Int. Res. J. of Science & Engineering, Volume 8 (2) March-April, 2020
ISSN: 2322-0015

SJIF Impact Factor 6.70

RESEARCH ARTICLE

OPEN ACCESS

SARS COV 2 Identity matrix

Rithu BS

Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India
Email: rithu2516@gmail.com

Manuscript Details

Received: 29.03.2020
Accepted: 08.04.2020
Published: 09.04.2020

Available online on http://www.irjse.in
ISSN: 2322-0015

Cite this article as:
Rithu BS. SARS COV 2 Identity matrix, *Int. Res. Journal of Science & Engineering*, 2020. Volume 8 (2) : 41-57.

Abstract

Covid-19 is a infectious disease caused by severe acute respiratory syndrome corona virus 2 (SARS COV 2), which has become a global pandemic leading to life threatening illness. RNA viral genome is considered to undergo mutation faster that DNA virus either due to replication error or recombination. This variation occurring in viral genome can be either fatal or favourable to the host. Hence it is important to identify the variant's present across the world to understand the virulence and manifestation of disease. In our study we have performed multiple sequence alignment of genome sequence collected from different geographical location explore the divergence. To gain further insights we also performed multiple sequence alignment of 3 target namely N gene, S gene & RdRP gene. Our study revealed the variants present across the world as well within a country to help the researchers working on development of universal diagnostic kits and drug discovery.

Keywords: SARS COV 2, multiple sequence alignment, N gene, S gene & RdRP gene.

1. Introduction

SARS COV 2:

In December 2019, the first case was reported at Wuhan (China), which was later declared pandemic as global emergency due to its severity and high mortality rate across the world [1][2]. Recent study revealed similarity between SARS COV 2 with bat SARS COV [3]. SARS COV 2 belongs to *coronaviridaie* family, *betacoronavirus* genus and was titled as COVID-19[4]. scientist across the world are initiating work related to genome sequencing to understand the rate of mutation and variant's present across the world [5].

N gene:

Coronavirus nucleocapsid (N) is a structural protein coded by N gene. The main role of the protein coded by this gene is to form complexes with genome to initiate host interaction with viral membrane protein during viral replication and assembly [6], it also has an important role in enhancing the rate of viral replication within the host system [7]. The major function of Cov N is to pack viral genome into ribonucleoprotein called as nucleocapsid [8]. This capsid acts as protection shield to viral genome [9]. Some supporting features of Cov N is it helps in interaction with ER-Golgi body of host to initiate budding of virus [10].

S gene:

S gene codes for spike glycoprotein and plays a major role in understanding the epidemiology of the disease [11]. Spike protein is considered as main surface antigen of SARS COV 2 [12]. The main function of this protein is to attach the virion to host cell membrane by ligand-receptor interaction to manifest the infection in host body [13]. They mainly involve in merging viral envelope during viral penetration [14].

RdRP gene:

RdRP gene codes for the enzyme called as RNA dependent RNA polymerase which has a vital role in replicating the viral genome and initiating transcription [15]. This enzyme has an error rate copying 1/10000 base pair due to lack of proof-reading capacity leading to mutation during replication [16]. Another special feature of this enzyme is it can initiate recombination (re-arrangement) with the co-infected viral genome or host genome leading to development of new variants [17].

Mutation in virus:

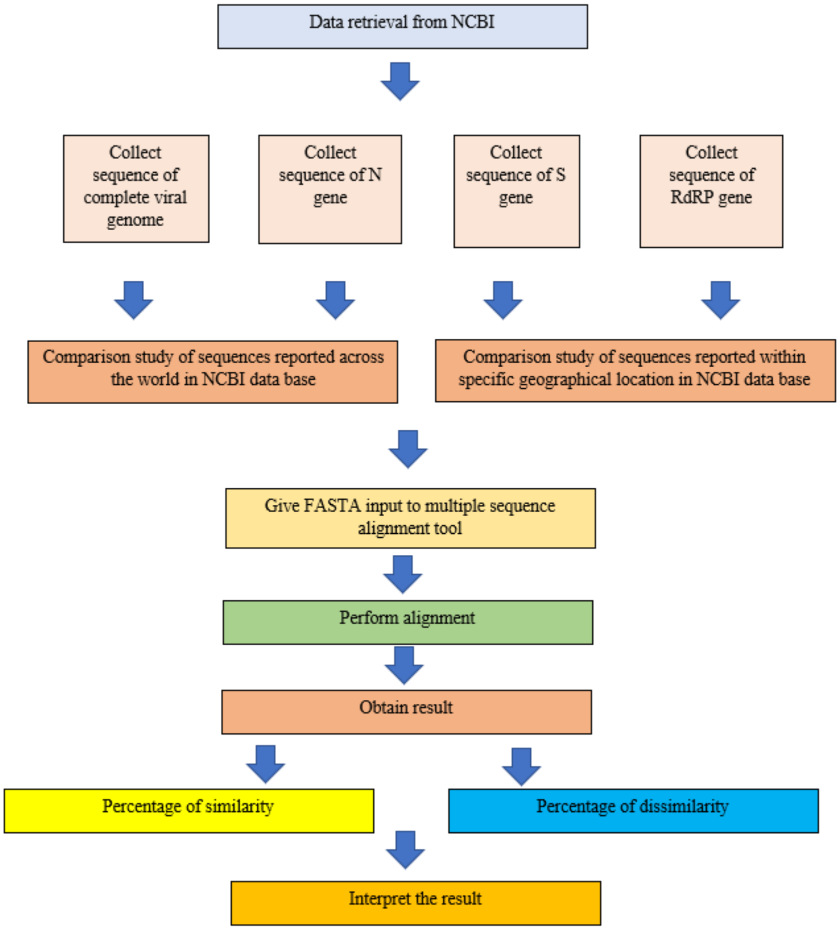
Mutation is considered as building block of evolutionary change leading to development of new traits [18]. This mutation which occurs are mostly beneficial for the virus increase its persistence in host [19]. When it comes to RNA virus the rate of mutation in viral genome is trillion times faster than the host system [20].

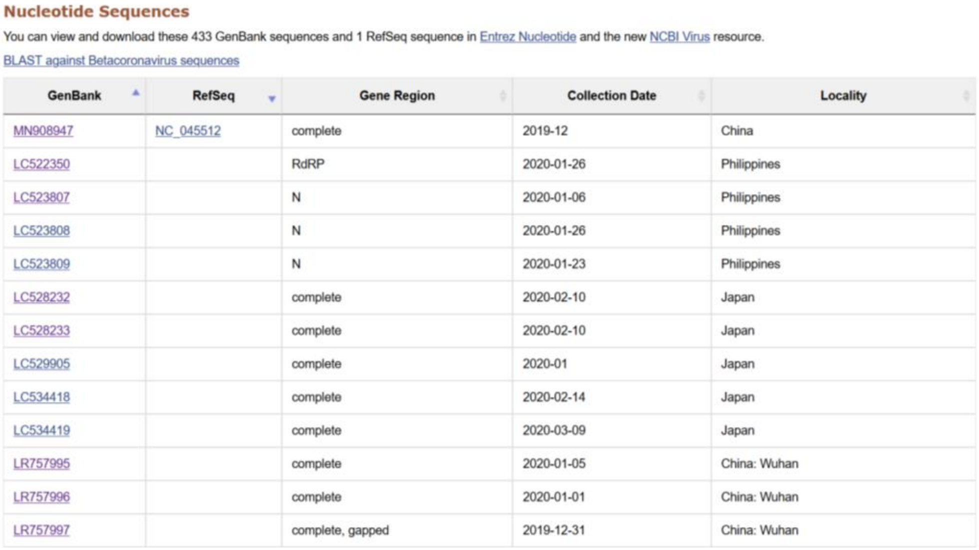
2. Methodology

Data retrieval:

2.1. Data retrieval to perform Complete genome comparison across the world

A list of 14 SARS COV 2 viral genome is downloaded from national center for biotechnology information (NCBI). (www.ncbi.nlm.nih.giv/). This NCBI link opens the home page of NCBI data base. Select the dropdown menu and NUCLIOTIDE from the choices provided by NCBI is selected. Then "SARS COV 2 "is searched which can be seen in figure 1. As per date 1st April 2020 there were totally 433 entries out of which we selected the first sequence reported from 14 countries across the world [1]. table 1 gives the list of genome sequence data collected from respective country along with the date were considered for the study.


Data retrieval from NCBI
Collect sequence of complete viral genome
Collect sequence of N gene
Collect sequence of S gene
Collect sequence of RdRP gene
Comparison study of sequences reported across the world in NCBI data base
Comparison study of sequences reported within specific geographical location in NCBI data base
Give FASTA input to multiple sequence alignment tool
Perform alignment
Obtain result
Percentage of similarity
Percentage of dissimilarity
Interpret the result



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
MN908947	NC_045512	complete	2019-12	China
LC522350		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
LC529905		complete	2020-01	Japan
LC534418		complete	2020-02-14	Japan
LC534419		complete	2020-03-09	Japan
LR757995		complete	2020-01-05	China: Wuhan
LR757996		complete	2020-01-01	China: Wuhan
LR757997		complete, gapped	2019-12-31	China: Wuhan

Figure 1: NCBI data base showing results of SARS COV 2

Table1: list of countries which were selected for genome comparison along with gene ID

NCBI genome ID	Country	Date sequence reported
MT007544	Australia Victoria	25-01-2020
LC528232	Japan	10-02-2020
MT240479	Pakistan	04-03-2020
MT012098	India Kerala	27-02-2020
MT072688	Nepal	13-01-2020
MT039890	South Korea	01-2020
MT198652	Spain	04-03-2020
MT050493	India Kerala -2	31-01-2020
MT126808	Brazil	28-02-2020
MT027062	USA:CA	29-01-2020
MN908947	china	12-2019
MT192772	Vietnam	22-01-2020
MN985325	USA: WA	19-01-2020
MT066175	Taiwan	31-12-2020

2.2 Data retrieval to perform complete genome comparison within a specific geographical location which was worst hit due to the global pandemic:

2.2.1 China

To find the variability exhibited among known SARS-COV 2 genome sequence reported in NCBI, a set of 8 genome sequence data was retrieved which lie between the time span of December 2019- January 2020.

Table2: list of genome sequence selected from NCBI within specific location: china

NCBI genome ID	Country	Date sequence reported
LR757997	China Wuhan	31-12-2020
MT259226	China Wuhan	10-01-2020
MT259228	China Wuhan	26-01-2020
MT019530	China Wuhan	30-12-2019
MT019529	China Wuhan	23-12-2019
LR757995	China Wuhan	05-01-2020
LR757998	China Wuhan	26-12-2020

2.2.2 Spain

To find the variability exhibited among known SARS-COV 2 genome sequence reported in NCBI, a set of 4 genome sequence data was retrieved which lie between the time span of February 2019- January 2020.

Table3: list of genome sequence selected from NCBI within specific location: Spain

NCBI genome ID	Country	Date sequence reported
MT256918	Spain	06-03-2020
MT233522	Spain	02-03-2020
MT233519	Spain	27-02-2020
MT233523	Spain	04-03-2020

2.2.3 USA

To find the variability exhibited among known SARS-COV 2 genome sequence reported in NCBI, a set of 8 genome sequence data was retrieved which lie between the time span of January 2019- February 2020.

Table 4: list of genome sequence selected from NCBI within specific location: USA

NCBI genome ID	Country	Date sequence reported
MT263386	USA	22-03-2020
MT246449	USA	13-03-2020
MN994467	USA	23-03-2020
MN985325	USA	19-01-2020
MT163719	USA	01-03-2020
MT159705	USA	17-02-2020
MT027064	USA	29-01-2020
MT118835	USA	23-02-2020

2.3 Data retrieval to perform comparison of N gene sequence reported across the world

A list of 8 N gene sequence is downloaded from national center for biotechnology information (NCBI). (www.ncbi.nlm.nih.giv/). This NCBI link opens the home page of NCBI data base. Select the dropdown menu and NUCLEOTIDE from the choices provided by NCBI is selected. Then "SARS COV 2 N gene "is given as keyword. From the list 8 sequences reported across the world from the time span between December 2019- march 2020 were considered for study.

Table 5: list of N gene sequence ID selected for comparison study across the world

NCBI genome ID	Country	Date sequence reported
MT163714	INDIA	04-03-2020
MT163715	INDIA	04-03-2020
MT186676	Iran	10-03-2020
MT192758	Italy	13-03-2020
LC523807	Philippines	11-02-2020
MT081059	China-1	13-02-2020
MT081063	China-2	13-02-2020
MT081068	China-3	13-02-2020

2.4 Data retrieval to perform N gene sequence comparison within a specific geographical location which was worst hit due to the global pandemic:

2.4.1 Philippines

To find the variability exhibited among N Gene sequence reported in NCBI, a set of 4 gene sequence data was retrieved which was reported by Philippines.

Table 6: list of N gene sequence selected from NCBI within specific location: Philippines

NCBI genome ID	Country
LC523807	Philippines
LC523808	Philippines
LC523809	Philippines

2.4.2 China

To find the variability exhibited among N Gene sequence reported in NCBI, a set of 10 gene sequence data was retrieved which was reported by china.

Table 7: list of N gene sequence selected from NCBI within specific location: China

NCBI genome ID	Country
MT081059	China
MT081060	China
MT081061	China
MT081062	China
MT081063	China
MT081064	China
MT081065	China
MT081066	China
MT081067	China

2.4.2 India

To find the variability exhibited among N Gene sequence reported in NCBI, a set of 2 gene sequence data was retrieved which was reported by India.

Table 8: list of N gene sequence selected from NCBI within specific location: India

NCBI genome ID	Country
MT163714	India
MT163715	India

2.4.3 Iran

To find the variability exhibited among N Gene sequence reported in NCBI, a set of 2 gene sequence data was retrieved which was reported by Iran.

Table 9: list of N gene sequence selected from NCBI within specific location: Iran

NCBI genome ID	Country
MT186676	Iran
MT186677	Iran
MT186678	Iran
MT186679	Iran
MT186680	Iran
MT186681	Iran
MT186682	Iran

2.4.4 Italy

To find the variability exhibited among N Gene sequence reported in NCBI, a set of 2 gene sequence data was retrieved which was reported by Iran.

Table 10: list of N gene sequence selected from NCBI within specific location: Italy

NCBI genome ID	Country
MT187977	Italy
MT192758	Italy

2.5 Data retrieval to perform S gene sequence comparison across the world:

A list of 9 S gene sequence is downloaded from national center for biotechnology information (NCBI). (www.ncbi.nlm.nih.giv/). This NCBI link opens the home page of NCBI data base. Select the dropdown menu and NUCLIOTIDE from the choices provided by NCBI is selected. Then "SARS COV 2 S gene "is given as keyword. From the list 9 sequences reported across the world from the time span between December 2019- march 2020 were considered for study.

Table 11: list of S gene sequence selected from NCBI across the world

NCBI genome ID	Country
MT232871	Iran
MT232872	Iran
MN938387	China Shenzhen
MN938388	China Shenzhen
MN938389	China Shenzhen
MN938390	China Shenzhen
MN975266	China Wuhan
MN975267	China Wuhan
MN975268	China Wuhan

2.6 Data retrieval to perform RdRP gene sequence comparison across the world:

A list of 10 RdRP gene sequence is downloaded from national center for biotechnology information (NCBI). (www.ncbi.nlm.nih.giv/). This NCBI link opens the home page of NCBI data base. Select the dropdown menu and NUCLIOTIDE from the choices provided by NCBI is selected. Then "SARS COV 2 RdRP gene "is given as keyword. From the list 10 sequences reported across the world from the time span between December 2019- march 2020 were considered for study.

Table 12: list of RdRP gene sequence selected from NCBI across the world

NCBI genome ID	Country
MN970003	Thailand
MT066157	Malaysia
MT127116	Vietnam
MT042773	Wuhan
MT050414	Australia
MT159778	Nigeria
LC522350	Philippines
MT072668	Belgium
MN938385	China
MN975263	China
MT232869	Iran

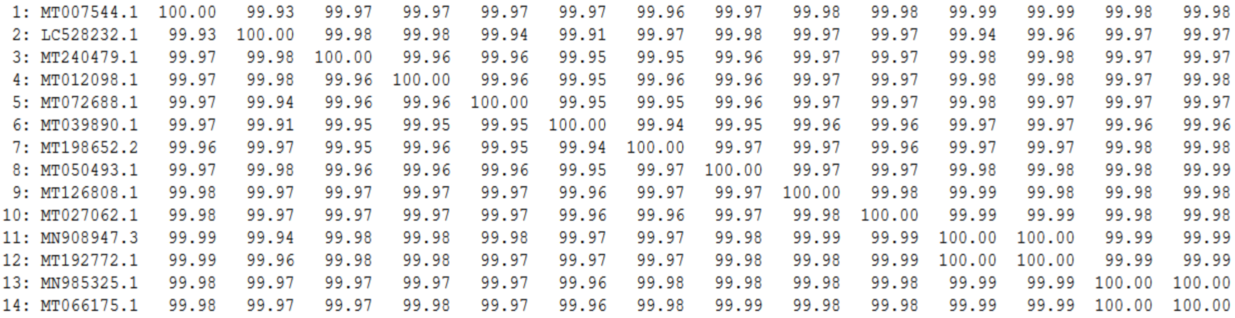
2.7 Multiple sequence alignment:

Clustal omega is a multiple sequence alignment program (www.ebi.ac.uk/Tools/msa/clustalo/) which uses seed guide tree and HMM technique to perform the alignment of series of sequence given as query and provide the output to interpret the result.

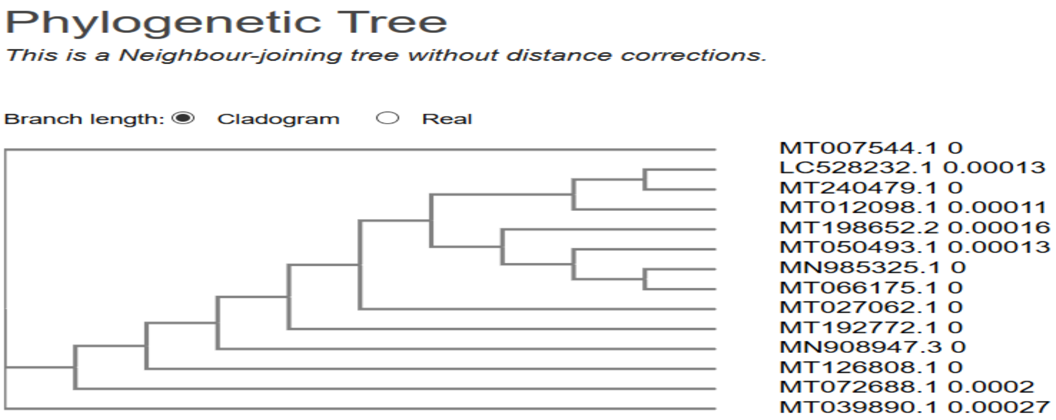
3. RESULTS & DISCUSSION

3.1 comparison of complete genome sequence if SARS COV 2 across the world

SARS COV 2 complete viral genome reported by 14 countries were downloaded and given as query sequence for Clustal Omega software to perform multiple sequence alignment and the output is observed in the form of percentage identity matrix and phylogenetic tree which was created by using neighbor joining method, from these results we can interpret the similarity between the sequence and determine the percentage of mutation occurred in viral genome.



```
 1: MT007544.1  100.00   99.93   99.97   99.97   99.97   99.97   99.96   99.97   99.98   99.98   99.99   99.99   99.98   99.98
 2: LC528232.1   99.93  100.00   99.98   99.98   99.94   99.91   99.97   99.98   99.97   99.97   99.94   99.96   99.97   99.97
 3: MT240479.1   99.97   99.98  100.00   99.96   99.96   99.95   99.95   99.96   99.97   99.97   99.98   99.98   99.97   99.97
 4: MT012098.1   99.97   99.98   99.96  100.00   99.96   99.95   99.96   99.96   99.97   99.97   99.98   99.98   99.97   99.98
 5: MT072688.1   99.97   99.94   99.96   99.96  100.00   99.95   99.95   99.96   99.97   99.97   99.98   99.97   99.97   99.97
 6: MT039890.1   99.97   99.91   99.95   99.95   99.95  100.00   99.94   99.95   99.96   99.96   99.97   99.97   99.96   99.96
 7: MT198652.2   99.96   99.97   99.95   99.96   99.95   99.94  100.00   99.97   99.97   99.96   99.97   99.97   99.98   99.98
 8: MT050493.1   99.97   99.98   99.96   99.96   99.96   99.95   99.97  100.00   99.97   99.97   99.98   99.98   99.98   99.99
 9: MT126808.1   99.98   99.97   99.97   99.97   99.97   99.96   99.97   99.97  100.00   99.98   99.99   99.98   99.98   99.98
10: MT027062.1   99.98   99.97   99.97   99.97   99.97   99.96   99.96   99.97   99.98  100.00   99.99   99.99   99.98   99.98
11: MN908947.3   99.99   99.94   99.98   99.98   99.98   99.97   99.97   99.98   99.99   99.99  100.00  100.00   99.99   99.99
12: MT192772.1   99.99   99.96   99.98   99.98   99.97   99.97   99.97   99.98   99.98   99.99  100.00  100.00   99.99   99.99
13: MN985325.1   99.98   99.97   99.97   99.97   99.97   99.96   99.98   99.98   99.98   99.98   99.99   99.99  100.00  100.00
14: MT066175.1   99.98   99.97   99.97   99.98   99.97   99.96   99.98   99.99   99.98   99.98   99.99   99.99  100.00  100.00
```



Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

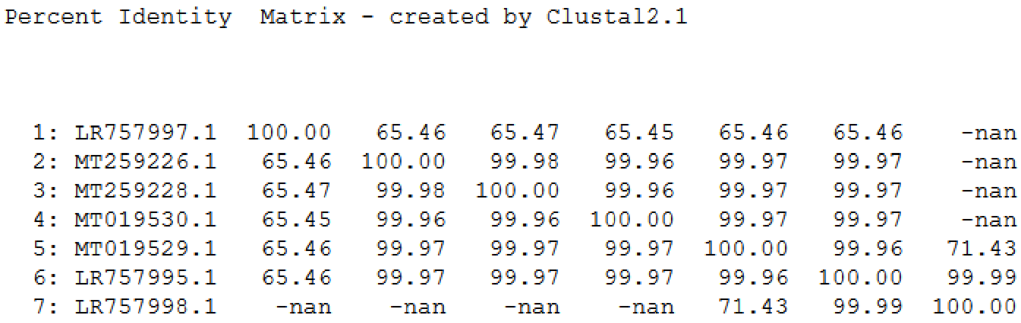
Branch length: ● Cladogram ○ Real

Figure 1: complete genome similarity identity matrix

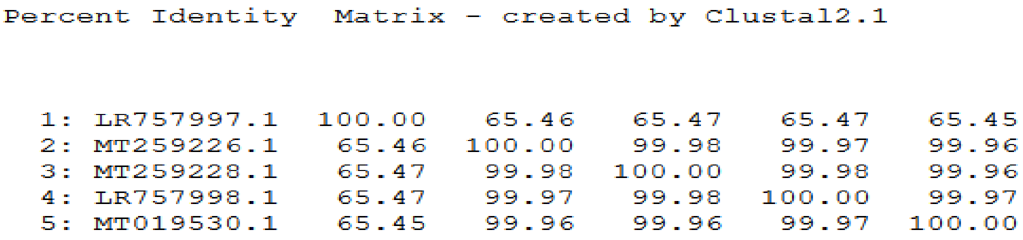
Figure 1 represents the identity matrix and phylogenetic tree obtained from clustal Omega submitting 14 query sequence reported across the world, from which we can interpret the percentage range of similarity in 14*14 identity matrix is ranging between 99.91% to 99.99 %, were 100% symbolizes absolute match which is found only between the sequence reported between Taiwan and USA, where as other sequences have variation which indicates incidences of mutation, phylogenetic tree is made my grouping the similar sequences.

3.2 Comparison of complete genome sequence of SARS COV 2 within china

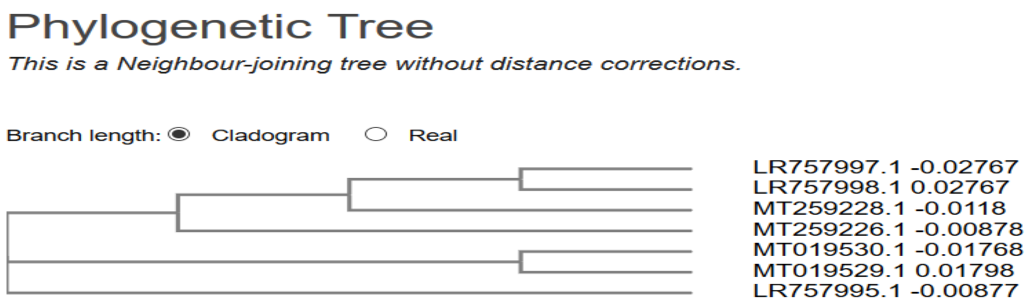
As we all know the SARS COV 2 outbreak was first seen in china-Wuhan later turning into global pandemic due to its severity and increasing the death troll of aged people, to understand the pathogenesis of the infection , genome sequencing plays a vital role, and identification of similarity between sequences is also important to support research in developing diagnostic tools and vaccine development. In our study we took 7 complete genome sequence reported by china and results revealed that there was no 100 % similarity which is a clear indication that the virus is undergoing constant mutation to increase the adaptability in host system, similarity range is seen between 65.47% to 99.99 % as given in figure 2.



```
Percent Identity  Matrix - created by Clustal2.1


  1: LR757997.1  100.00   65.46   65.47   65.45   65.46   65.46    -nan
  2: MT259226.1   65.46  100.00   99.98   99.96   99.97   99.97    -nan
  3: MT259228.1   65.47   99.98  100.00   99.96   99.97   99.97    -nan
  4: MT019530.1   65.45   99.96   99.96  100.00   99.97   99.97    -nan
  5: MT019529.1   65.46   99.97   99.97   99.97  100.00   99.96   71.43
  6: LR757995.1   65.46   99.97   99.97   99.97   99.96  100.00   99.99
  7: LR757998.1    -nan    -nan    -nan    -nan   71.43   99.99  100.00
```



```
Percent Identity  Matrix - created by Clustal2.1


  1: LR757997.1  100.00   65.46   65.47   65.47   65.45
  2: MT259226.1   65.46  100.00   99.98   99.97   99.96
  3: MT259228.1   65.47   99.98  100.00   99.98   99.96
  4: LR757998.1   65.47   99.97   99.98  100.00   99.97
  5: MT019530.1   65.45   99.96   99.96   99.97  100.00
```



Phylogenetic Tree

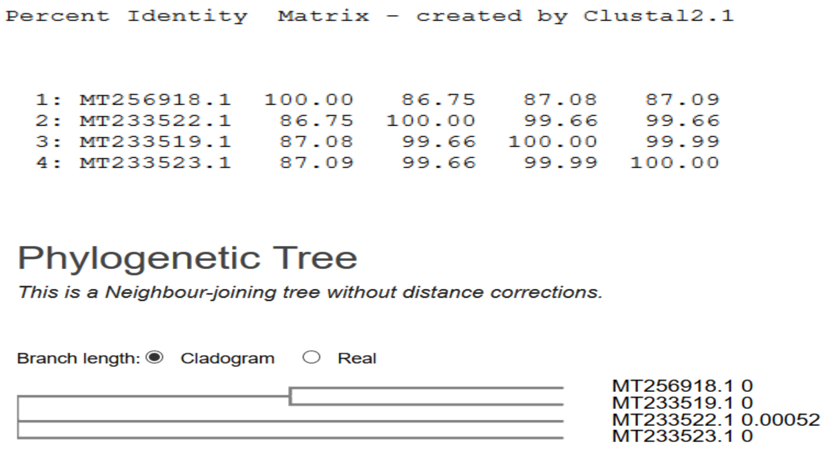
This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

Figure 2: percentage identity matrix of sequences submitted within china

3.3 comparison of complete genome sequence of SARS COV 2 within Spain

Spain is also in one of the countries which was affected with high mortality rate and fast spread of infection, Spain had reported 4 complete genome sequence, which we considered for our study and results revealed there was no 100% similarity between 4 sequences, and the range was found between 86.75%-99.99% as given in figure 3.



```
Percent Identity  Matrix - created by Clustal2.1


  1: MT256918.1   100.00    86.75    87.08    87.09
  2: MT233522.1    86.75   100.00    99.66    99.66
  3: MT233519.1    87.08    99.66   100.00    99.99
  4: MT233523.1    87.09    99.66    99.99   100.00
```

Phylogenetic Tree

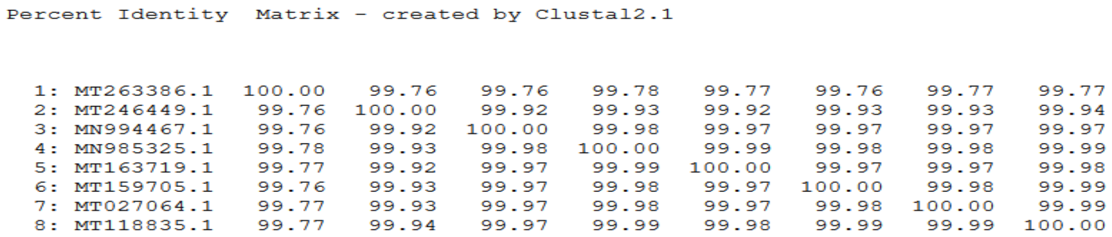
This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

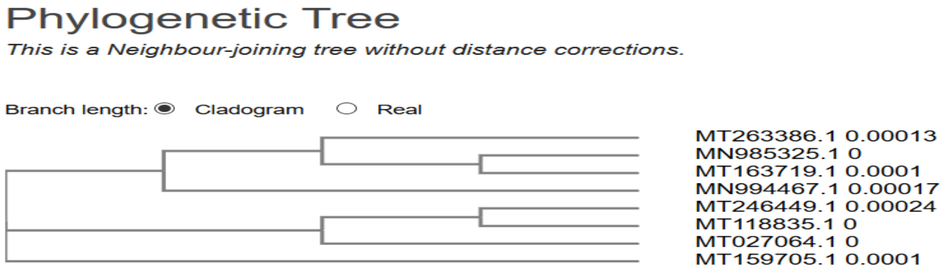
Figure 3: percentage identity matrix of sequences submitted within Spain

3.4 comparison of complete genome sequence of SARS COV 2 within USA

USA is considered as one among the country which is highly developed and it is believed to provide golden standard health care facilities, irrespective of these facility USA was not able to control the spread of COVIR-19, leading huge number of people getting infected and crossed the mortality rate higher than china, in or study we had downloaded the genome sequences reported from January 2020 to march 2020 and results obtained revealed that there was no 100% similarity and the similarity range between 8 sequences was in the range 99.77% to 99.99% as given in figure 4.



```
Percent Identity  Matrix - created by Clustal2.1


  1: MT263386.1  100.00   99.76   99.76   99.78   99.77   99.76   99.77   99.77
  2: MT246449.1   99.76  100.00   99.92   99.93   99.92   99.93   99.93   99.94
  3: MN994467.1   99.76   99.92  100.00   99.98   99.97   99.97   99.97   99.97
  4: MN985325.1   99.78   99.93   99.98  100.00   99.99   99.98   99.98   99.99
  5: MT163719.1   99.77   99.92   99.97   99.99  100.00   99.97   99.97   99.98
  6: MT159705.1   99.76   99.93   99.97   99.98   99.97  100.00   99.98   99.99
  7: MT027064.1   99.77   99.93   99.97   99.98   99.97   99.98  100.00   99.99
  8: MT118835.1   99.77   99.94   99.97   99.99   99.98   99.99   99.99  100.00
```



Phylogenetic Tree

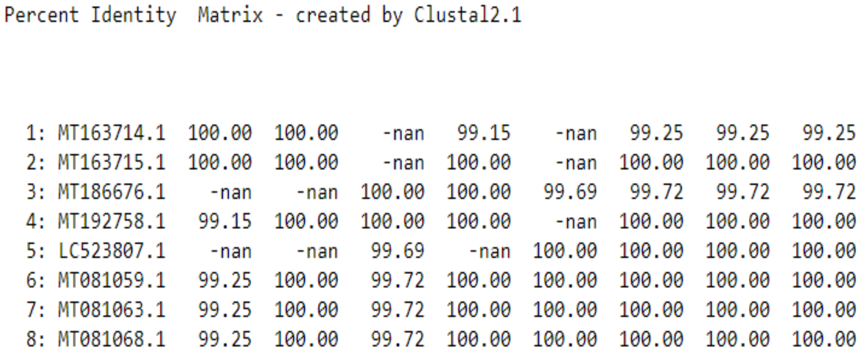
This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

Figure 4: percentage identity matrix of sequences submitted within USA

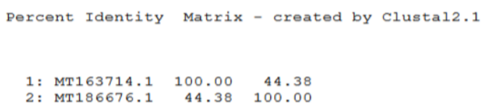
3.5 comparison of N gene sequence reported cross the world:

N gene codes protein N which plays a major role in assembling of virion during the replication of viral genome, they also interact with protein M and help in enhancing the transcription. Due to its importance , N gene Is been commonly selected as the target site for development of detection / diagnostic kit using the gene data reported in website, due to the pandemic situation the diagnostics kits developed by various countries are exchanged as a token of humanity to support the worst hit countries, if there is a mutation in the target site of N gene the kit may lack the detection leading to fetal spread of disease , to avoid this condition in our study we have taken 8 sequences of N gene reported across the world and compared by keeping threshold as 99%. The study revealed that 4 countries namely Italy, India, Iran & Philippines had very less similarity between each other and hence (-nan) was observed, apart from that the similarity of sequences ranges from 99.25% to 99.72%, and in many cases an absolute match of 100% is observed as shown in figure 5. Now our area of interest lied in the matrix region were we obtained (-nan) and separately performed the alignment to interpret the similarity range, surprisingly in case I (India vs Philippines) showed only 44.38% similar , case II (India Vs Iran) showed only 45.9% similarity, case III (Italy Vs Philippines) showed 49.17% similarity and lastly (India vs Italy) showed 49.01% similarity as shown in figure 6. Hence this result gives us the clue of rapid evolution of virus across the world.



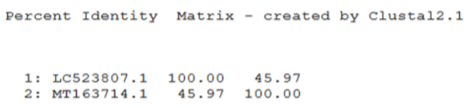
```
Percent Identity  Matrix - created by Clustal2.1



  1: MT163714.1  100.00  100.00    -nan   99.15    -nan   99.25   99.25   99.25
  2: MT163715.1  100.00  100.00    -nan  100.00    -nan  100.00  100.00  100.00
  3: MT186676.1    -nan    -nan  100.00  100.00   99.69   99.72   99.72   99.72
  4: MT192758.1   99.15  100.00  100.00  100.00    -nan  100.00  100.00  100.00
  5: LC523807.1    -nan    -nan   99.69    -nan  100.00  100.00  100.00  100.00
  6: MT081059.1   99.25  100.00   99.72  100.00  100.00  100.00  100.00  100.00
  7: MT081063.1   99.25  100.00   99.72  100.00  100.00  100.00  100.00  100.00
  8: MT081068.1   99.25  100.00   99.72  100.00  100.00  100.00  100.00  100.00
```

Figure 5: percentage identity matrix of N gene sequences submitted across the world



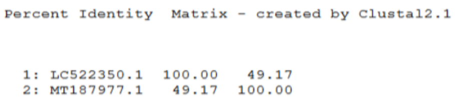
```
Percent Identity  Matrix - created by Clustal2.1


  1: MT163714.1  100.00   44.38
  2: MT186676.1   44.38  100.00
```

- India vs Philippines



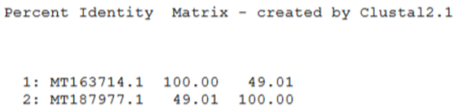
```
Percent Identity  Matrix - created by Clustal2.1


  1: LC523807.1  100.00   45.97
  2: MT163714.1   45.97  100.00
```

- India vs Iran



```
Percent Identity  Matrix - created by Clustal2.1


  1: LC522350.1  100.00   49.17
  2: MT187977.1   49.17  100.00
```

- Italy vs Philippines



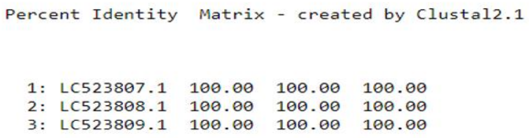
```
Percent Identity  Matrix - created by Clustal2.1


  1: MT163714.1  100.00   49.01
  2: MT187977.1   49.01  100.00
```

- India vs Italy

Figure 6: percentage identity matrix of N gene sequences submitted in selected countries

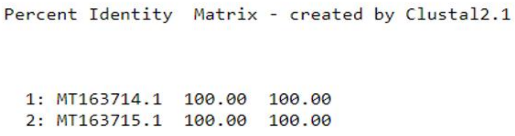
3.6 comparison of N gene sequence reported within a country:

Across the world very few countries namely India, Iran, Philippines, China & Italy, have submitted the sequence of N gene which we considered for our study, results revealed that among these countries India , Iran & Philippines have 100% similarity as shown in figure 7, where in china among the 10 sequences 8 showed 100% similarity and 2 showed a similarity of 99.92% indicating 0.08%, divergence and in other hand in Italy only 2 sequences was reported and they show only 45% similarity indicating 55% divergence as shown in figure 8, which is not acceptable if N gene is considered as target site for diagnostics and treatment.



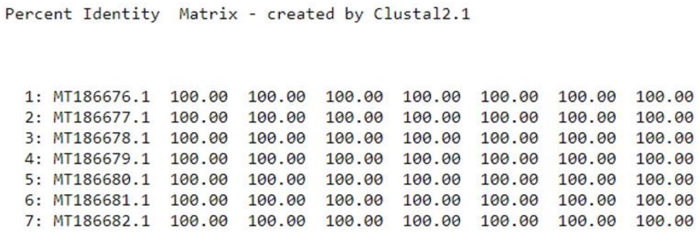
```
Percent Identity  Matrix - created by Clustal2.1


  1: LC523807.1  100.00  100.00  100.00
  2: LC523808.1  100.00  100.00  100.00
  3: LC523809.1  100.00  100.00  100.00
```

Philippines



```
Percent Identity  Matrix - created by Clustal2.1


  1: MT163714.1  100.00  100.00
  2: MT163715.1  100.00  100.00
```

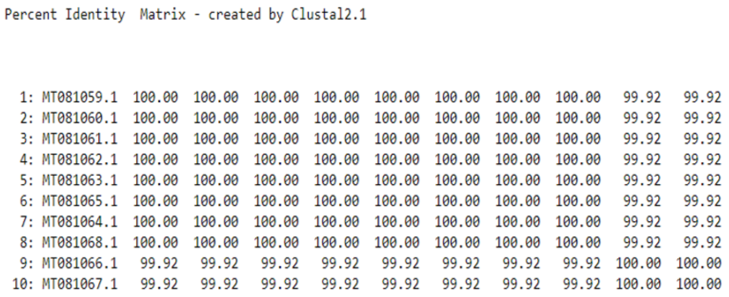
India



```
Percent Identity  Matrix - created by Clustal2.1


  1: MT186676.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  2: MT186677.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  3: MT186678.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  4: MT186679.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  5: MT186680.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  6: MT186681.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  7: MT186682.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00
```

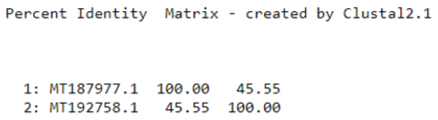
Iran

Figure 7: percentage identity matrix of N gene sequences submitted within a country



```
Percent Identity  Matrix - created by Clustal2.1


  1: MT081059.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  2: MT081060.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  3: MT081061.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  4: MT081062.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  5: MT081063.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  6: MT081065.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  7: MT081064.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  8: MT081068.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00   99.92   99.92
  9: MT081066.1   99.92   99.92   99.92   99.92   99.92   99.92   99.92   99.92  100.00  100.00
 10: MT081067.1   99.92   99.92   99.92   99.92   99.92   99.92   99.92   99.92  100.00  100.00
```

china



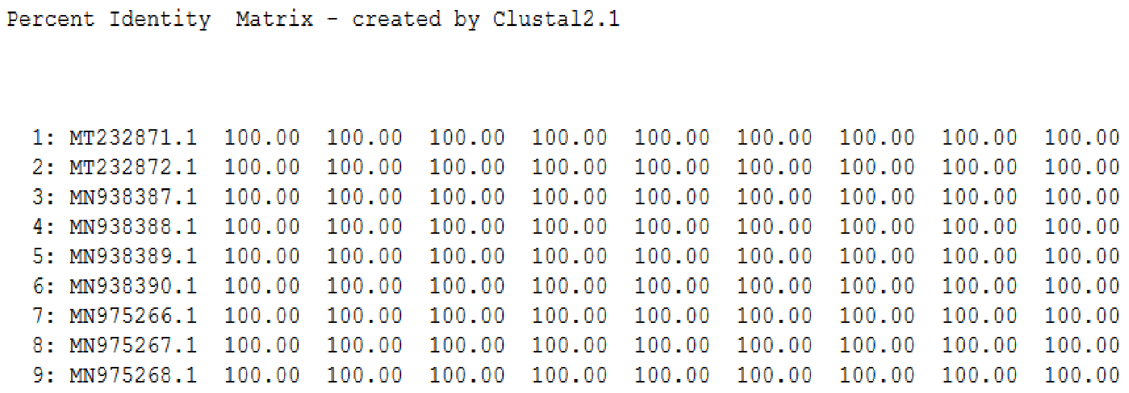
```
Percent Identity  Matrix - created by Clustal2.1


  1: MT187977.1  100.00   45.55
  2: MT192758.1   45.55  100.00
```

Italy

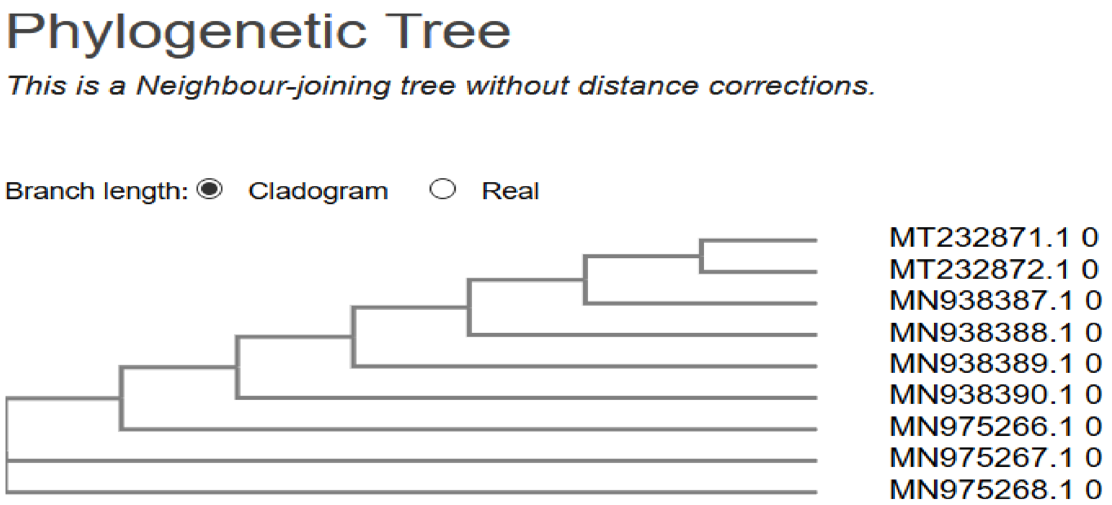
Figure 8: percentage identity matrix of N gene sequences submitted within a country in Italy & china

3.7 Comparison of S gene sequence reported across the world:

S gene in SARS COV 2 code for S protein called as the spike glycoprotein which plays a major role in binding the virion to the host cell using the receptor ligand interaction to initiate the infection, since this protein plays a major role in manifestation of infection it is considered one among the list of important target for diagnostics and drug development. In NCBI data base the genome of S gene was reported only by 3 countries as of 01-04-2020 ,namely china-Wuhan, china-Shenzhen and Iran , in our study we had considered 8 sequences reported by these 3 countries and found that there was 100% similarity as shown in figure 9, indicating that there is no mutation occurred and hence can be selected as one of the major target for diagnostics and drug development. We also performed the comparison study within the country to conform the similarity and found 100% result as shown in figure 10.



```
Percent Identity  Matrix - created by Clustal2.1


  1: MT232871.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  2: MT232872.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  3: MN938387.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  4: MN938388.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  5: MN938389.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  6: MN938390.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  7: MN975266.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  8: MN975267.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  9: MN975268.1  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
```

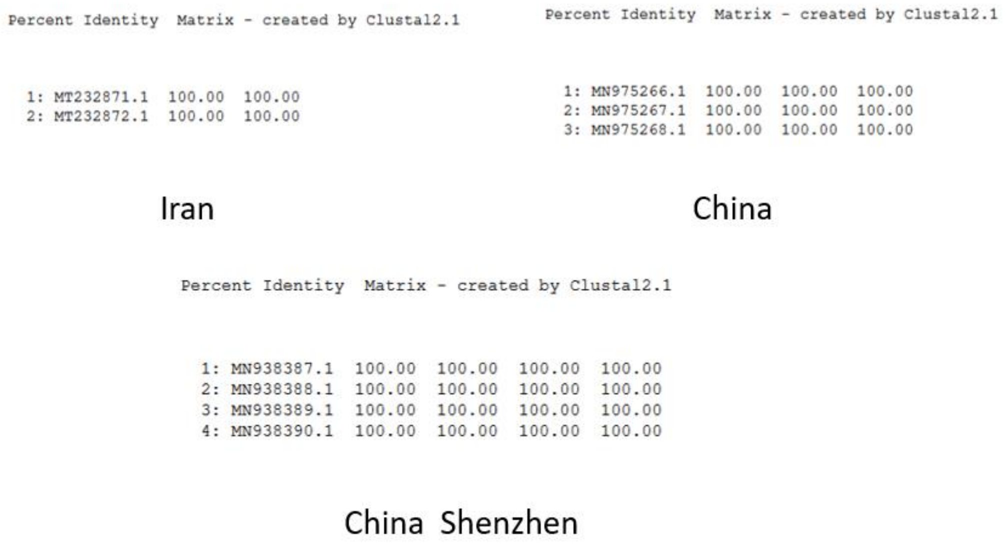


Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

Figure 9: percentage identity matrix of S gene sequences submitted across the world



```
Percent Identity  Matrix - created by Clustal2.1


 1: MT232871.1  100.00  100.00
 2: MT232872.1  100.00  100.00
```

Iran

```
Percent Identity  Matrix - created by Clustal2.1


 1: MN975266.1  100.00  100.00  100.00
 2: MN975267.1  100.00  100.00  100.00
 3: MN975268.1  100.00  100.00  100.00
```

China

```
Percent Identity  Matrix - created by Clustal2.1


 1: MN938387.1  100.00  100.00  100.00  100.00
 2: MN938388.1  100.00  100.00  100.00  100.00
 3: MN938389.1  100.00  100.00  100.00  100.00
 4: MN938390.1  100.00  100.00  100.00  100.00
```

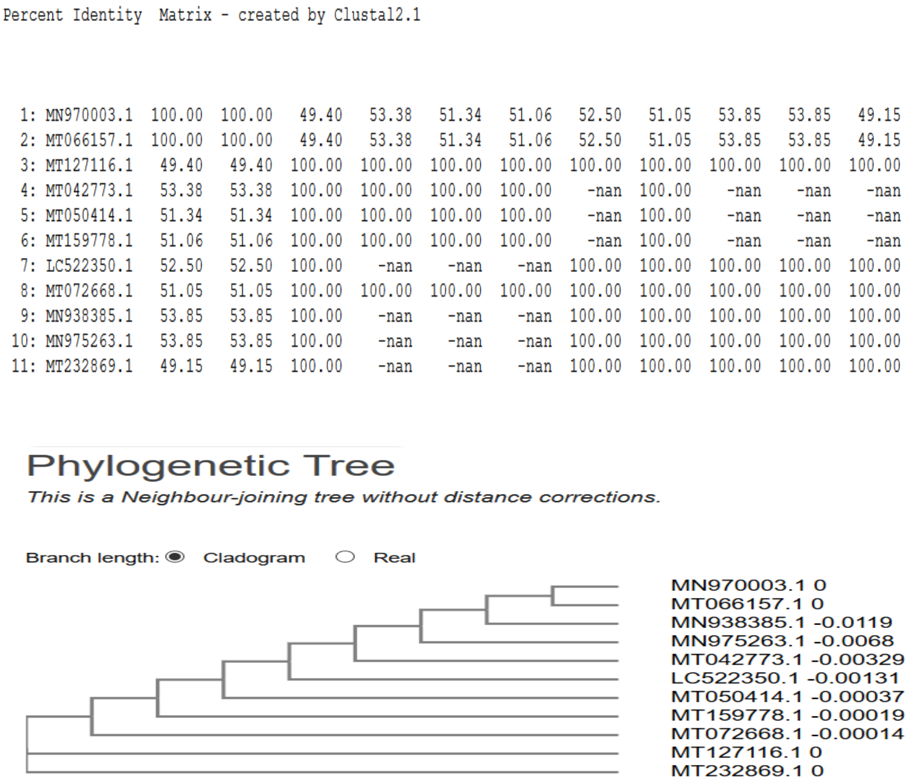
China Shenzhen

Figure 10: percentage identity matrix of S gene sequences submitted within a country

3.8 Comparison of RdRP gene sequence reported across the world:

RdRP gene codes for RNA dependent RNA polymerase enzyme which plays a important role in replication of RNA template, the protein coded by this gene is present in all RNA viruses , since it plays a vital role in replication of viral genome in host it is used as a target for research in diagnostics and vaccine development and may also help in development of drugs to control the viral load in host system during the manifestation of the infection. In our study we collected 11 sequences and results revealed that there was very less similarity between the sequences due to mutation. As shown in figure 11 we can understand that the range of similarity lies between 49.15% to 54% and in few cases 100% similarity was observed. In some places (-nan) is obtained as a result because it lies below the threshold value and hence further to find the similarity we again performed the alignment of those selected sequences in category of 3 cases , In case 1 we considered 4 countries namely (Philippines, Wuhan, Australia , Nigeria) and we found that gene sequence reported by Philippines was having divergence of 48.18%, 51.69% &

47.95% and rest other 3 sequence was having 100% similarity respectively . In case II we considered 4 countries namely (china, Wuhan, Australia & Nigeria) and we found that sequence reported by china was having divergence of 52.27%, 48.94% and 51% and other 3 showed 100% similarity respectively. In case III we considered 4 countries namely (Iran, Wuhan, Australia & Nigeria) and we found that sequence reported in Iran was having divergence of 47.37%, 49.25% and 48.60% and other 3 sequences showed 100% similarity respectively as shown in figure 12.



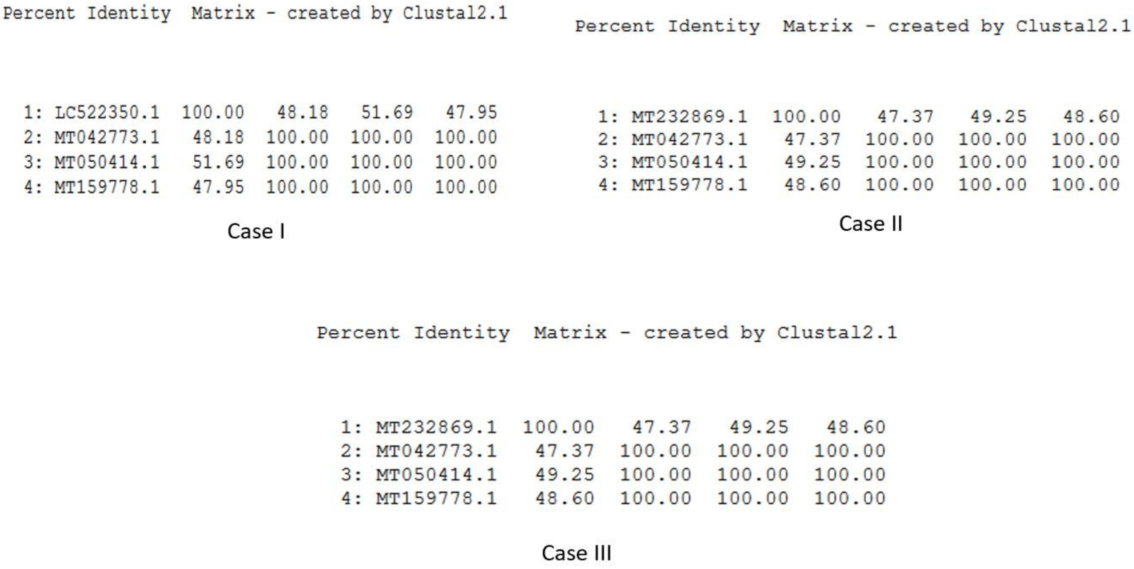
```
Percent Identity  Matrix - created by Clustal2.1


  1: MN970003.1  100.00  100.00   49.40   53.38   51.34   51.06   52.50   51.05   53.85   53.85   49.15
  2: MT066157.1  100.00  100.00   49.40   53.38   51.34   51.06   52.50   51.05   53.85   53.85   49.15
  3: MT127116.1   49.40   49.40  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  4: MT042773.1   53.38   53.38  100.00  100.00  100.00  100.00    -nan  100.00    -nan    -nan    -nan
  5: MT050414.1   51.34   51.34  100.00  100.00  100.00  100.00    -nan  100.00    -nan    -nan    -nan
  6: MT159778.1   51.06   51.06  100.00  100.00  100.00  100.00    -nan  100.00    -nan    -nan    -nan
  7: LC522350.1   52.50   52.50  100.00    -nan    -nan    -nan  100.00  100.00  100.00  100.00  100.00
  8: MT072668.1   51.05   51.05  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00  100.00
  9: MN938385.1   53.85   53.85  100.00    -nan    -nan    -nan  100.00  100.00  100.00  100.00  100.00
 10: MN975263.1   53.85   53.85  100.00    -nan    -nan    -nan  100.00  100.00  100.00  100.00  100.00
 11: MT232869.1   49.15   49.15  100.00    -nan    -nan    -nan  100.00  100.00  100.00  100.00  100.00
```

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

Figure 11: percentage identity matrix of RdRP gene sequences submitted across the world



```
Percent Identity  Matrix - created by Clustal2.1


 1: LC522350.1  100.00   48.18   51.69   47.95
 2: MT042773.1   48.18  100.00  100.00  100.00
 3: MT050414.1   51.69  100.00  100.00  100.00
 4: MT159778.1   47.95  100.00  100.00  100.00
```

Case I

```
Percent Identity  Matrix - created by Clustal2.1


 1: MT232869.1  100.00   47.37   49.25   48.60
 2: MT042773.1   47.37  100.00  100.00  100.00
 3: MT050414.1   49.25  100.00  100.00  100.00
 4: MT159778.1   48.60  100.00  100.00  100.00
```

Case II

```
Percent Identity  Matrix - created by Clustal2.1


 1: MT232869.1  100.00   47.37   49.25   48.60
 2: MT042773.1   47.37  100.00  100.00  100.00
 3: MT050414.1   49.25  100.00  100.00  100.00
 4: MT159778.1   48.60  100.00  100.00  100.00
```

Case III

Figure 12: percentage identity matrix of RdRP gene sequences selected in specific cases

4. Conclusion

After carrying out various combination of multiple sequence alignment we got a brief the similarity between the viral genome across the world and as well within a specific geographic location. This data generated by our research can give a brief on idea on sites on genome where the mutation has occurred. Our research gave a brief on the similarity between the complete viral genome, N gene sequence, S gene sequence & RdRP gene sequence. As per the results obtained we can conclude that S gene can be a ideal target because it shows 100% similarity according to the data recorded till 1st April 2020, we can interpret that This data provided by can be handful for the researchers working on drug target selection and target selection for developing universal diagnostic kit for SARS COV 2.

5. Future Scope

We have carried out the study with limited data available in NCBI data base, As we all know that RNA viruses undergo mutation at higher speed compared to DNA virus, Increase in sequencing of genome is recommended to understand the rate of change in virus, details regarding the manifestation of infection & to understand the variants present within a population.

Conflicts of interest: The authors stated that no conflicts of interest.

6. References

1. uang C et al, *2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395:497–506. doi:10.1016/S0140-6736(20)30183-5*
2. Rodriguez-Morales AJ et al, *2020. Going global: travel and the 2019 novel coronavirus. Travel Med Infect Dis 33:101578. doi:10.1016/j.tmaid.2020.101578.*
3. Rodriguez-Morales AJ et al , *2020. History is repeating itself, a probable zoonotic spillover as a cause of an epidemic: the case of 2019 novel coronavirus. Infez Med 28:3–5*
4. Rodriguez-Morales AJ et al, *2020. History is repeating itself, a probable zoonotic spillover as a cause of an epidemic: the case of 2019 novel coronavirus. Infez Med 28:3–5*
5. Chu DK et al, *2020. Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia. Clin Chem. doi:10.1093/clinchem/hvaa029.*
6. Snijder E.J., Bredenbeek P.J., Dobbe J.C., Thiel V., Ziebuhr J., Poon L.L., Guan Y., Rozanov M., Spaan W.J., Gorbalenya A.E. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. 2003;331:991–1004. doi: 10.1016/S0022-2836(03)00865-9.
7. Thiel V., Ivanov K.A., Putics A., Hertzig T., Schelle B., Bayer S., Weissbrich B., Snijder E.J., Rabenau H., Doerr H.W., et al. Mechanisms and enzymes involved in SARS coronavirus genome expression. J. Gen. Virol. 2003;84 Pt 9:2305–2315.
8. Laude H., Masters P. Coronaviruses and Arteriviruses. Plenum Press; New York, NY, USA: 1995. The coronavirus nucleocapsid protein; pp. 141–163.
9. Parker M.M., Masters P.S. Sequence comparison of the N genes of five strains of the coronavirus mouse hepatitis virus suggests a three domain structure for the nucleocapsid protein. Virology. 1990;179:463–468.
10. Huang Q., Yu L., Petros A.M., Gunasekera A., Liu Z., Xu N., Hajduk P., Mack J., Fesik S.W., Olejniczak E.T. Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein. Biochemistry. 2004;43:6059–6063.
11. Raj VS, Mou H, Smits SL, Dekkers DH, Muller MA, Dijkman R, Muth D, Demmers JA, Zaki A, Fouchier RA, Thiel V, Drosten C, Rottier PJ, et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature. 2013;495:251–4.
12. Lu G, Hu Y, Wang Q, Qi J, Gao F, Li Y, Zhang Y, Zhang W, Yuan Y, Bao J, Zhang B, Shi Y, Yan J, et al. Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. Nature. 2013;500:227–31.
13. Mou H, Raj VS, van Kuppeveld FJ, Rottier PJ, Haagmans BL, Bosch BJ. The receptor binding domain of the new Middle East respiratory syndrome coronavirus maps to a 231-residue region in the spike protein that efficiently elicits neutralizing antibodies. J Virol. 2013;87:9379–83.

14. Lu L, Liu Q, Zhu Y, Chan KH, Qin L, Li Y, Wang Q, Chan JF, Du L, Yu F, Ma C, Ye S, Yuen KY, et al. Structure-based discovery of Middle East respiratory syndrome coronavirus fusion inhibitor. Nat Commun. 2014;5:3067.

15. Elena S.F., Sanjuan R. Adaptive value of high mutation rates of RNA viruses: Separating causes from consequences. J. Virol. 2005;79:11555–11558. doi: 10.1128/JVI.79.18.11555-11558.2005.

16. Crotty S., Cameron C.E., Andino R. RNA virus error catastrophe: Direct molecular test by using ribavirin. Proc. Natl. Acad. Sci. USA. 2001;98:6895–6900.

17. Lai M.M. Cellular factors in the transcription and replication of viral RNA genomes: A parallel to DNA-dependent RNA transcription. Virology. 1998;244:1–12.

18. Baer CF. Does mutation rate depend on itself. PLoS Biol. 2008;6(2):e52 10.1371/journal.pbio.0060052

19. Loewe L, Hill WG. The populations of mutations: good, bad and indifferent. Philosophical transactions of the Royal Society of London. 2010;365(1544):1153–67.

20. Gago S, Elena SF, Flores R, Sanjuan R. Extremely high mutation rate of a hammerhead viroid. Science. 2009;323(5919):1308 10.1126/science.1169202 .