

# The Distribution of COVID 19 based on Phylogeny Construction in Silico Sequences SARS-CoV-2 RNA at Genbank NCBI

*by* Yermia Mocosuli

---

**Submission date:** 05-Jan-2023 06:35AM (UTC+0700)

**Submission ID:** 1988665549

**File name:** Article.pdf (1.06M)

**Word count:** 2323

**Character count:** 12304

## The Distribution of COVID 19 based on Phylogeny Construction in *Silico* Sequences SARS-CoV-2 RNA at Genbank NCBI

Revolson Alexius MEGE<sup>1</sup>, Herry Sinyo SUMAMPOUW<sup>2</sup>, Dewa Nyoman OKA<sup>3</sup>,  
Nonny MANAMPIRING<sup>4</sup> and Yermia Samuel MOKOSULI<sup>4,5,\*</sup>

4

<sup>1</sup>Department of Biology Faculty of Mathematics and Natural Sciences, Manado State University, Tondano, Minahasa, North Sulawesi, Indonesia

<sup>2</sup>Department of Biology Education, Faculty of Mathematics and Natural Sciences, Manado State University, Tondano, Minahasa, North Sulawesi, Indonesia

<sup>3</sup>Department of Biology Education, Faculty of Mathematics and Natural Sciences, IKIP Saraswati Tabanan, Bali, Indonesia

<sup>4</sup>Department of Biology Faculty of Mathematics and Natural Sciences, Manado State University, Tondano, Minahasa, North Sulawesi, Indonesia

<sup>5</sup>Bioactivity and Molecular Biology Laboratory, Department of Biology, Faculty of Mathematics and Natural Sciences, Manado State University, Tondano, Minahasa, North Sulawesi, Indonesia

(\* Corresponding author's e-mail: yermiamokosuli@unima.ac.id)

Received: 10 April 2020, Revised: 10 July 2020, Accepted: 20 July 2020

### Abstract

The Covid-19 pandemic, due to severe acute respiratory coronavirus (SARS-CoV-2) virus, has an effect on human civilization today. With high fatality infections, SARS Covid-19 has influenced the global economic, socio-cultural, and even political order. This study aims to construct the phylogeny of the SARS corona virus that causes Covid-19 in various countries in the world by using the SARS Covid-19 gene database from the NCBI GenBank. The results of this study can trace the origin of SARS Covid-19, which is then called SARS-CoV-2, the gene characteristics, and the evolutionary relationship of these genes to various countries in the world. This research uses *in silico* method with gene sequence sources from the NCBI GenBank ([www.ncbi.nih.gov](http://www.ncbi.nih.gov)). A total of 433 SARS Covid-19 sequences reported by 21 countries as of April 2nd, 2020 were the subject of the study. Sequences representing each country were analyzed using the MEGA 7.0 program. The results showed that the phylogeny trees formed were obtained by 2 main monophyletic groups. The first major monophyletic group consisted of 11 nodes, with 19 SARS-CoV-2 gene sequences from 23 countries. The second major monophyletic group consisted of 5 nodes with 5 countries of origin of SARS-CoV sequence 19. The spread of Covid-19 from the epicenter in Wuhan, China to the world has taken place randomly. This has happened because of the migration of people from the Chinese epicenter. The location of the countries adjacent to China did not determine the closest phylogenetic relationship. The number of phylogenetic nodes formed showed mutases which caused very high variations of the SARS CoV 2 RNA gene sequence. The results of this study reinforce that efforts to limit the spread of human viruses to humans must be done. The presence of sequences from China in the 2 main monophyletic groups confirms that this virus originated in the Chinese epicenter.

**Keywords:** SARS Covid-19 gene, Genbank, NCBI, Spread

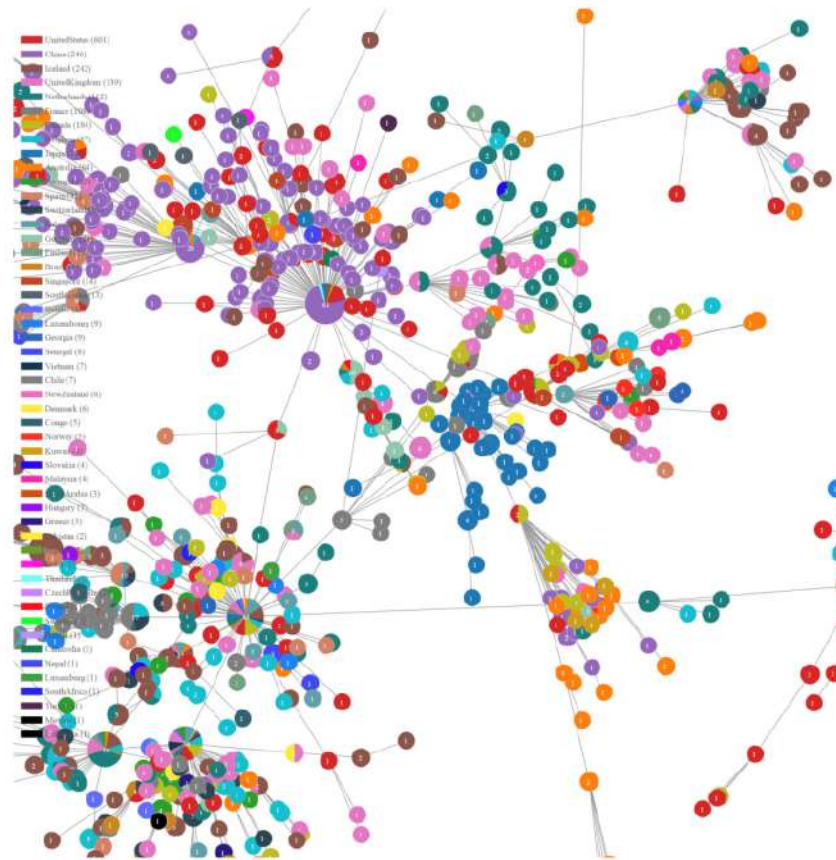
## Introduction

At the beginning of April 2020, the total number of positively infected SARS CoV-19 in the world was 750,890. The number of victims who died was 36,405 people. In Indonesia in early April 2020, the total number of positive SARS CoV-19 was 1,677, with a total death toll of 157. The SARS CoV-19 global fatality rate is around 4.85 %, while in Indonesia it is 9.36 % [1].

Coronavirus is a species of virus that belongs to the Coronavirinae subfamily in the Coronaviridae family, in the order of Nidovirales. Coronaviruses are viruses that are enveloped in a positive single-stranded RNA genome, belonging to a large family of viruses that are widely available in nature. Certain coronaviruses can infect humans and cause illness, such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS), whose symptoms can range from the common cold to severe lung infections [2]. Coronavirus is classified as an RNA virus that infects many animal species, including humans, other mammals, and birds. After infection of the host cell, respiratory, intestinal, liver, and neurological diseases follow [3]. Coronaviruses are members of the Nidovirales and subfamily Orthocoronavirinae. This subfamily is divided into 4 genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. In general, Alphacoronavirus and Betacoronavirus tend to infect mammals, whereas Gammacoronavirus and Deltacoronavirus usually infect birds. However, some Gammacoronavirus and Deltacoronavirus can infect mammals under certain conditions [4].

SARS-CoV and MERS-CoV show 79 and 50 % similarity, respectively, with SARS CoV-19. These findings indicate that there is no close evolutionary relationship between SARS COV-19 with SARS-CoV and MERS-CoV. Thus, SARS COV-19 is considered the seventh novel human Betacoronavirus [5]. Furthermore, reported by Gao *et al.* [6,7], SARS CoV-2019 has the closest genetic relationship with Pangolin-CoV [8].

The origin of the 2019 SARS CoV is still being debated by scientists today. However, the search for SARS CoV-19, which turned into SARS CoV-20 and which has become a pandemic in the world, needs to be done. In addition to knowing the pattern of spread, the characteristics of the SARS CoV-20 gene sequence also determine mutations that occur. One approach that can be done is to use the SARS CoV-20 RNA sequence that has been reported and recorded in the gene bank. Based on data from the China National Center for Bioinformation until March 26<sup>th</sup>, 2020 [2], there are 2,103 strains and 1,135 haplotypes of SARS CoV-20. Haplotype maps of novel coronavirus worldwide and from China, respectively, are built on the basis of knowledge on genome variation obtained from available sequences of high quality genomes (Figure 1). The spread of SARS CoV-20 is very fast and massive compared to other viruses that have become pandemic.



**Figure 1** SARS CoV-2020 strains from 2020-03-26 with 2,103 strains and 1,135 haplotypes found. (<https://bigd.big.ac.cn/nCoV/network?lang=en> citation on April 2<sup>nd</sup>, 2020).

Thus, genetic tracing is needed to provide data on variations of SARS CoV-20 from various countries that have reported their sequences. The big question is how the spread of SARS CoV-2020 is seen from the aspect of phylogeny, based on sequence data that has been reported and validated in the NCBI Genbank. This study aims to obtain the characteristics and phylogeny evolutionary relationship of SARS CoV-2020, which has become a pandemic today.

**Materials and methods**

**5** The 16 RNA SARS Covid-19 gene sequences obtained from the NCBI Genbank (<https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/>) of 343 recorded sequences were reported in the NCBI GenBank (**Table 1**). A total of 25 RNA SARS COV-2 sequences were used, representing various countries from various continents. Sequences downloaded from the NCBI Genbank in txt format were

converted to MEGA format using the MEGA 7.0 Program [9,10]. Next, the sequence groups were analyzed for substitution to determine the phylogeny tree model. Substitution analysis results were obtained by the Neighbor Joining model with Bootstrap 1000x.

**Table 1** Sources of Covid-19 SARS RNA gene sequences from NCBI GenBank [11].  
<https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs> (Accessed April 1<sup>st</sup>, 2020).

**Nucleotide Sequences**

You can view and download these 433 GenBank sequences and 1 RefSeq sequence in [Entrez Nucleotide](#) and the new [NCBI Virus](#) resource.  
[BLAST against Betacoronavirus sequences](#)

GenBank	RefSeq	Gene Region	Collection Date	Locality
MN008947	NC_045512	complete	2019-12	China
LC522359		RdRP	2020-01-26	Philippines
LC523807		N	2020-01-06	Philippines
LC523808		N	2020-01-26	Philippines
LC523809		N	2020-01-23	Philippines
LC528232		complete	2020-02-10	Japan
LC528233		complete	2020-02-10	Japan
LC529005		complete	2020-01	Japan
LC534418		complete	2020-02-14	Japan
LC534419		complete	2020-03-09	Japan
LR157995		complete	2020-01-05	China: Wuhan
LR157996		complete	2020-01-31	China: Wuhan
LR157997		complete, gapped	2019-12-31	China: Wuhan
LR157998		complete	2019-12-26	China: Wuhan
MN938384		complete	2020-01-10	China: Shenzhen
MN938385		RdRP	2020-01	China: Shenzhen
MN938386		RdRP	2020-01	China: Shenzhen
MN938387		S	2020-01	China: Shenzhen
MN938388		S	2020-01	China: Shenzhen
MN938389		S	2020-01	China: Shenzhen
MN938390		S	2020-01	China: Shenzhen
MN970803		RdRP	2020-01-08	Thailand
MN970804		RdRP	2020-01-13	Thailand
MN975262		complete	2020-01-11	China
MN975263		RdRP	2020-01	China

Showing 1 to 25 of 433 entries

1 2 3 4 5 ... 18

**Results and discussion**

As of April 1<sup>st</sup>, 2020, 433 SARS-CoV-2 sequences have been reported. The phylogeny construction is based on the country of origin reporting the sequence. The United States is the country with the most SARS-CoV-2 sequences of 282 sequences, followed by China with 77 sequences, and Iran with 14 sequences. As of April 1<sup>st</sup>, 2020, 23 countries have reported sequences of SARS-CoV-2, both complete and partial gene regions. No data has been found yet on the SARS-CoV-2 sequence from Indonesia. Italy, as a country with 433 positive cases of SARS-CoV-2 sequences, has the most relatively small sequence report data (**Table 2**).



**Table 2** SARS-CoV-2 sequences from GenBank NCBI (April 1<sup>st</sup>, 2020).

No	Country	Region	Number
1	China	Hangzhou	26
		Shenzhen	7
		Wuhan	26
		Yunnan	1
		Shanghai	1
		Guangzhou	4
		China	12
2	Philippines	Philippines	1
3	Japan	Japan	5
4	Finland	Finland	1
5	Thailand	Thailand	2
6	United States of America	United States of America	27
		Massachusetts	1
		Wisconsin	1
		Washington	229
		Minnesota	3
		Illinois	2
		California	17
		Texas	1
		Arizona	1
		North Korea	North Korea
8	Australia	Victoria	1
		Queensland	6
9	Italy	Roma	4
		Cagliari	2
10	India	India	3
11	Malaysia	Malaysia	3
12	Taiwan	Taiwan	3
13	Belgium	Belgium	2
14	Nepal	Nepal	1
15	Sweden	Sweden	1
16	Brazil	Brazil	1
17	Vietnam	Vietnam	6
18	Iran	Iran	14
19	Nigeria	Nigeria	1
20	Spain	Valencia	12
21	Pakistan	Pakistan	2
22	Colombia	Colombia	1
23	Peru	Peru	1
		<b>Total</b>	<b>433</b>

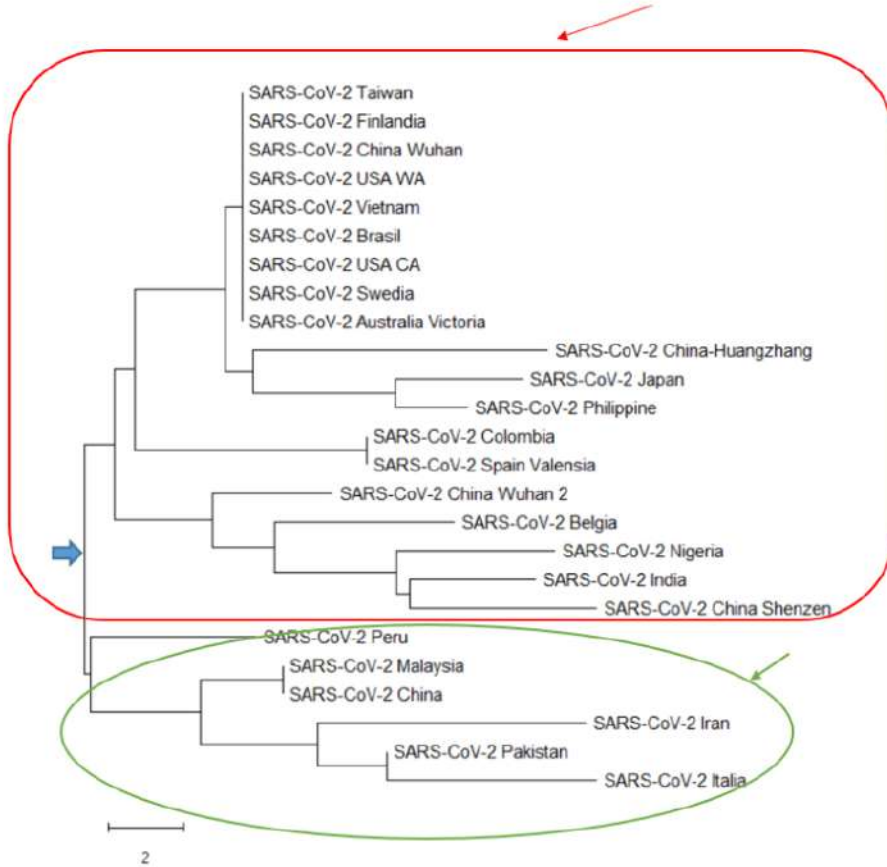


*Nucleotide-Composition	
1	
2	Data Filename: revinsi negara 1 april 2020.meg
3	Data Title:
4	Nucleotide Frequencies
5	Filter Used: All selected
6	All frequencies are given in percent:
7	Domain: Data
8	
9	SARS-CoV-2 Peru 32.1 18.4 29.9 19.6 29856 36.7 15.0 28.6 18.8 9952 25.9 18.3 30.8 25.0 9952 33.8 21.6 39.2 15.0 9952
10	SARS-CoV-2 Colombia 32.1 18.4 29.8 19.7 28920 36.4 15.9 28.5 19.1 9640 25.9 18.3 30.9 24.9 9640 34.0 20.5 30.1 15.0 9637
11	SARS-CoV-2 Pakistan 32.1 18.4 29.9 19.6 29836 36.7 15.9 28.6 18.9 9946 25.9 18.3 30.8 25.0 9945 33.8 21.0 39.2 15.0 9945
12	SARS-CoV-2 Saoxoi Valencia 32.1 18.4 29.8 19.6 29860 36.7 15.9 28.5 18.9 9937 25.9 18.3 30.8 25.0 9937 33.8 21.0 39.2 15.0 9936
13	SARS-CoV-2 Vietnam 32.1 18.4 29.9 19.6 29891 36.7 15.9 28.6 18.8 9968 25.9 18.2 30.9 25.0 9964 33.7 21.0 39.3 15.0 9963
14	SARS-CoV-2 India 21.1 25.1 31.1 22.7 399 14.3 26.3 30.8 28.6 133 12.8 30.1 31.6 25.6 133 36.1 19.1 31.6 12.8 133
15	SARS-CoV-2 Nigeria 33.3 18.4 28.1 20.1 483 24.2 14.9 31.1 29.8 163 30.4 22.4 29.8 17.4 161 45.3 18.6 21.6 11.0 161
16	SARS-CoV-2 Iraq 26.1 21.1 26.5 27.3 322 42.6 34.1 19.4 15.0 109 10.3 25.2 31.5 43.0 107 25.2 14.6 35.8 26.2 107
17	SARS-CoV-2 Brazil 32.1 18.4 29.9 19.6 29876 36.7 15.9 28.6 18.8 9959 25.9 18.3 30.9 25.0 9959 33.8 21.0 39.2 15.0 9958
18	SARS-CoV-2 Sweden 32.1 18.4 29.9 19.6 29866 36.7 15.9 28.6 19.0 9943 25.9 19.3 30.9 25.0 9942 33.7 21.0 39.3 15.0 9942
19	SARS-CoV-2 China Hubei 21.0 25.0 31.7 22.2 1260 15.0 24.0 31.4 29.5 420 16.4 28.3 33.3 21.9 420 31.7 22.4 30.5 15.2 420
20	SARS-CoV-2 Belgium 39.4 17.0 26.7 16.9 870 31.7 13.8 29.9 24.6 228 29.9 22.0 24.1 13.9 225 46.6 15.2 26.0 12.1 223
21	SARS-CoV-2 Taiwan 32.1 18.4 29.9 19.6 29870 36.7 15.9 28.6 18.9 9957 25.9 18.2 30.8 25.0 9957 33.8 21.0 39.2 15.0 9956
22	SARS-CoV-2 Malaysia 24.1 15.0 27.1 15.7 190 40.2 19.6 27.0 15.5 97 47.0 18.6 29.3 24.7 97 34.4 19.8 27.1 14.9 96
23	SARS-CoV-2 Finlandia 32.1 18.4 29.9 19.6 29806 36.7 15.9 28.5 18.9 9936 25.9 18.3 30.8 25.0 9935 33.7 21.0 39.2 15.0 9935
24	SARS-CoV-2 Iceland 21.7 23.9 22.4 22.0 222 39.0 25.0 21.2 15.7 109 18.8 26.2 29.0 28.0 107 40.2 20.4 25.8 22.4 107
25	SARS-CoV-2 Australia Victoria 32.1 18.4 29.9 19.6 29893 36.7 15.9 28.5 18.9 9945 25.9 18.2 30.9 25.0 9944 33.7 21.0 39.4 15.0 9944
26	SARS-CoV-2 USA CA 32.1 18.4 29.9 19.6 29882 36.7 15.9 28.6 18.8 9943 25.9 18.3 30.8 25.0 9941 33.8 21.0 39.3 15.0 9940
27	SARS-CoV-2 USA WA 32.1 18.4 29.9 19.6 29882 36.7 15.9 28.6 18.8 9943 25.9 18.3 30.8 25.0 9941 33.8 21.0 39.3 15.0 9940
28	SARS-CoV-2 Cina 24.1 15.0 27.1 15.7 190 40.2 19.6 25.8 15.5 97 27.8 18.6 28.9 24.7 97 34.4 19.4 27.1 18.8 96
29	SARS-CoV-2 Cina Shenzhen 32.1 18.4 29.9 19.6 29838 36.8 15.0 28.2 15.0 9946 36.7 15.9 28.5 18.9 9946 25.9 18.3 30.8 25.0 9946
30	SARS-CoV-2 Cina Wuhan 32.1 18.4 29.9 19.6 29903 36.8 15.9 28.7 18.8 9968 25.9 18.2 30.9 25.0 9968 33.7 21.0 39.3 15.0 9967
31	SARS-CoV-2 Cina Wuhan 2 32.1 18.4 29.9 19.6 29872 36.7 15.9 28.6 18.8 9958 25.9 18.3 30.8 25.0 9957 33.7 21.0 39.3 15.0 9957
32	SARS-CoV-2 Japan 32.1 18.4 29.9 19.6 29902 36.7 15.9 28.6 18.8 9968 25.9 18.3 30.8 25.0 9967 33.7 21.0 39.3 15.0 9967
33	SARS-CoV-2 Filipina 28.6 19.2 33.0 15.2 182 39.3 21.3 31.1 8.2 61 18.0 13.1 41.3 27.9 61 28.3 23.3 26.7 21.7 60
34	Avg. 32.1 18.4 29.9 19.6 19244.4 34.4 16.2 18.7 16.7 6415.3 26.5 18.1 30.7 24.6 6414.8 33.3 20.8 30.3 15.6 6414.3

**Figure 2** The phylogeny tree construction process with the MEGA 7.0 Program. (a) Alignment of SARS CoV-19; (b) sequences, phylogeny tree construction; (c) SARS COV nucleotide compositions from sequences of countries around the world.

In the phylogeny tree formed, 2 main monophyletic groups were obtained. The first monophyletic group (red circle) consisted of 11 nodes. The first monophyletic group contained 19 SARS-CoV gene sequences from 15 countries. The second major monophyletic group (green circle) consisted of 5 nodes with 5 countries of origin of SARS - CoV sequence 19. Based on the phylogeny tree, the second monophyletic group is closer to the ancestor virus (arrows in purple). It is interesting that SARS CoV-2 from Peru is closer to the initial node (the oldest ancestor). Most SARS CoV-2 originating from the spreading epicenter of China is in the first monophyletic group. In the second monophyletic group, only 1 sequence originated from China (**Figure 3**).





**Figure 3** SARS CoV-2 phylogeny tree built using the 1000× Neighbor Joining Bootstrap model.

This phylogenetic tree obtained describes that the pattern of the spread of SARS CoV-19/20 took place randomly in various countries. SARS CoV-2 is the oldest sequence from Peru, indicating that the sequences from Peru originated directly from the epicenter in China. From the table above, the United States reported the most SARS CoV-19 sequences compared to China as the epicenter center. Thus, many SARS CoV-19 sequences are not or have not been recorded in the NCBI Genbank. In the phylogeny tree, it can also be explained that SARS CoV-19 from China has a similarity of sequence close to 100 %. The same was also found in the first monophyletic group, SARS CoV-19, reported from Taiwan, Finland, Wuhan China, USA WA, Vietnam, Brazil, USA CA, Sweden, and Australia Victoria (**Figure 4**).



**Figure 4** Countries with the most similar SARS CoV-19 gene sequences. (a) In the first monophyletic group and (b) in the second monophyletic group.

### Conclusions

The spread of Covid-19 from the epicenter in Wuhan, China, to the rest of the world was random. This is thought to have occurred due to human movements from the epicenter of China. The location of the countries adjacent to China does not determine the phylogeny relationship of the closest virus. The results of this study reinforce that efforts to limit the spread of the human viruses to humans must be done. The presence of virus sequences from China in the 2 main monophyletic groups confirms that this virus originated from the Chinese epicenter.

### Acknowledgements

The authors thank researchers in the bioactivity and molecular biology laboratory, FMIPA Unima, who helped analyze the results of this research.

### References

- [1] Data COVID-19, Available at: <https://www.covid19.go.id>, accessed April 2020.
- [2] China National Center for Bioinformatics, Available at: <https://bigd.big.ac.cn/ncov/about?lang=en>, accessed April 2020.
- [3] J Cui, F Li and ZL Shi. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 2019; **17**, 181-92.
- [4] K Subbaram, H Kannan and MK Gatasheh. Emerging developments on pathogenicity, molecular virulence, epidemiology and clinical symptoms of current middle east respiratory syndrome coronavirus (MERS-CoV). *Hayati. J. Biosci.* 2017; **24**, 53-6.
- [5] Y Chen, Q Liu and D Guo. Emerging coronaviruses: Genome structure, replication, and pathogenesis. *J. Med. Virol.* 2020; **92**, 418-23.
- [6] L Gao, J Qi, H Wei, Y Sun and B Hao. Molecular phylogeny of coronaviruses including human SARS-CoV. *Chinese Sci. Bull.* 2003; **48**, 1170-4.
- [7] Y Gao, T Li and L Luo. *Phylogenetic study of 2019-nCoV by using alignment-free method. Ic.* 2019.
- [8] ZW Yc, S Yuan, KS Yuen, SY Fung, CP Chan and DY Jin. Zoonotic origins of human

- coronaviruses. *Int. J. Biol. Sci.* 2020; **16**, 1686-97.
- [9] BG Hall. Building phylogenetic trees from molecular data with MEGA. *Mol. Biol. Evol.* 2013; **30**, 1229-35.
- [10] S Kumar, G Stecher and K Tamura. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 2016; **33**, 1870-4.
- [11] National Center for Biotechnology Information, Available at: <https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs>, accessed March 2020.

# The Distribution of COVID 19 based on Phylogeny Construction in Silico Sequences SARS-CoV-2 RNA at Genbank NCBI

## ORIGINALITY REPORT

12%

SIMILARITY INDEX

8%

INTERNET SOURCES

10%

PUBLICATIONS

%

STUDENT PAPERS

## PRIMARY SOURCES

- 1** [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) 3%  
Internet Source
- 2** [covid.portugene.com](http://covid.portugene.com) 3%  
Internet Source
- 3** Braulio Cervantes-Paz, Elhadi M. Yahia, José de Jesús Ornelas-Paz, Claudia Inés Victoria-Campos et al. "Bioaccessibility of fat-soluble bioactive compounds (FSBC) from avocado fruit as affected by ripening and FSBC composition in the food matrix", Food Research International, 2021 3%  
Publication
- 4** Simanjuntak Suddin, Yermia Samuel Mocosuli, Welerubun Marcelina, Naharia Orbanus, Kapahang Ardi. "Molecular Barcoding Based 16S rRNA Gene of Thermophilic Bacteria from Volcanic Sites, Linow Lake, Tomohon", Materials Science Forum, 2019 2%  
Publication



Exclude quotes      On

Exclude matches      < 2%

Exclude bibliography      On