IJPSR (2021), Volume 12, Issue 4 **(Research Article)**

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 30 December 2020; received in revised form, 25 February 2021; accepted, 16 March 2021; published 01 April 2021

GEOGRAPHICAL BASED GENOMIC AND PATHOGENIC STUDY OF THE SARS-COV-2

G. Dhanalakshmi * and Anadi

Department of Biochemistry, Padmashree Institute of Management and Sciences, Bangalore - 560060, Karnataka, India.

Keywords:

SARS-CoV-2, COVID 19, GC content, ORFs

Correspondence to Author:
G. Dhanalakshmi

Assistant Professor,
Department of Biochemistry,
Padmashree Institute of Management
and Sciences, Bangalore - 560060,
Karnataka, India.

E-mail: dhanu.bio@gmail.com

ABSTRACT: The current pandemic COVID 19 caused by SARS-CoV-2 is very contagious and causes respiratory disease. Genomic analysis of different strains in SARS-CoV-2 were found to infect different geographical locations of the world (India, USA, and Brazil) and also in different parts of India (Kerala, Surat, and Ahmedabad). These sequences were downloaded from the viral genome sequence database of the National Center for Biotechnology Information (NCBI), and comparative analysis is performed using various computational tools. Comparison of genome size, G+C content, and identified ORFs is performed for different sequences of SARS-CoV-2 isolated from different geographical locations of the world and also from different cities/states of India using various bioinformatics tools. This study concludes that all the sequences isolated from different geographical locations are highly similar to the reference strain of Wuhan, with an identity percentage ranging from 99.56-99.98% and also with the same GC content of 37%. The difference in the identified ORFs of length >150 and a small difference in the identity percentages shows the presence of mutation between different strains of SARS-CoV-2.

INTRODUCTION: The coronavirus disease 2019 (COVID-19) pandemic is caused by severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2). This outbreak started in Wuhan, China in December 2019. On 30 January 2020, the World Health Organization declared the outbreak as a Public Health Emergency of International concern. And, the COVID-19 was declared as a pandemic on 11 March, 2020 [1]. Coronaviruses are named after their characteristic crown-shaped microscopic appearance [2]. They are positive-sense single stranded RNA viruses. Taxonomically, they form part of the Coronaviridae sub-family (Coronaviridae family and Nidovirales order) [3, 4]. The SARS-CoV-2 spreads primarily through droplets of saliva or discharge from the nose. Hence, asked to cover nose and mouth while coughing or sneezing and also to maintain distance in public places [5].

There is a difference in infection and death cases recorded from different geographical locations of the world and also from different states of India. Total cases reported till 4th September 2020 in Kerala and Gujarat were 79,625 and 100,213, respectively. The death cases reported were 315 in Kerala and 3062 in Gujarat (Center for Systems Science and Engineering, 2020) [7]. Currently, the search for vaccines or treatments for COVID-19 is going on. To understand the root cause of differences in the infection and death rates, this study was planned to analyze the genomic and pathogenic similarities and differences between the

QUICK RESPONSE CODE

DOI:
10.13040/IJPSR.0975-8232.12(4).2411-16

This article can be accessed online on
www.ijpsr.com

DOI link: http://dx.doi.org/10.13040/IJPSR.0975-8232.12(4).2411-16

strains of SARS-CoV-2, isolated from different geographical locations of the world and different parts.

METHODOLOGY: Different genome sequence from the viral genome sequences database of the National Center for Biotechnology Information (NCBI) and different comparative analyses were performed. For this analysis, we used the SARS-CoV-2 reference genome, NC_045512, isolated from China. We downloaded three more sequences of the SARS-CoV-2 isolated from MT477885 - India, MT886288 -USA, and MT 835383 from Brazil. For a further detailed study of SARS-CoV-2, sequences isolated from different parts of India were downloaded: MT050491 isolated from Kerala, MT806104 from Surat, and MT799970 from Ahmedabad.

Sequence Alignment: Samples were aligned to the reference genome in the nucleotide BLAST using Basic Local Alignment Search Tool (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to extract the differences between the genome variants and the reference genome.

The parameter was set to a highly similar sequence. The sequences of SARS-CoV-2 isolated from different geographical locations of the world and India was compared with the reference genome of SARS-CoV-2.

Comparison of the GC Content: The genomic GC content was calculated among all the sequences by using GC content calculator of Biologics International Corp (BIC) (https:/www.biologics-corp.com/ tools/). FASTA format sequence was entered in the box provided (in FASTA format). The window size was selected for 30 and the sequence was submitted to obtain the results. GC content is also called G+C ratio or GC-ratio and it was calculated as a percentage by using the formula Count

$$(G + C)/Count\ (A + T + G + C) * 100\%.$$

Identified ORFs (Length > 150 bps): The ORF finder is a program available at NCBI website. It identifies all the open reading frames or the possible protein-coding region in sequence. The Open Reading Frame, ORF tool of NCBI (https://www.ncbi.nlm.nih.gov/orffinder/) was used to identify all the ORFs present in different sequences of SARS-CoV-2 isolated from different geographical locations of the world and India. ORF finder searches for open reading frames (ORFs) in the DNA sequence entered. FASTA format sequence, which starts with a '>' symbol followed by the sequence ID was entered in the box provided (in FASTA format). In the list of the genetic codes "standard" was selected with "AUG" as the start codon. Nested ORFs were ignored. The size of ORFs greater than 150 was selected. The ORF finder button was clicked to run the operation.

RESULTS AND DISCUSSION:

Sequence Comparison of SARS-Cov-2 (NC_045512.2) with Different Strains of SARS-Cov-2 Isolated from Different Geographical Locations - India, USA and Brazil: The reference genome of SARS-CoV-2 was aligned separately with different strains of SARS-CoV-2 isolated from India, USA and Brazil. This comparison was carried out to confirm if different strains of SARS-CoV-2 were infecting the population of different countries as the death rates were highly varied.

TABLE 1: SEQUENCE COMPARISON OF SARS-COV-2 (NC_045512.2) WITH DIFFERENT STRAINS OF SARS-COV-2 ISOLATED FROM DIFFERENT GEOGRAPHICAL LOCATIONS: INDIA, USA AND BRAZIL

Strains	GeneBank Accession Number	Country	Host	Year Isolated	Length (bp)	Query Cover (%)	E Value	Identify with Reference Sequence (%)
SARS-CoV-2	NC_045512.2	China	Homo sapians	2020	29903	100	0	100
SARS-CoV-2	MT477885	India	Homo sapians	2020	29899	99	0	99.98
SARS-CoV-2	MT886288	USA	Homo sapians	2020	29743	99	0	99.98
SARS-CoV-2	MT835383	Brazil	Homo sapians	2020	29858	99	0	99.98

All the three sequences were highly similar, with the reference sequence from China. They all showed 99.98% identity with the reference sequence **Table 1**. A high homology was observed between the sequences but a high variation in the death cases reported by each country.

As of data till 4th September 2020, the death cases reported by China were 4,634; India was 68,598; USA was 191, 060 and Brazil was 124,729. Mohanty *et al.*, linked the difference in the death rates to factors such as a large proportion of the old population, obesity or existing illness, climate, and the genetic differences [8]. The different ethnicity may have led to high immunity and weaker host-pathogen interaction of the individual [9].

Sequence Comparison of SARS-CoV-2 (NC_045512.2) with Different Strains of SARS-CoV-2 Isolated from Kerala, Surat and Ahmedabad: The reference genome of SARS-CoV-2 was aligned separately with the strains isolated from Kerala, Surat, and Ahmedabad. The identity percentage of a strain isolated from Surat and Ahmedabad was highly similar to the reference strain resulted in 99.97% and 99.96%, respectively **Table 2**. And the strain isolated from Kerala resulted in an identity percentage of 99.56% lesser than the other two strains found that the SARS-CoV-2 strain infecting Gujarat is different from the rest of India. She showed two missense mutations in the SARA-CoV-2 strain of Gujarat. The first deleterious mutation, C28854T in nucleocapsid (N) gene, which leads to high mortality in Gujarat, and second G25563T, located in Orf3a plays crucial roles in viral pathogenesis [11].

TABLE 2: SEQUENCE OF SARS-COV-2 (NC_045512.2) COMPARED WITH DIFFERENT STRAINS OF SARS-COV-2 ISOLATED FROM DIFFERENT PARTS OF INDIA: KERALA, SURAT AND AHMEDABAD

Strains	GeneBank Accession Number	Country	Host	Year Isolated	Length (bp)	Query Cover (%)	E Value	Identify with Reference Sequence (%)
SARS-CoV-2	NC_045512.2	China	Homo sapians	2020	29903	100	0	100
SARS-CoV-2	MT050491	Kerala	Homo sapians	2020	29850	95	0	99.56
SARS-CoV-2	MT806104	Surat	Homo sapians	2020	29800	99	0	99.97
SARS-CoV-2	MT799970	Ahmedabad	Homo sapians	2020	29800	99	0	99.96

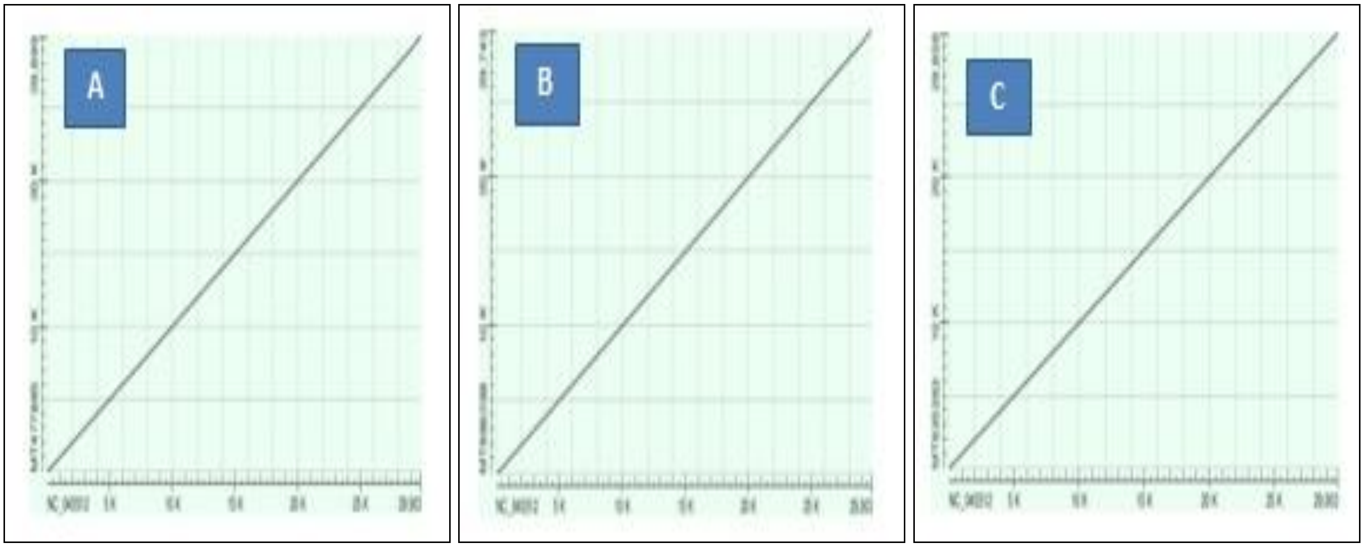


FIG. 1: DOT PLOT OF SEQUENCE ALIGNMENT A) SARS-COV-2 REFERENCE SEQUENCE (NC_045512.2) AND SARS-COV-2 ISOLATED FROM INDIA (MT477885). B) SARS-COV-2 REFERENCE SEQUENCE (NC_045512.2) AND SARS-COV-2 ISOLATED FROM USA (MT886288). C) SARS-COV-2 REFERENCE SEQUENCE (NC_045512.2) AND SARS-COV-2 ISOLATED FROM BRAZIL (MT835383)

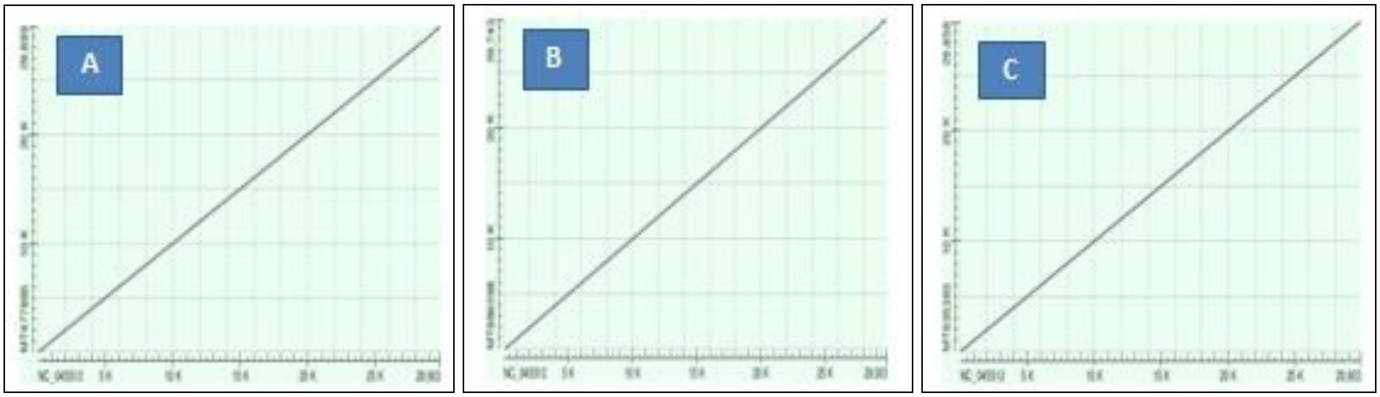


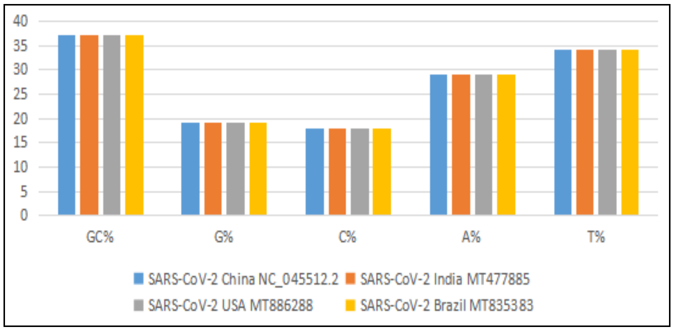
FIG. 2: DOT PLOT OF SEQUENCE ALIGNMENT A) SARS-COV-2 REFERENCE SEQUENCE (NC_045512.2) AND SARS-COV-2 ISOLATED FROM KERALA (MT050491). B) SARS-COV-2 REFERENCE SEQUENCE (NC_045512.2) AND SARS-COV-2 ISOLATED FROM SURAT (MT806104). C) SARS-COV-2 REFERENCE SEQUENCE (NC_045512.2) AND SARS-COV-2 ISOLATED FROM AHMEDABAD (MT799970)

The same was also supported by CSIR-Institute of Genomics and Integrative Biology by representing strains of SARS-Cov-2 of Kerala and Gujarat in different clades *i.e.* B4 clade and A3 clade, respectively [12] Rahila Sardar *et al.*, [13] also found a high rate of mutations within India strains (Rahila Sardar, 2020). The difference in the identity percentage may be due to the genome length variation of SARS-CoV-2 strains **Table 2**. The length of reference strain was 29903, while the genome length of a strain isolated from Kerala was 29850 giving low query value of 95%; thus giving low identity percentage.

The genome length of strains from Surat and Ahmedabad is 29800 with a query cover of 99%; resulting in a high identity percentage. The dot plots of SARS-CoV-2 strain from China China (reference sequence) and SARS-CoV-2 isolated from India **Fig. 1A,** USA **Fig. 2B** and Brazil **Fig. 2C** showed a complete diagonal line concluding high similarity between the genomes. The same was concluded with their 99.98% identity percentage **Table 2.**

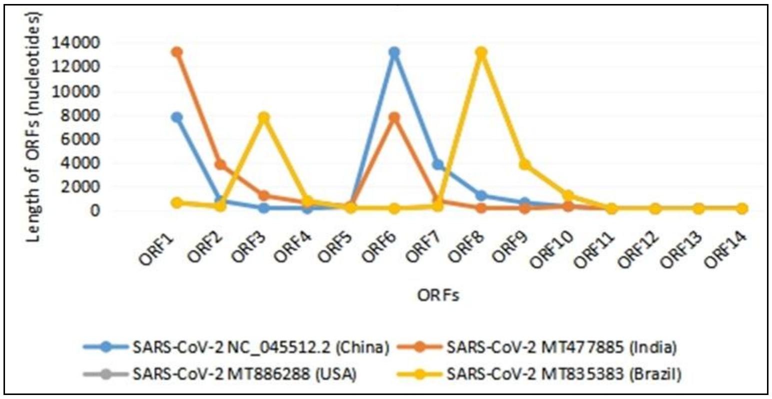
The dot plot of SARS-CoV-2 strain isolated from China (reference sequence) and SARS-CoV-2 isolated from Kerala did not form a complete diagonal line concluding a low identity percentage of 99.56% **Table 2**. The insertions or deletions were observed as the breaks or discontinuities in the diagonal lines **Fig. 2A**. The dot plot of SARS-CoV-2 strain isolated from reference sequence with sequence isolated from Surat **Fig. 2B** and Ahmedabad **Fig. 2C** showed a complete diagonal line hence are highly similar. The same was concluded with the identity percentage of 99.97% and 99.96% res-pectively **Table 2.**

Comparison of GC Content: The GC content of SARS-CoV-2 from all the four geographical locations- China, India, USA and Brazil showed the same GC content of 37% **Graph 1**. Also, the GC content of all the SARS-CoV-2 strains isolated from different parts of India showed the same GC content of 37% **Graph 1**. Hence, the coding regions of the core genome could be the same in all the strains of SARS-CoV-2 isolated from different geographical locations.



GRAPH 1: GC CONTENT TO DIFFERENT GENOMES OF SARS-COV-2 ISOLATED FROM DIFFERENT GEOGRAPHICAL LOCATION

Identified ORFs (Length > 150 bps): Comparing the length of different ORFs of various strains of SARS-CoV-2 isolated from different geographical locations *i.e.*, China, India, USA, and Brazil **Table 3** it can be concluded that the strains isolated from China (NC_045512.2) and India (MT477885) are similar to each other with lengths. But, India strain (MT477885) was highly assorted from Brazil (MT835383) and USA (MT886288) strains.



GRAPH 2: TOTAL ORFs AND LENGHTH OF ORFs WAS COMPARED WITH DIFFERENT STAINS OF SARS-COV2 ISOLATED FROM INDIA, USA, BRAZIL

TABLE 3: TOTAL ORFS AND LENGTH OF ORFS WAS COMPARED WITH DIFFERENT STRAINS OF SARS-COV-2 ISOLATED FROM INDIA, USA AND BRAZIL

Strain	GeneBank Acession Number (Country)	Total ORFs (>150)	Length of ORFs (Nucleotides)													
SARS-CoV-2			ORF1	ORF2	ORF3	ORF4	ORF 5	ORF6	ORF7	ORF8	ORF9	ORF10	ORF11	ORF12	ORF13	ORF14
	NC_045512.2 (China)	14	7788	828	228	186	366	13218	3849	1260	669	366	156	162	192	177
	MT477885 (India)	14	13218	3849	1260	669	366	7788	828	228	186	366	156	162	192	177
	MT886288 (USA)	13	669	366	7788	828	228	186	366	13218	3849	1260	177	162	192	
	MT835383 (Brazil)	14	669	366	7788	828	228	186	366	13218	3849	1260	177	156	162	192

ORFs of length greater than 150bps that begin with a start codon AUG and end with stop codons UAA, UAG or UGA were detected in different sequences of SARS-CoV-2 [14]. USA and Brazil have the highest number of COVID-19 cases and high death rates [19]. Observing the obtained result **Graph 2,** the same highly virulent strain of SARS-CoV-2 infects the population of the USA and Brazil. B Korber *et al.*, tracked Spike (S) real-time protein mutation in SARS-CoV-2. In his study, he found that the strains spreading so quickly in Europe and the U.S. have a mutated S “spike” protein that makes it about 10 times more infectious than the strain that was originally identified in Asia. This mutation does not make the virus more deadly, but it makes it more contagious. The original strain in China is D614, while the one found in North America is dubbed G614 [15]. The Indian microbiologist Lal R and with his team, studied the complete genomes of 95 strains of SARS-CoV-2 reported that the strain found in India almost matches with the sequence of SARS-CoV-2 found in Wuhan, but it is less virulent [16].

CONCLUSION: SARS-CoV-2 strains of China and India are highly similar in their ORF length was different from USA and Brazil. However, in the same way, length and identity percentage of the Kerala strain of SARS-CoV-2 was different from the reference sequence of china. From this outcome, we speculate, various factors causing the difference in reported death cases could be a result of mutation that has been observed in the sequence alignment data. Other possible factors could be large proportion of the age factor, chronic respiratory disease, Cardio vascular disease, Cancer, Diabetes, Climatic condition and ethnicity. An incomplete diagonal line in the dot plot of SARS-CoV-2 (Kerala) indicated insertion or deletion in the sequence. But, there is no variation in the GC content percentage of all of the sequences.

The ORF comparison study further revealed that the SARS-CoV-2 strains are completely similar in the USA and Brazil. This could be linked with the high virulence capacity of the strain, causing a high number of COVID-19 cases and deaths in these countries. The SARS-CoV-2 reference strains of India are most similar to the reference strain isolated from Wuhan while the death rate of India is still very less. This provides insight for further studies on linking the low COVID-19 death rate in India with the underlying immunity.

Based on this speculation, still more research work is required in a genomic, pathogenic, and immunological approach based on geo-graphical location of the world, and in India, it must identify the exact solution; hence people can get rid of the pandemic situation.

ACKNOWLEDGEMENT: Nil

CONFLICTS OF INTEREST: No conflicts of interest.

REFERENCES:

1. WHO, Coronavirus disease (COVID-19) pandemic. [Online] Available at: https: // www. who. Int / emergencies /diseases/novel-coronavirus-2019 [Accessed 23 May 2020].
2. Goldsmith CS, Tatti KM and Ksiazek TG: Ultrastructural characterization of SARS coronavirus. Emerg Infect Dis 2004; 10(2): 320-26.
3. Auewarakul P: Composition bias and genome polarity of RNA viruses. Virus Res 2005; 109(1): 33-37.
4. Coronaviridae. ICTV, 2020. [Online] Available at: talk.ictvonline.org/ictv-reports/ictv_9th_report/positive-sense-rna-viruses-2011 / w / posrna _ viruses / 222 / coronaviridae.
5. Andersen KG, Rambaut A, Lipkin WI, Holmes EC and Garry RF: The proximal origin of SARS-CoV-2. Nat Med 2020; 26(4): 450-52.
6. Kumar D, Malviya R and Sharma KP: Corona virus: a review of COVID-19. EJMO 2020; 4(1): 8-25.
7. Center for Systems Science and Engineering (CSSE), 2019. "COVID-19: Novel coronavirus." Data repository. [Online] Available at: https://github.com/CSSEGIS and Data/COVID-19.
8. Mohanty SK, Satapathy A and Naidu MM: Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) and coronavirus disease 19 (COVID-19) - anatomic pathology perspective on current knowledge. Diagn Pathol 2020; 15: 103.
9. Sharma G and Mehra NK: Indian population could have intrinsic immunity to resist COVID-19 challenge: International Union of Immunological Sciences (https://www. immunopaedia.org.za/breaking-news/indian-population-could-have-intrinsic-immunity-to-resist-covid-19-challenge/) 2020.
10. Santos AM: Covid-19 state wise status. [Online] Available at: https://www.mygov.in/corona-data/covid19-statewise-status/ 2020.
11. Joshi M, Puvar A, Kumar D, Ansari A, Pandya M, Raval J and Patel Z: Genomic variations in SARS-CoV-2 genomes from Gujarat: Underlying role of variants in disease epidemiology. Bio Rxiv 2020; 07-10: 197095.
12. Kumar P, Pandey R, Sharma P, Dhar MS and Vivekanand A and Uppili B: Integrated genomic view of SARS-CoV-2 in India. Bio Rxiv 2020; 6(4): 128751.
13. Sardar R, Satish DS, Birla S and Gupta D: Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis, bioRxiv 2020; 24-27.
14. WHO, World Health Organisation. [Online] Available at: https://www.who.int/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov), 2019.
15. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Foley B, Giorgi EE, Bhattacharya T, Parker MD, Partridge DG, Evans CM and de Silva TI: Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. Bio Rxiv 2020; 04(29): 069054.
16. Lal R, Kumar R, Verma H, Singhvi N, Sood U and Gupta V: Comparative genomic analysis of rapidly evolving SARS-CoV-2. Reveals Mosaic Pattern of Phylogeographical Distribution Systems 2020; 5(4): 00505-2.

How to cite this article:
Dhanalakshmi G and Anadi: Geographical based genomic and pathogenic study of the SARS-Cov-2. Int J Pharm Sci & Res 2021; 12(4): 2411-16. doi: 10.13040/IJPSR.0975-8232.12(4).2411-16.