CoV-2 from Asia showed that there were four important clades in Asia, which three of them included in the 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 our study, Clade S with 22 sequences was located in the base of the tree which is 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 prevalent clade in USA and China. The sequences which are clustered in this clade were collected in January and February. These results indicate the early frequent transmission of the virus from china to south Korea. Previous studies have shown that clade G is the most prevalent clade in European countries which our results showed that this clade defined as an important type in Asia as the same. In our study, this clade 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) which

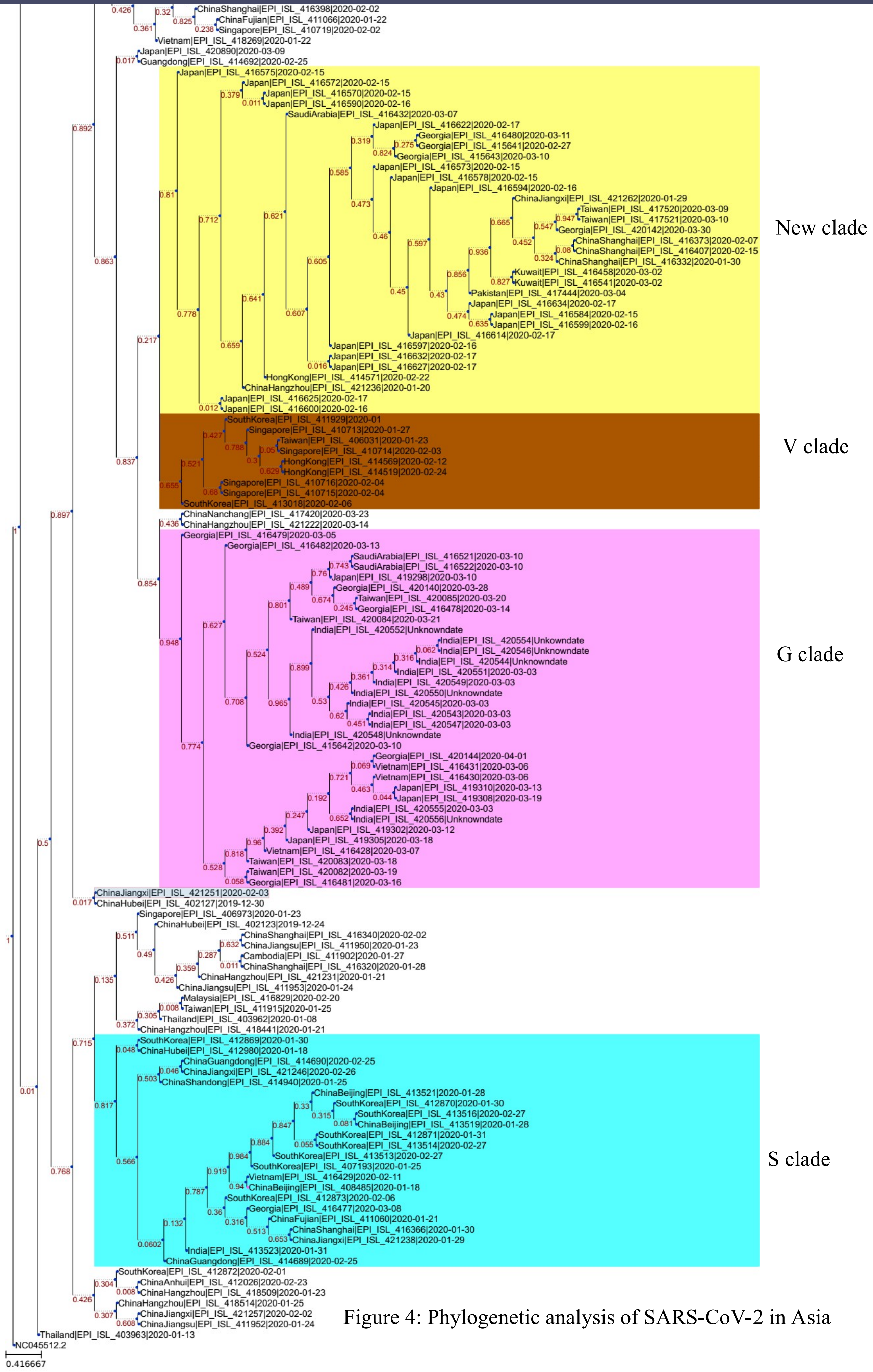
ChinaShanghai|EPI_ISL_416398|2020-02-02
ChinaFujian|EPI_ISL_411066|2020-01-22
Singapore|EPI_ISL_410719|2020-02-02
Vietnam|EPI_ISL_418269|2020-01-22
Japan|EPI_ISL_420890|2020-03-09
Guangdong|EPI_ISL_414692|2020-02-25
Japan|EPI_ISL_416575|2020-02-15
Japan|EPI_ISL_416572|2020-02-15
Japan|EPI_ISL_416570|2020-02-15
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
New clade
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
V clade
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
G clade
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
S clade
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 in Asia

all the sequences were collected in March and April. These results show that transmission to India with the most number of sequences in this clade is done recently. Clade v which is shown that related to Europe and Asia is the limited clade in our study. This clade was including eleven 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). Our study shows that clade V has low variability and has not included in the co-mutation list which is compatible with the early collection date of the included sequences. New clade outside of GISAID classification was included the sequences which carried 11083G>T mutation. This mutation was co-evolved with 1397G>A, 28688T>C, and 29742G>T. This new 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). Our results show that the most frequent variants which are circulated in Asia belong to this new clade.

In summary, we report that there are fifteen frequent mutations in Asia. Most of the sequences have been collected in clade G, S, and new clade related to the 11083G>T. All of these three clades include sequences with co-mutations. Clade V has less variability in comparison to the mentioned three clades without any co-mutations that a small number of our studied sequences were located in this clade. Taken together, all of this evidence indicates the circulated viruses in Asia have evolved and have differences with the early ancestors.

Conclusion

The results of this study can provide an overview of the most prevalent mutations and evolutionary relationships among the sequences from Asian countries. The Results show the geographical variations that can be used to find transmission roots. Although, the lack of complete genome sequence from some countries and imbalanced frequency of sequences in the available countries would affect the outcome. Therefore, future analysis with the sequence availability from all of the countries in Asia will complete these results.

Acknowledgments

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

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 Res. 2018 Jan;46(D1):D708–17.
2. Zeng Z-Q, Chen D-H, Tan W-P, Qiu S-Y, Xu D, Liang H-X, et al. 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. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2018 Feb;37(2):363–9.
3. Vabret A, Mourez T, Gouarin S, Petitjean J, Freymuth F. An outbreak of coronavirus OC43 respiratory infection in Normandy, France. Clin Infect Dis an Off Publ Infect Dis Soc Am. 2003 Apr;36(8):985–9.
4. Pene F, Merlat A, Vabret A, Rozenberg F, Buzyn A, Dreyfus F, et al. Coronavirus 229E-related pneumonia in immunocompromised patients. Clin Infect Dis an Off Publ Infect Dis Soc Am. 2003 Oct;37(7):929–32.
5. Pyrc K, Berkhout B, van der Hoek L. The novel human coronaviruses NL63 and HKU1. J Virol. 2007 Apr;81(7):3051–7.
6. Lau SKP, Woo PCY, Yip CCY, Tse H, Tsoi H, Cheng VCC, et al. Coronavirus HKU1 and other coronavirus infections in Hong Kong. J Clin Microbiol. 2006 Jun;44(6):2063–71.
7. Zhong NS, Zheng BJ, Li YM, Poon, Xie ZH, Chan KH, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People’s Republic of China, in February, 2003. Lancet (London, England). 2003 Oct;362(9393):1353–8.
8. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012 Nov;367(19):1814–20.

9. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 Apr;5(4):536–44.

10. Peiris JSM, Yuen KY, Osterhaus ADME, Stöhr K. The severe acute respiratory syndrome. N Engl J Med. 2003 Dec;349(25):2431–41.

11. WHO. Middle East respiratory syndrome coronavirus (MERS-CoV) [Internet]. Available from: https://www.who.int/emergencies/mers-cov/en/

12. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019 Mar;17(3):181–92.

13. Amanat F, Krammer F. SARS-CoV-2 Vaccines: Status Report. Immunity. 2020 Apr;

14. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, et al. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. Cell Mol Immunol. 2020 Mar;

15. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. medRxiv [Internet]. 2020; Available from: http://arxiv.org/abs/2003.10965

16. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772–80.

17. Okonechnikov K, Golosova O, Fursov M. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012 Apr;28(8):1166–7.

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 Nov;35(21):4453–5.

19. Lemoine F, Domelevo Entfellner J-B, Wilkinson E, Correia D, Dávila Felipe M, De Oliveira T, et al. Renewing Felsenstein’s phylogenetic bootstrap in the era of big data. Nature. 2018 Apr;556(7702):452–6.

20. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016 Jun;33(6):1635–8.

21. 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 [Internet]. 2020 Mar