Emergence of European and North American mutant variants of SARS-CoV-2 in Southeast Asia

Ovinu Islam[1], Hassan Al-emran[2], Md. Hasan[1], Azraf Anwar[3], Md. Jahid[1], and M. Hossain[4]

[1]Jashore University of Science & Technology
[2]Jashore University of Science and Technology
[3]Affiliation not available
[4]University of Dhaka

June 2, 2020

Abstract

The SARS-CoV-2 strain of the coronavirus is responsible for the current COVID-19 pandemic, with an ongoing toll of over 5 million infections and 333 thousand deaths worldwide within the first 5 months. Insight into the phylodynamics and mutation variants of this strain is vital to understanding the nature of its spread in different climate conditions. The incidence rate of COVID-19 is increasing at an alarming pace within subtropical Southeast Asian nations with high temperatures and humidity. To understand this spread, we analyzed 60 genome sequences of SARS-CoV-2 available in GISAID platform from 6 Southeast Asian countries. Multiple sequence alignments and maximum likelihood phylogenetic analyses were performed to analyze and characterize the non-synonymous mutant variants circulating in this region. Global mutation distribution analysis showed that the majority of the mutations found in this region are also prevalent in Europe and North America, and the concurrent presence of these mutations at a high frequency in Australia and Saudi Arabia indicate possible transmission routes. Unique spike protein and non-structural protein mutations were observed circulating within a localized area. We divided the circulating viral strains into 4 major groups and 2 sub-groups on the basis of the most frequent non-synonymous mutations. Strains with a unique set of 4 co-evolving mutations were found to be circulating at a high frequency within India, specifically, group 2 strains characterized by two co-evolving NS mutants which alter in RdRp (P323L) and spike protein (D614G) common in Europe and North America. These European and North American variants (Nextstrain clade A2) have rapidly emerged as dominant strains within Southeast Asia, increasing from a 0% presence in January to an 85% presence by May 2020. These variants may have an evolutionary advantage over their ancestral types and could present the largest threat to Southeast Asia for the coming winter.

Introduction

The SARS-CoV-2 ssRNA virus was initially detected in December 2019 within China's Hubei province and its associated COVID-19 pandemic is currently ongoing, with a global toll of over 5 million infections and 333 thousand deaths so far (WHO, Corona virus disease (COVID-19) situation report-124, published on 23rd May, 2020). North American and European nations have already suffered a crippling blow from this pandemic, and the infection numbers within Southeast Asian, South American and African nations are growing. However, the number of infections in resource-poor, low-income countries may be underestimated due to their limited test facilities and skilled personnel. A number of earlier studies indicated that high temperatures and high humidity could decrease the spread parameter of the virus (Demongeot, Flet-Berliac & Seligmann, 2020; Sajadi et al., 2020; Wang, Tang, Feng, & Lv, 2020). Despite having such climate conditions, COVID-19 cases are increasing at an alarming rate in relatively hot and humid subtropical Southeast Asian countries. As of 22nd May, a total of 182,278 cases of SARS-CoV-2 infected cases were identified in Southeast Asia, with a death toll of over 5,500 (WHO, Corona virus disease (COVID-19) situation report-124, published on 23rd May, 2020).

Scientists from all over the world are making an unprecedented effort to expose the genomic profiles, and characterize the mutation variants, of the circulating virus to get insight into its evolutionary patterns and driving force. Tang, X.L. et al. analyzed 103 genomic sequences and indicated that the circulating SARS-CoV-2 strains have two major lineages, one with a synonymous mutation (NSP4_S75S) and the other with a non-synonymous mutation (NS8_L84S) (Tang et al., 2020). In another study, Forster, Peter et al. found 3 central variants by analyzing 160 sequences and claimed that B type viruses (with amino acid substitution, NS8_L84S) were common in East Asia, whereas A type (ancestral lineage) and C type (NS3_-G251V variant) were prevalent in Europe and North America (Forster, Forster, Renfrew, & Forster, 2020). Changchuan Yin analyzed 558 genome sequences and found 15 high frequency single SNP genotypes (Yin, 2020). He suggested four major groups with either single or co-evolving mutations; group 1 with NSP6_-L37F, group 2 with NS3_G251V, group 3 with co-evolving mutation at 2 sites (NSP4_S75S and NS8_L84S) and group 4 with co-evolving mutations at 4 sites (241C > T- leader sequence, NSP3_F105F, NSP12_-P323L and spike_D614G). In addition, GISAID differentiated COVID-19 into three major clades; Clade S (prevalent in North America), Clade V (prevalent in Asia and Europe) and Clade G (prevalent in Europe), having non-synonymous mutations with amino acid substitutions at NS8_L84S, NS3_G251V and spike_D614G respectively (Fuertes et al., 2020). Finally, based on genomic data available from the GISAID, Nextstrain's SARS-CoV-2 global genomic epidemiology analysis show 10 major clades for this virus. In this study, we investigated 60 sequenced genomes of SARS-CoV-2 available from Southeast Asia to provide a proximate insight into the genomic divergence, phylogeny and recurrent mutations in this region.

Methods and Materials

Data Collection

In Southeast Asia, a total of 411 complete or near complete genomes from India (n=355), Bangladesh (n=20), Indonesia (n=9), Thailand (n=122), Sri Lanka (n=4), Nepal (n=1) and Myanmar (n=1) have been submitted to GISAID (www.gisaid.org) as of 23rd May, 2020. This study analyzed 60 genome sequences of SARS-CoV-2 with 30 genome sequences coming from India, 10 from Bangladesh, 8 from Indonesia, 7 from Thailand, 4 from Sri Lanka and one from Nepal. While another sequence from Myanmar was available in GISAID, this sequence was excluded from this study due to the poor quality of the data. The ratio of sequence by country was determined based on the available number of sequences and the total number of Covid-19 cases up to 23rd of May, 2020. The selection of sequences analyzed was based on genome quality, discreteness of their location and random sample collection dates. hCoV19 / Wuhan / WIV04 / 2019 (Accession: EPI_ISL_-402124) was used as a reference genome for phylogenetic and mutation analysis. Accessions ID, collection dates and locations are provided in supplementary Table-s1.

Phylogenetic and evolutionary analyses

All selected sequences were aligned using the MAFFT (Multiple Alignment using Fast Fourier Transform) algorithm (Katoh, Misawa, Kuma, & Miyata, 2002) and the alignment was visualized in JalView 2.11.0. The aligned sequences were used for the construction of a phylogenetic tree using the neighbor-joining method, along with 500 bootstrap replications and a 95% site coverage cutoff value, in Molecular Evolutionary Genetics Analysis (MEGA X) software (Kumar, Stecher, Li, Knyaz, & Tamura, 2018). Interactive Tree Of Life (iTOL) v5 (Letunic & Bork, 2019) was used to adjust the branch and label color of the phylogenic tree. Non-synonymous mutations with their specific mutation sites and their global frequencies were determined using 'CoVsurver enabled by GISAID' based on viral sequences in GISAID's EpiCoV database. This data was checked and validated carefully against aligned sequences. 3D structural visualization of the spike glycoproteins was performed using the same application that annotates the structural positions of mutations and amino acid substitutions based on processing coronavirus crystal structures in PDB. Countries with a high frequency of specific mutations were marked on a geographic heat map using the online tool Maptive. A mutation-time plot for Southeast Asia was prepared by analyzing the sequence during the first 15 days of each month available in GISAID. Additionally, the numbers of COVID-19 infections were collected for each month from January to May 2020.

Phylogenetic analysis and transmission patterns (performed on 23nd May, 2020) were observed by analyzing 329 genome sequences from Southeast Asia (India 220, Bangladesh 13, Thailand 80, Indonesia 9, Sri Lanka 6, Nepal 1) using country filters on Nextstrain, an online based real-time pathogen evolution tracking tool (Hadfield et al., 2018).

Results

Phylogenetic and mutation analysis of 60 SARS-CoV-2 sequences from Southeast Asia revealed 78 non-synonymous mutations. Most of them (n=52) were found in non-structural (NS) proteins. Other mutations with amino acid (aa) substitutions were present in spike protein (n=13), N protein (n=9), M protein (n=3) and E protein (n=1). The Nepal SARS-CoV-2 genome, which was sequenced early, had no NS mutations compared to our reference sequence (Table-1).

This study identified 21 NS mutations including 4 aa alterations in spike proteins that solely observed in Southeast Asia (Table 2). The majority of these unique mutations (n=15) have arisen once to-date. The remaining 6 were present more than once, but each of these variants circulated in a specific country or region. Moreover, we found 13 mutations with amino acid replacement in spike protein across Southeast Asia. Seven of them (L54F, T76I, S116C, A243S, E471Q, T572I and D614G) were present in the S1 domain and the remaining 6 (L822F, A829T, A930V, S939F, F1109L and G1124V) were present in the S2 domain of the spike protein. Only one aa substitution (E471Q) occurred in the receptor binding motif of spike RBD. A 3D structural visualization is presented in Figure 1, with 13 aa substitution sites represented; of them, 3 aa substitutions (S116C, E471Q and A930V) are highlighted in the trimeric spike glycoprotein.

The recurrent mutations found in Southeast Asia were presented country wise in figure 2. We identified the 10 most frequent mutations with amino acid substitution (N_R203K, N_G204R, N_P13L, NS3_Q57H, NS3_-Q57H, NS8_L84S, NSP12_A97V, NSP12_P323L, NSP3_T1198K, NSP6_L37F, Spike_D614G) and separated the variants into 4 major groups and 2 subgroups accordingly (Figure 3). Group 1 consists of 11 sequences, including the reference sequence hCoV19 / Wuhan / WIV04 / 2019 (Accession: EPI_ISL_ 402124). All sequences within in this group were observed earlier in this year (13 January to 1 April). Group 2 consists of the co-evolving mutations NSP12_P323L and spike_D614G. Most of the sequences (n=24) belong to this group and their isolation dates range from 10 March to 4 May, 2020. We further separated this group into two subgroups; those in 2a have additional N_R203K, N_G204R (28881-28883: GGG>AAC) trinucleotide mutations and those in 2b have mutations with amino acid substitution at NS3 (Q57H). In group 3, 4 co-evolving mutations (NSP12_A97V, N_P13L, NSP3_T1198K, NSP6_L37F) were found in 12 Indian sequences which were collected between 29 March to 26 April, 2020. Group 4 consists of 8 variants from Bangladesh, India and Thailand with a common amino acid substitution at NS8(L84S).

In the cluster based time-plot, 187 available sequences were studied over the 15 days of the month from January until May 2020. The group 2 cluster was not observed until February (0%, n=0), 38% (n=19) in March, 37% (n=20) in April and 85% (n=61) in May. The group 3 cluster was 0% to 8% in those months, with a sudden increase of 54% (n=29) in April. The infections were increased from 17 cases in January up to 128,257 in May, 2020 (Figure 4).

A geographic heat map (Figure 5) revealed that most of the 78 NS mutations found in this study were also common in Europe and North America. The transmission map, generated with Nextstrain using 329 genome sequences (Figure 6), revealed that Group 2 sequences (A2 clade of Nextstrain) from this study were found to be dominant among strains circulating in Southeast Asia. These sequences were also found to be prevalent in Europe and North America.

Discussion:

The SARS-CoV-2 pandemic has cost more than 333 thousand lives already, with many more deaths being reported each day. During this period, the global community has yet to predict its virulence, seasonal variation, carriage and immunity. However, it is clear that the fatality rate varies by region and that the degree of virulence varies from person to person. Some regions in Europe and North America were affected

the most, while most of Asia, Africa and Australia remain less affected. A close analysis of this ssRNA genome has now become an elementary scientific need.

This study has characterized the SARS-CoV-2 virus circulating in Southeast Asia into 4 major groups and 2 sub-groups by studying common non-synonymous mutations. Group 1 consists of 5 out of 7 Indonesian sequences, 3 out of 8 sequences from Thailand and the only sequence from Nepal. Group 2 involves 40% of the variants in this study. Strains belong to this group coevolved with characteristic NS mutation, NSP12_-P323L and Spike_D614G. These variants were initially prevalent in Europe and North America, and now constitute 68% of the virus all over the world. A recent study analyzed 95 sequences and also found NSP12_-P323L variants to be at a higher frequency, and reported that this variant was mostly found outside of Wuhan, China (Khailany, Safdar, & Ozaslan, 2020). Another study suggests that RNA dependent RNA-polymerase (RdRp) aa substitution at the 323rd position (NSP12_P323L) causes RdRp fidelity, which, in turn, increases the number of mutations within the virus and causes co-evolved mutations (Pachetti et al., 2020). NSP12_P323L was co-evolved with Spike_D614G; this particular non-silent spike protein mutation generates an additional elastase cleavage site near the S1-S2 junction and thus facilitates fusion and cell entry (Koyama, Weeraratne, Snowdon, & Parida, 2020). This variant (Spike_D614G) was first observed in January 28, 2020 and was initially prevalent in Europe. Within 4 months, this variant has now rapidly outcompeted its ancestral subtype all over the world (Bhattacharyya et al., 2020). This explains the frequency of Group 2 variants in Southeast Asia and why these variants have subdivided into additional sub-groups involving co-evolving mutations.

We differentiated Group 2 into 2 subgroups, 2a and 2b, which involve N_203-204: RG> KR and NS3_Q57H amino acid substitutions, respectively, along with NSP12_P323L and Spike_D614G. Several studies (Ayub, 2020; Lorusso et al., 2020; Yin, 2020) mention trinucleotide block mutations in nucleotides (28881-28883: GGG>AAC) which resulted in 2 amino acid changes (N_203-204: RG> KR) and affected the Serine-Arginine-rich motif of N protein. This trinucleotide block mutations were found in 8 sequences, 3 of them were from Dhaka, Bangladesh. NS3_Q57H mutation variants have been commonly found in the USA (Mercatelli & Giorgi, 2020) and Europe and are predicted to be deleterious (Issa, Merhi, Panossian, Salloum, & Tokajian, 2020).

Unlike the others, Group 3 was unique, with 4 coevolving mutations. Of these, the NSP6_L37F mutation variant was common (Mercatelli & Giorgi, 2020); this mutation variant has also been frequently found in the UK, USA, Australia and India. The other 3 mutations are relatively less common and found mostly in India and Australia. Group 4, on the other hand, consists of a characteristic NS8_L84S mutation variant, which was declared as S type by Tang, X.L. et al. (Tang et al., 2020). This mutation was later reported as C type by another group (Forster, Forster, Renfrew, & Forster, 2020) and were clustered as S clade by GISAID (Fuertes et al., 2020). Group 4 included 4 Bangladeshi variants, isolated from the Chittagong district in May, 2020, along with 3 strain sequences out of 8 from Thailand and only one strain sequence from India.

A recent study conducted with 10,014 sequences identified 13 frequent non-synonymous mutations (Mercatelli & Giorgi, 2020), while we found only 7 of them, along with 3 less common mutation, at high frequency in this region (Figure 2). Most of the spike protein mutations identified in this region were also observed in Europe and North America. Spike protein mutations with aa substitution at 614 position, found in 40% of the studied strains in this region, were also prevalent in Europe and North America. On the contrary, the amino acid substitutions found at the 1109th position of the spike protein found in one Bangladeshi strain was found to be globally common with one strain from Switzerland. We observed another amino acid substitution in the spike protein at the 76th (Spike_T76I) position in an Indonesian strain, which was also found in two strains from West Bengal, India (data non shown). This specific amino acid substitution was identified on 55 occasions according to the global database. Among them, 49 were from Australia, suggesting that this variant might have transmitted from Australia. Another spike protein amino acid substitution (Spike_E471Q) was found in the receptor binding domain of the spike protein. Glutamate (E) was replaced by Glutamine (Q) resulting in a conservative replacement that may not contribute largely in binding to the ACE2 receptor.

Additionally, global mutation distribution statistics showed that Spike_A829T mutation was observed in 31 sequences, all of them from Thailand (Table 2). NSP2_I120F mutation was found in 9 of the 12 cases from Dhaka, Bangladesh and NSP2_D92G mutation was present in 4 out of the 5 sequences from Chittagong, Bangladesh (Data not shown). These cities are separated by a distance of 250 kilometers, suggesting that those viruses carrying novel mutations were circulating in an area-specific manner.

NS mutation and phylogenetic analysis conducted through the Nextstrain database was particularly useful in getting a closer look at mutation variants and their possible routes of transmission. We found a common N_203-204: RG> KR amino acid substitution (9 out of 12 strains) in Dhaka, Bangladesh. However instead of the common N_203-204: RG> KR amino acid substitution, a less common aa substitution was observed at the 202nd position of N protein (N_S202N) among the most (5 out of 7) strains of Chittagong. The mutation distribution database showed that strains having trinucleotide block mutation in N protein were prevalent in Europe and that the N_S202N mutant was found more commonly in recent strains of Saudi Arabia. Phylogenetic analysis by Nextstrain also revealed that the Chittagong strains (belongs to Nextstrain B4 clade and group 4 of our study) have close relationship with the Saudi Arabian strains, while Dhaka strains (A2 clade in nextstrain, group 2 in this study) are similar to the European ones.

The geographical heat-map (Figure 5) of these non-synonymous mutations indicate that most of these mutations were also frequently found in the UK, USA, Australia, Saudi Arabia and other European countries, revealing possible transmission routes to Southeast Asia. Phylogenetic analysis with 329 genomes from this region by Nextstrain produced a similar transmission route map (Figure 6). This study also confirmed, through phylogenetic and mutation analysis, that a high percentage of Group 2 strains are linked to European and North American strains (A2 clade in Nextstrain analysis) in India and Bangladesh.

We could not analyze the strains from Maldives, Bhutan and Timor-Leste because they do not have whole genome sequence data of the virus at the time of our analysis. Among the six countries with available genome sequences of a good quality, only India, Bangladesh and Indonesia have reported a higher number of SARS-CoV-2 infections. The frequencies of infection have increased exponentially from mid-April, 2020. In our study, we additionally analyzed 187 sequences (Figure 4) of which 100 (53%) sequences from India, Bangladesh, Thailand, Indonesia and Srilanka showed characteristic NSP12_P323L and spike_D614G mutations, which put them in the Group 2 cluster (Nextstrain clade A2). It was also shown that Group 2 variants were not found earlier than the 10^{th} of March in this study. The time plot data delineates that this Group 2 cluster is emerging rapidly from 0% in January and February, to 85% in May 2020. In contrast, group 1 strains (similar to the ancestral strain) were not found after the 1^{st} of April, suggesting that the European and North American strains are the most recent predominant strains in this subcontinent. A study conducted in early March reported that NSP12_P323L (14408C>T) and spike_D614G (23403A>G) mutations were recurrent in Europe and had not been detected in Asia until then, supporting our statement (Pachetti et al., 2020). Along with other co-evolving mutations, NSP12_P323L and Spike_D614G probably provide variants with an evolutionary advantage over their ancestral types, allowing them to survive and circulate in this densely populated region.

Although a number of earlier studies hypothesized that high temperatures and high humidity could result in reduced SARS-CoV-2 transmission, the infection rate of SARS-CoV-2 is already increasing in this subcontinent. Given that the European and North American variants (Nextstrain clade A2) are emerging rapidly and that winter is approaching, the next wave of SARS-CoV-2 may take place in Southeast Asia.

Data Availability

Complete and near-complete genome sequences of SARS-CoV-2 are available in the GISAID database. The accession numbers of the genome sequences are available in supplementary table-1.

Acknowledgment:

We would like to acknowledge Dr. Mohammad Mahfuzur Rahman, Chairman, Dept. of Climate and Disaster Management, Jashore University of Science and Technology for his valuable suggestion on graphical

presentation.|

Ethics statement

'The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to. No ethical approval was required as this study didn't collect any samples or questionnaires from animals or humans.

Conflicts of interests:

All authors declared that they do not have any conflicts of interests or relationship with this study financially or otherwise.

References:

Ayub, M. I. (2020). Reporting Two SARS-CoV-2 Strains Based on A Unique Trinucleotide-Bloc Mutation and Their Potential Pathogenic Difference. Published online April, 19.

Bhattacharyya, C., Das, C., Ghosh, A., Singh, A. K., Mukherjee, S., Majumder, P. P., . . . Biswas, N. K. (2020). Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. bioRxiv.

Demongeot, J., Flet-Berliac, Y., & Seligmann, H. (2020). Temperature Decreases Spread Parameters of the New Covid-19 Case Dynamics. Biology, 9(5), 94.

Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences, 117(17), 9241-9243.

Fuertes, F. D., Caballero, M. I., Monzón, S., Jiménez, P., Varona, S., Cuesta, I., . . . Pérez, J. G. (2020). Phylodynamics of SARS-CoV-2 transmission in Spain. bioRxiv.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., . . . Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), 4121-4123.

Issa, E., Merhi, G., Panossian, B., Salloum, T., & Tokajian, S. T. (2020). SARS-CoV-2 and ORF3a: Non-Synonymous Mutations and Polyproline Regions. bioRxiv.

Katoh, K., Misawa, K., Kuma, K. i., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic acids research, 30(14), 3059-3066.

Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. Gene reports, 100682.

Koyama, T., Weeraratne, D., Snowdon, J. L., & Parida, L. (2020). Emergence of drift variants that may affect COVID-19 vaccine development and antibody treatment. Pathogens, 9(5), 324.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution, 35(6), 1547-1549.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research, 47(W1), W256-W259.

Lorusso, A., Calistri, P., Mercante, M. T., Monaco, F., Portanti, O., Marcacci, M., . . . Di Pasquale, A. (2020). A "One-Health" approach for diagnosis and molecular characterization of SARS-CoV-2 in Italy. One Health, 100135.

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations.

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., . . . Gallo, R. C. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine, 18, 1-9.

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). Temperature and latitude analysis to predict potential spread and seasonality for COVID-19. Available at SSRN 3550308.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., . . . Qian, Z. (2020). On the origin and continuing evolution of SARS-CoV-2. National Science Review.

Wang, J., Tang, K., Feng, K., & Lv, W. (2020). High temperature and high humidity reduce the transmission of COVID-19. Available at SSRN 3551767.

WHO, Corona virus disease (COVID-19) situation report-124, published on 23rd May, 2020 (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200523-covid-19-sitrep-124.pdf?sfvrsn=9626d639_2)

Yin, C. (2020). Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics.

	India	Bangladesh	Indonesia	Thailand	Sri Lanka	Nepal
Total Covid-19 Cases	125101	30205	20796	3040	1068	548
Cases/ million people	90	184	76	44	50	18
Total Deaths	3720	432	1326	56	9	3
Collection dates	31 Jan-6 May	18 Apr-13 May	17 Mar-14 Apr	22 Jan-3 Apr	4-31 Mar	13 Jan
Sequences for study	30	10	8	7	4	1
Non Synonymous (N-S) Mutations	39	22	12	8	13	0
Unique N-S mutations	11	4	2	3	1	0
Spike protein N-S mutations	8	2	4	2	1	0
E protein N-S mutation	1	0	0	0	0	0
M protein N-S mutations	1	0	0	1	1	0
N protein N-S mutations	4	6	1	2	3	0
NS3 N-S mutations	1	3	1	1	2	0
NS7b N-S mutations	0	0	0	1	0	0
NS8 N-S mutations	1	2	0	1	0	0
NSP2 N-S mutations	8	1	0	1	1	0
NSP3 N-S mutations	5	2	1	2	1	0
NSP4 N-S mutations	2	0	0	0	0	0
NSP5 N-S mutations	0	2	0	0	0	0
NSP6 N-S mutations	1	0	1	0	1	0
NSP8 N-S mutations	0	1	0	0	0	0
NSP12 N-S mutations	3	1	4	1	1	0
NSP13 N-S mutations	2	1	0	0	0	0
NSP14 N-S mutations	2	1	0	1	1	0
NSP15 N-S mutations	0	0	0	1	1	0

Table 1: The frequency of SARS-CoV-2 cases identified in Southeast Asia region by country and their possessed mutations among selected 60 strains. Case reports were until 23rd May.

Mutation sites	Country	No. of virus (study)	No. of virus (total)
E_L65M	India	1	1
M_L29J	India	1	1
N_K347N	Indonesia	2	2
NS3_I263M	Srilanka	1	1
NSP2_G212D	Thailand	1	1
NSP2_I120F	Bangladesh	4	9
NSP2_L266I	India	1	1

NSP3_L1553J	India	1	1
NSP3_N1337S	Bangladesh	1	1
NSP3_S1485Y	India	1	1
NSP5_D92G	Bangladesh	3	4
NSP12_V880I	India	1	4
NSP13_K40R	India	1	2
NSP13_T214I	India	1	1
NSP14_D415G	Thailand	1	1
NSP14_S434N	India	1	1
NSP14_V459I	Bangladesh	1	1
Spike_A829T	Thailand	3	31
Spike_A930V	India	1	1
Spike_E471Q	India	1	1
Spike_S116C	Indonesia	1	1

Table 2: Unique mutations with amino acid substitutions observed solely in Southeast Asia until 23rd of May, 2020.

Figure legends

Figure 1: mutation sites with amino acid substitution in spike protein found in Southeast Asia

Figure 2: The 10 most recurrent mutations identified from the randomly selected 60 strains in Southeast Asia with percentage of mutations compared with world-wide frequencies.

Figure 3: Phylogenetic relations of selected 60 SARS-CoV-2 strain sequences separated into four clusters

Figure 4: The cluster based time-plot of 187 complete genome sequences available in GISAID from January to May 2020. Bar charts indicate the frequency of mutants identified in first 15 days of the respective month. Line graph indicates the number of COVID-19 infections of the respective whole month.

Figure 5: A total of 78 non-synonymous mutations found in 60 sequences in Southeast Asia. Red marks in the heat map showing the countries having these mutations in high frequencies.

Figure 6: Transmission map constructed by Nextstrain with 329 sequences showing possible transmission routes into 6 Southeast Asian countries. This map is also separated the sequences of different clades.

Appendices

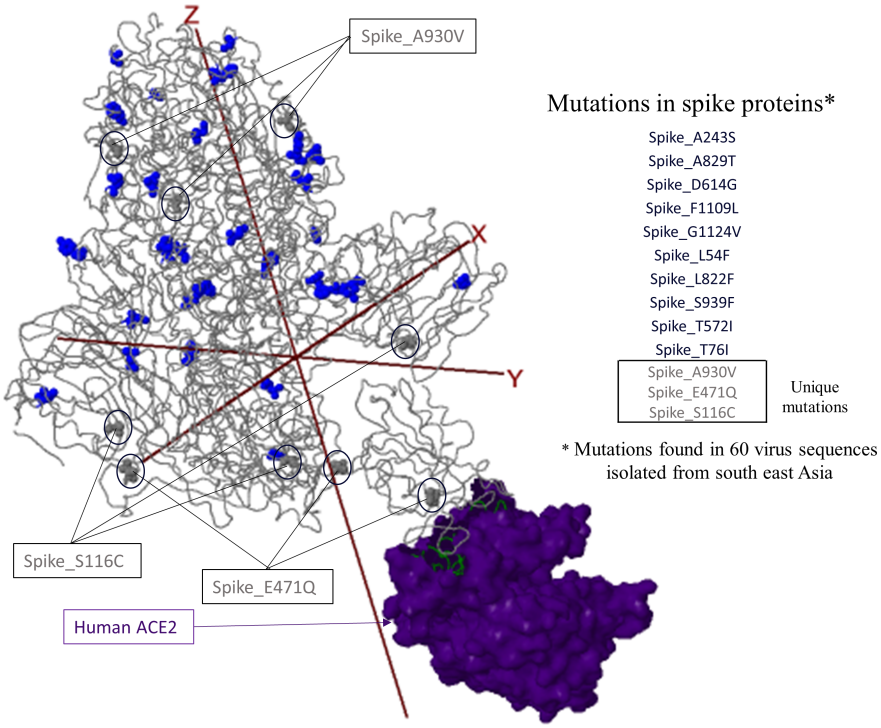
Supplementary Table -1: Accession Number and location of studied sequence.

Hosted file

`table 1.docx` available at https://authorea.com/users/325049/articles/456169-emergence-of-european-and-north-american-mutant-variants-of-sars-cov-2-in-southeast-asia

Hosted file

`table 2.docx` available at https://authorea.com/users/325049/articles/456169-emergence-of-european-and-north-american-mutant-variants-of-sars-cov-2-in-southeast-asia



Mutations in spike proteins*

Spike_A243S
Spike_A829T
Spike_D614G
Spike_F1109L
Spike_G1124V
Spike_L54F
Spike_L822F
Spike_S939F
Spike_T572I
Spike_T76I

Spike_A930V
Spike_E471Q
Spike_S116C

Unique mutations

* Mutations found in 60 virus sequences isolated from south east Asia