Original Article

Microbial Biosystems 5(2) (2020) 2020.1026
10.21608/mb.2020.46676.1026

Contents lists available at Egyptian Knowledge Bank
Microbial Biosystems
Journal homepage: http://mb.journals.ekb.eg/

MICROBIAL BIOSYSTEMS
INTERNATIONAL SCIENTIFIC JOURNAL OF MICROBIAL BIOLOGY

SARS-CoV-2 mutation hotspots incidence in different geographic regions

Jivan Q. Ahmed, Gahin A. Tayib, Teroj A. Mohamed*

Pathology and Microbiology Department, College of Veterinary Medicine, University of Duhok, Iraq.

ARTICLE INFO

Article history
Received 16 October 2020
Received revised 24 October 2020
Accepted 25 October 2020
Available online 3 December 2020

Corresponding Editor:
Tariq FJ
Abdel-Wareth MT

Keywords
SARS-CoV-2
Mutations
Covid-19
Evolution

ABSTRACT

SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) is RNA virus with a positive-sense single-strand that belongs to the beta-coronavirus group that causes COVID-19 (Coronavirus Disease 2019) which originally emerged in China. Viruses with RNA genomes are known by a high mutation rate potential. The mutation rate determines genome variability and evolution of the virus; therefore, allowing viruses to evade the immune system, gain more infectivity potentials, virulence modifications, and probably resistance development to antivirals. A total of 311 SARS-CoV-2 virus whole genome sequences have been retrieved from the GISAID database from 1st of January 2020 to 31th of August 2020. The sequences were analyzed for sequence purity and multiple sequence alignment together with reference sequence was conducted through using Clustal Omega that is imbedded in Jalview software and Blast tools. We recorded the occurrence of 4 newly incident high frequently occurring mutations in all six geographic regions, namely at positions 2416, 18877, 23401, and 27964. The majority of all recorded hotspots were detected in Asia, Europe, and North America. The findings of our study suggest that the SARS-CoV-2 is in continuous evolution. For the impact of these mutations, further investigations are required and to understand whether these mutations would lead to the appearance of Drug-resistance viral strains, strains with increased infectivity and pathogenicity, and also their effect on the vaccine development and immunogenesis.

Published by Arab Society for Fungal Conservation

Introduction

The recent emergence of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) that results in COVID-19 (Coronavirus Disease 2019) was initially reported for the first time in late December of 2019 from Wuhan, China. The virus is a positive-sense single-stranded RNA of the beta-coronavirus group and genome length is 29,903 nucleotides (Wu et al. 2020).

Consequently, due to the rapid spread of the SARS-CoV-2 on both national and international levels, the World Health Organization (WHO) declared a state of health emergency around the globe by late January 2020. Subsequently, the WHO declared the pandemic of the SARS-CoV-2 outbreak by March 11th, 2020. In a couple of weeks, the outbreak resulted in the thousand fatalities

*** Corresponding author**
E-mail address: teroj.mohamed@uod.ac (Teroj Abdulrahman Mohamed)

worldwide that heavily affected the globe in most sectors and strongly hit the global economy. The virus has spread to more than 210 countries and territories in around 4 months and has reached to over 34 million reported infections and more than 1 million fatalities around the globe as of October 5th, 2020 (World Health Organization). A variation in death rates in different countries could be due to demographic divergence and the different control measures implemented in many countries (Dowd et al. 2020).

Variations in infection rates worldwide could be attributed to many reasons: several strategies implemented to limit the public movements, quarantine, and diversity of the population genetics as well as herd immunity. Nevertheless, the reasons behind a wide difference in mortality rate from region to region due to SARS-CoV-2 are not yet fully understood, but monitoring of genetic mutations of the virus, as well as viral abilities of evolution might be significant.

The incidence of mutations and its rate in the genome of RNA viruses is relatively greater than that of their targeted host by nearly a million times. This huge rate is associated with viral virulence variation and evolution capabilities, feature considered helpful for a virus to adjust to its targeted host (Duffy, 2018). Studies have been recently conducted to characterize the mutations incidence in the genome of SARS-CoV-2 in 18 spots that were frequently detected on ORF1ab, S, ORF3a, ORF8 and N genome regions, located at positions 1397, 2891, 3036, 8782, 11083, 14408, 17746, 17857, 18060, 23403, 26143, 28144 and 28881 (Wrapp et al. 2019; Khailany et al. 2020; Pachetti et al. 2020; Wang et al. 2020) . Furthermore, these studies have demonstrated that SARS-CoV-2 is quickly spreading among countries and a virus is acquiring fresh mutation hotspots (Khailany et al. 2020; Mercatelli and Giorgi, 2020; Pachetti et al. 2020; Tang et al. 2020; Wang et al. 2020). Nevertheless, as more whole-genome sequences are becoming available, the necessity of defining a more precise geographical circulation of different variants has become essential for defining medical and political approaches. Regardless of many studies suggesting fairly low differences in SARS-CoV-2 genome sequences (Ceraolo and Giorgi, 2020; Lu et al. 2020), has to be investigated whether the observed variation in mortality rate and quick spread documented in different countries might be results of mutations in the SARS-CoV-2 genome sequences (Brufsky, 2020). The influence of SARS-CoV-2 mutation hotspot upsurge on immunogenesis and the predictions for vaccine development could be significant and requires more investigations and monitoring (Jackson et al. 2020).

This study focused on mutations of SARS-CoV-2 to assess whether the newly developed mutations are circulating among countries. SARS-CoV-2 genome characterization of mutations hotspots could be helpful for developing antivirals and vaccines, understanding the effect of these mutations on the infectivity and pathogenesis of the virus as well as its advantages for diagnostic techniques.

Materials & Methods

A total of 311 SARS-CoV-2 virus whole genome sequences have been retrieved from the GISAID database (https://www.gisai d.org/) from 1st of January 2020 to 31st of August 2020. A formerly known as Wuhan seafood market pneumonia virus (WSM) that was submitted on the Genbank data base with accession number NC_045512 (Wu et al. 2020) has been considered as a reference sequence for analysis. Complete genome sequences of the SARS-CoV-2 virus were collected from 44 different countries (i.e. China, South Korea, Indonesia, Taiwan, Japan, Hong Kong, Saudi Arabia, Singapore, Vietnam, Malaysia, Germany, England, Italy, Sweden, Spain, Switzerland, Ireland, Belgium, Norway, Portugal, Russia, Nigeria, South Africa, Morocco, Benin, Senegal, Madagascar, Zambia, Sierra Leone, Kenya, Egypt, USA, Costa Rica, Canada, Mexico, Australia, New Zealand, Brazil, Peru, Colombia, Venezuela, Ecuador, Chile, Argentina) (table 1). The collected sequences were clustered into six geographic areas namely Asia, North America, Europe, Oceania, Africa, and South America as done earlier by the GISAID consortium. Complete genome sequences (28000–30,000 bps) were taken for analysis.

Retrieved sequences were analysed for sequence purity and multiple sequence alignment together with reference sequence was conducted through using the Clustal Omega which is embedded in Jalview software along with Blast tools. Jalview was used to identify the mutation spots when comparing With the WSM reference sequence. Furthermore, the aligned sequences were translated into amino acid by using Jalview to determine the potential amino acid substitutions of a given mutation spot. The statistical analysis and graphics were performed using prism software (Version 8.4.3).

Results

A total of 311 complete genome sequences of SARS-CoV-2 were randomly recovered from the GISAID consortium website and aligned together with the reference genome sequence. The dataset was divided into 6 geographic regions: Asia, North America, Europe, Africa, South America, and Oceania (Fig. 1). Genome sequence alignment was performed for each region along with the reference genome sequence.

The mutation distribution of SARS-CoV-2 was calculated from January to August in different geographic regions (Fig. 2), calculation of the mutation frequency in the six mentioned regions, by counting a total number of a given mutation in all analysed genome sequences for each geographic region table 1. This study evaluated mutations of 10< time-frequency in all analyzed genome sequences of our data sets, mutations of less than 10-time incidence were not reported. Here we confirm the recurrent mutations that have been recorded earlier at positions 1059, 3037, 8782, 11083, 14408, 20268, 22444, 23403, 25563, 26735, 26144, 28144, 28854, 28881, 28882, and 28882 (Khailany et al. 2020; Mercatelli and Giorgi, 2020; Pachetti et al. 2020; Tang et al. 2020; Weber et al. 2020). Furthermore, we recorded the occurrence of 4 newly incident high frequently occurring mutations in all six geographic regions, namely at positions 2416, 18877, 23401, and 27964 (Table 1). The identified mutations were found on Open reading frame 1ab (ORF1ab) gene (nsp2 1059nt, nsp2 2416nt, nsp3 3037nt, nsp4 8782nt, nsp6 11083nt, RNA-dependent RNA polymerase (RdRp) 14408nt, nsp14 18877nt, nsp15 20268nt), S gene (Spike protein 23401nt, Spike protein 23403nt), ORF3a gene (ORF3a protein 25563nt, ORF3a protein 26144nt), M gene (Membrane Glycoprotein 26735nt), ORF8 gene (ORF8 protein 27964nt, ORF8 protein 28144nt), and N gene (Nucleocapsid Phosphoprotein 28854nt, Nucleocapsid Phosphoprotein 28881nt, Nucleocapsid Phosphoprotein 28882 nt, Nucleocapsid Phosphoprotein 28883nt) .

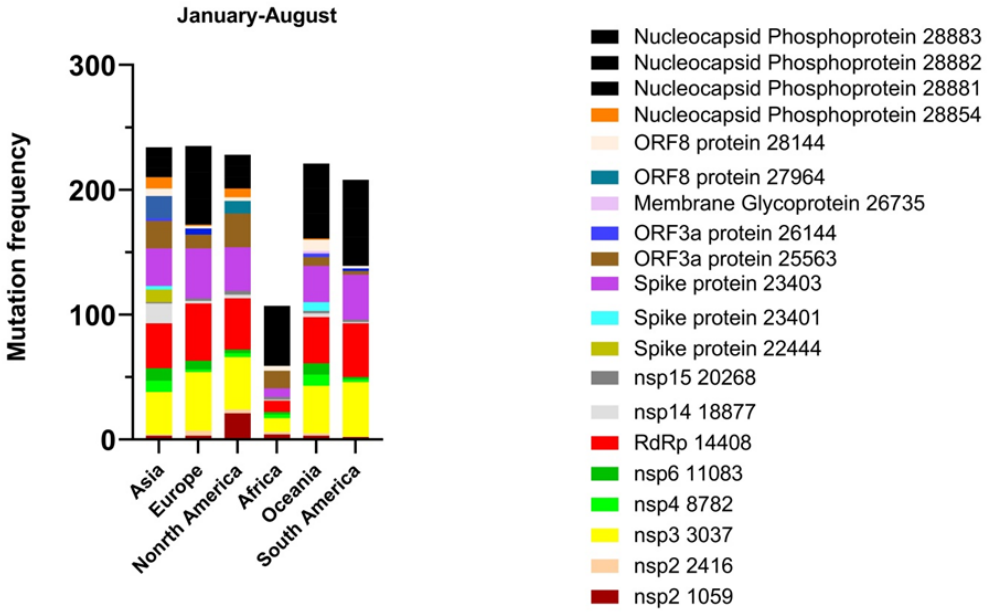

Fig 1. SARS-CoV-2 mutation in six geographic regions with the frequency of occurrence from January to August 2020. Ten novel recurring hotspots mutations in positions: 1059, 3037, 8782, 11083, 14408, 20268, 23403, 26144, 28144 and 28881 and ten novel hots.

The results revealed that 4 out of 20 most frequent mutations at positions 2416, 3037, 8782, and 20268 appeared to be silent. Conversely, the rest of the mutations spots lead to amino acid substitution at positions: 1059, 2416, 11083, 14408, 18877, 22444, 23401, 23403, 25563, 26144, 26735, 27964, 28144, and 28854, the corresponding amino acid changes are: P>S, C>F, P>L, S>F, P>S, G>W, D>G, Q>H, G>V, T>I, H>Y, Y>H, and S>L, respectively (Table 1). The mutations on a triple nucleotide 28881, 28882, and 28883 that are adjacent to each other and are associated with a double codon change, in a way changing (R to K) and (G to R) (Table 1). It's observed that the number of each mutation and frequency of incidences in all eight months is increasing particularly starting from March. Furthermore, the frequency of recorded mutations is highly incident in Asia, North America, Europe, South America, and Oceania comparing to lower mutation frequency in Africa. Moreover, mutations on positions 3036, RdRp 14408 and 23403 were among those of highly incident spots on the SARS-Cov-2 genome, some other mutations were frequently detected in the early months and started regressing in the last months of this study and even some were not detected this was noted at positions: 1059, 8782, 23401, 23403, 26144, 28144. Additionally, triple mutation

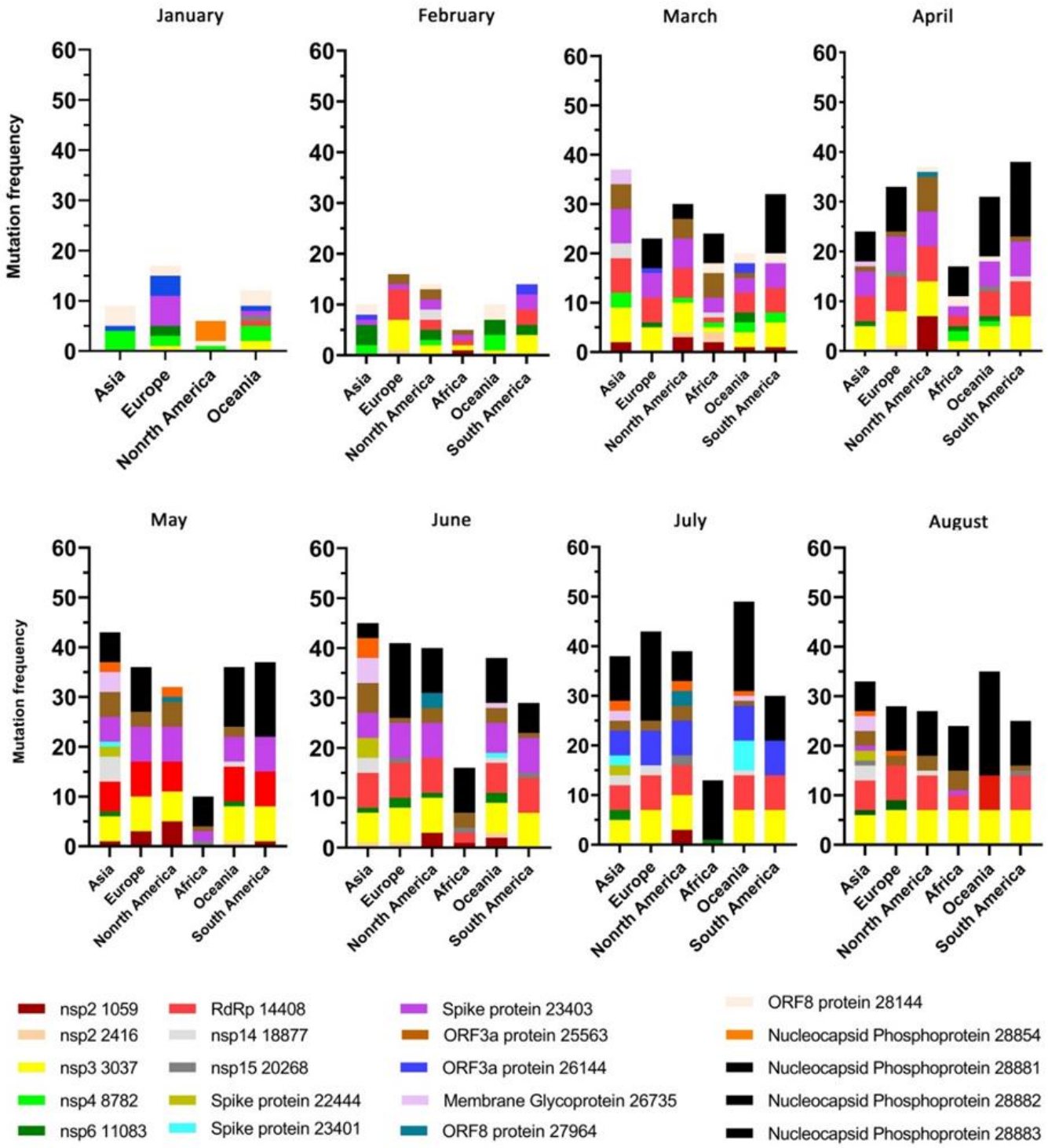

Fig 2. Illustration of SARS-CoV-2 mutations in six geographic regions with the frequency of occurrence through January to August 2020.

at positions 28881, 28882, and 28883 (black) was occurred in March and increased through other subsequent months. Regarding the base changes, the highest recorded base substitution is C>T.

Discussion

This study was conducted to compare submitted SARS-CoV-2 whole genome sequences on the GISAID database with the reference genome for the purpose of gaining significant vision on mutation hotspot, their incidence within different geographical regions over eight months post-emergence of the virus.

We recorded the incidence of 4 more high frequently occurring mutations in six geographic regions, namely at positions 2416, 18877, 23401, and 27964 beside confirming earlier reported mutations at positions 1059, 3037, 8782, 11083, 14408, 20268, 22444, 23403, 25563, 26735, 26144, 28144, 28854, 28881, 28882, and 28882 (Khailany et al. 2020 ; Mercatelli and Giorgi, 2020; Pachetti et al. 2020; Tang et al. 2020; Weber et al. 2020) (Table 1).

Table 1 Mutation hot spots identified in SARS-CoV-2 genomes sequences

Nucleotide position	Gene	Mutation	Amino acid substitution	Frequency of mutation	References
NSP2 (NT 1059)	ORF1ab	C>T	P>S	37	Mercatelli and Giorgi (2020)
NSP2 (NT2416)	ORF1ab	C>T	Silent	12	This study
NSP3 (NT3037)	ORF1ab	C>T	Silent	216	Mercatelli and Giorgi (2020)
NSP4 (NT8782)	ORF1ab	C>T	Silent	28	Khailany et al. (2020), Mercatelli and Giorgi (2020), Tang et al.(2020)
NSP6 (NT11083)	ORF1ab	G>T	C>F OUT	33	Mercatelli and Giorgi (2020), Pachetti et al. (2020), Tang et al., (2020)
NSP12 (RDRP, NT14408)	ORF1ab	C>T	P>L	175	Mercatelli and Giorgi (2020), Pachetti et al. (2020)
NSP14 (NT18877)	ORF1ab	C>T	S>F	26	This study
NSP15 (NT20268)	ORF1ab	A>G	Silent	12	Mercatelli and Giorgi (2020)
Spike protein (NT22444)	S	C>T	P>S	10	Weber et al. (2020)
Spike protein (NT23401)	S	G>T	G>W	10	This study
Spike protein (NT23403)	S	A>G	D>G	177	Mercatelli and Giorgi (2020), Pachetti et al. (2020), Weber et al. (2020)
ORF3A protein (NT25563)	ORF3a	G>T	Q>H	84	Weber et al. (2020)
ORF3A protein (NT26144)	ORF3a	G>T	G>V	12	Mercatelli and Giorgi (2020)
Membrane Glycoprotein (NT26735)	M	C>T	T>I	20	Weber et al. (2020)
ORF8 protein (NT27964)	ORF8	C-T	H>Y	10	This study
ORF8 protein (NT28144)	ORF8	T>C	Y>H	26	Mercatelli and Giorgi (2020), Pachetti et al. (2020), Tang et al.. (2020)
Nucleocapsid phosphoprotein (NT28854)	N	C>T	S>L	18	Weber et al. (2020)
Nucleocapsid Phosphoprotein (NT28881)	N	G>A	R>K G>R	97	Pachetti et al. (2020)
Nucleocapsid Phosphoprotein (NT28882)	N	G>A		97	Weber et al. (2020)
Nucleocapsid Phosphoprotein (NT28883)	N	G>C		97	Weber et al. (2020)

Table 2 Mutation frequency incidence and percentage in each geographic region

Region	Total Mutation detected	Asia	Europe	North America	Africa	Oceania	South America
nsp2 (nt 1059)	36	8.3	8.3	58.3	11.1	8.3	5.5
nsp2 (nt2416)	12	8.3	33.3	25	16.6	16.6	0
nsp3 (nt3037)	216	15.7	21.7	19.4	5.0	17.5	20.3
nsp4 (nt8782)	28	32.1	7.1	10.7	10.7	32.1	7.1
nsp6 (nt11083)	33	30.3	21.2	9.	6.0	27.2	6.
nsp12 (RdRp, nt14408)	212	16.9	21.6	19.3	4.2	17.4	20.2
nsp14 (nt18877)	26	61.5	7.6	11.5	3.8	11.5	3.8
nsp15 (nt20268)	12	8.3	16.6	25	16.6	16.6	16.6
Spike protein (nt22444)	10	100	0	0	0	0	0
Spike protein (nt23401)	10	30	0	0	0	70	0
Spike protein (nt23403)	177	16.9	22	19.7	3.9	16.3	20.3
ORF3a protein (nt25563)	84	26.1	13	32	16.6	8.3	3.5
ORF3a protein (nt26144)	12	16.6	41.6	0	0	25	16.6
Membrane Glycoprotein (nt26735)	20	90	0	0	0	10	0
ORF8 protein (nt27964)	10	0	0	100	0	0	0
ORF8 protein (nt28144)	26	23	7.6	11.5	15.3	34.6	7.6
Nucleocapsid Phosphoprotein (nt28854)	18	50	5.5	38.8	0	5.5	0
Nucleocapsid Phosphoprotein (nt28881)	97	8.2	21.6	9.2	16.4	20.6	23.7
Nucleocapsid Phosphoprotein (nt28882)	97	8.2	21.6	9.2	16.4	20.6	23.7
Nucleocapsid Phosphoprotein (nt28883)	97	8.2	21.6	9.2	16.4	20.6	23.7

The newly detected hotspots were unevenly identified in 6 geographic regions. Mutation at position 2416 was detected in Europe (33%) higher than the other regions except for South America where it has not been reported in table 2. Mutation 18877 was detected in all study regions with the majority of incidence in Asia (61.5%). Furthermore, the hotspot at position 23401 was only identified in Asia (30%) and Oceania (70%) regions. Interestingly, mutation at position 27964 was only seen in genome sequences from Asia (100%) (Table 2).

We noted that the bulk of mutation hotspots ascended following SARS-CoV-2 spread to countries outside China and have been given a chance to replicate in diverse environments and populations confirming the earlier suggestions (Weber et al. 2020) (Fig. 1).

In January there were a few pointed mutations in all regions except Africa and South America that there have been no genome submissions. The majority of these mutations were occurred in Europe than other regions (in position 3036, 8287, 23403, 26143, 26143, 28144 and 28854). Afterward, the mutation rate started to gradually increase starting from February onward as it has been connected to a mutation in RdRp at position 14408 by (Pachetti et al. 2020) however, we have detected this mutation in January in the Oceania region as well (Fig 2). This is because RdRp is involved in genome proofreading during the replication process as this point mutation might have affected its proofreading abilities.

Some frequently incident mutations were only detected in one region like positions 22444 and 27964 that were only detected in Asia and North America, respectively figure 1 and 2. By looking at figure 2 some mutations started to reappear and some others has diminished like the incidence of mutation at 28881-28883 in March and subsequent increase in the following months, and mutations at positions 1059, 8782, 23401, 26144, 27964, and 28144 have not been detected in August. Conversely, mutations at positions 3036 and 14408 remained consistent. In general, the rate of mutation frequency for detected spots was higher in Europe (19%) comparing to the lowest rate in Africa (8%), Asia and North America both were (18%) and (17%) respectively while was (16%) for either Oceania and South America respectively (Fig.1).

In this study, we have detected three mutations on S gene at positions 22444, 23401, and 23403. The importance of mutations on S gene and corresponding amino acid substitution were investigated and suggested in increase infectivity for variants containing D614G on Spike protein and amino acid substitution on S1 subunit of S protein which is a receptor-binding domain (Wrapp et al. 2019), resulted in decreasing infectivity (Li et al. 2020).

Therefore, further studies are essential to evaluate the impact of these mutations on vaccine development and cross reaction of developed immunity against newly detected variants.

The mutation of RdRp at position 14408 that has been recorded earlier (Pachetti et al. 2020), suggesting that the increased mutation rate from February and afterward was due to this alteration in the RdRp, with keeping its original catalytic action, that may have adjusted its binding abilities with some other cofactors like Non-structural protein (nsp)7, nsp8 or Exonuclease (ExoN) , thus altered the mutation rate.

On the other hand, there have been mutations detected on N gene that encodes for Nucleocapsid Phosphoprotein at positions 28881-28883 that resulted in the amino acid substitution of R>K and G>R table 1. This mutation was detected for the first time in North America, Europe, and South America in March and subsequently detected in other regions (Fig. 2), the incidence of this mutation in Europe (21%), and Oceania (20%) was higher compared to other regions (Fig. 1) and (Table 1). This mutation raises the demand of focus on the potential generation of the nuclear localization signal (Kalderon, 1984) and there is a lack of understanding of the activity of solid DNA-binding motifs of Nucleocapsid Phosphoprotein in the nuclei of an infected cell. A study has reported evidence of this protein localization in the nuclei of the infected cells (Cawood et al. 2007). In the meantime, it is not understood whether SARS-CoV-2 carrying this mutant would possibly modify the pathogenicity and biology of the virus, and does RNA-RNA recombination has a role in the mutation of three nucleotides at position 28881 to change GGG to AAC.

Conclusions

The findings of our study suggest that SARS-CoV-2 is a continuous evolution. Here we report 4 novel mutations hotspot in SARS-CoV-2 sequences in 6 geographic regions; North America, Europe, Asia, South America, Oceania, and Africa. The newly detected hotspots were unevenly identified in 6 geographic regions. Mutation at position 2416 was identified in all studied regions except Africa and spot 18877 was detected in all studied regions with the majority of incidences in Asia. Besides, the hotspot at position 23401 was only noticed in Asia and Oceania regions while mutation at position 27964 was only seen in genome sequences from Asia. The impact of these mutations requires further investigation and to understand whether these mutations of this study and previously identified hotspots would lead to the appearance of drug-resistance viral strains, strains with increased infectivity and pathogenicity, and also their effect on the vaccine development.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank all frontline medical workers working around the clock to fight this pandemic.

Funding

No funding was used to conduct this research.

References

Brufsky A (2020) Distinct viral clades of SARS-CoV-2: Implications for modeling of viral spread. J Med Virol 92(9): 1386-1390

Cawood R, Harrison SM, Dove BK, Reed ML, Hiscox JA (2007) Cell cycle dependent nucleolar localization of the coronavirus nucleocapsid protein. Cell Cycle. 6 (7): 863-867.

Ceraolo C, Giorgi FM (2020) Genomic variance of the 2019-nCoV coronavirus. J Med Virol 92(5):522-528.

Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, Liu Y, Mills MC (2020) Demographic science aids in understanding the spread and fatality rates of COVID-19. PNAS USA 117(18):9696-9698.

Duffy S (2018) Why are RNA virus mutation rates so damn high?. PLOS Biol 16(8):3000003.

Jackson, L.A., Anderson, E.J., Rouphael, N.G., Roberts, P.C., Makhene, M., Coler, R.N., McCullough, M.P., Chappell, J.D., Denison, M.R., Stevens, L.J. and Pruijssers, A.J (2020) An mRNA vaccine against SARS-CoV-2—preliminary report. New England Journal of Medicine, 383(20): 920-1931.

Kalderon D (1984) Roberts BL, Richardson WD, Smith AE. A short amino acid sequence able to specify nuclear location. Cell 39: 499-509.

Khailany, R.A., Safdar, M. and Ozaslan, M. (2020) Genomic characterization of a novel SARS-CoV-2. Gene Rep.19:100682.

Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, Zhao C, Zhang Q, Liu H, Nie L, Qin H (2020) The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell, 182 (5):1284-1294.

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, 395(10224):565-574.

Mercatelli D, Giorgi F M (2020) Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Front. Microbiol 1:1800.

Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi, M, Gallo RC, Zella D (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J. Transl. Med 18:1-9.

Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J (2020) On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. 7(6): 12–1023. doi: 10.1093/nsr/nwaa036.

Wang C, Liu Z, Chen Z, Huang X, Xu M, He T, Zhang Z (2020) The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol 92(6): 667-674.

Weber S, Ramirez C, Doerfler W (2020) Signal hotspot mutations in SARS-CoV-2 genomes evolve as the virus spreads and actively replicates in different parts of the world, Virus Res. 289: 198170.

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367(6483): 1260-1263.

Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML (2020) A new coronavirus associated with human respiratory disease in China. Nature, 579(7798):265-269.