SEQUENCE ANALYSIS FOR SNP DETECTION AND PHYLOGENETIC RECONSTRUCTION OF SARS-CoV-2 SEQUENCES ISOLATED FROM DIFFERENT COVID-19 SEQUENCES

Dissertation submitted in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE
IN
BIOTECHNOLOGY

By

Raj Laxmi Singh
Enrollment No. 217806

Under the Supervision of

Dr. Shikha Mittal

(Assistant Professor)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY WAKNAGHAT,

DEPARTMENT OF BIOTECHNOLOGY & BIOINFORMATICS

SOLAN-173234, HIMACHAL PRADESH

DECLARATION

I hereby declare that work reported in the M.Sc. project entitled "**Sequence analysis for SNP detection & phylogenetic reconstruction of SARS-CoV-2 isolated from different COVID-19 cases**" submitted at **Jaypee University of Information Technology, Waknaghat, India,** is an authentic record of my work carried out under the supervision of **Dr. Shikha Mittal**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my M.Sc. Project report.

Raj Laxmi Singh **Date:**
Enrollment Number-217806
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, India - 173234

CERTIFICATE

This is to certify that the work reported in the **M.Sc.** project report "**Sequence analysis for SNP detection & phylogenetic reconstruction of SARS-CoV-2 isolated from different COVID-19 cases"** submitted by **Ms. Raj Laxmi Singh** at **Jaypee University of Information Technology, Waknaghat, India**, is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Supervisor:

Dr. Shikha Mittal **Date:**

Assistant Professor
Department of Biotechnology &Bioinformatics
Jaypee University of Information Technology
Waknaghat, India-173234

ACKNOWLEDGMENT

I would like to express my profound gratitude to my guide **Dr. Shikha Mittal** for her guidance, support and constant encouragement throughout the course of this project work. She has been more than just my project guide; at times a mentor to rescue me out of my doubts. She has always helped me to work hard and also taught me how to implement different ideas to deal with the problem. Moreover, she taught me to not give up and many other valuable lessons.

I also want to mention the HOD of Biotechnology and Bioinformatics **Prof. (Dr.) Sudhir Kumar** has been a source of immense motivation and inspiration both for my academic and personal life. He was never, and I know will never be, more than just a phone call away. He has helped me in almost every aspect I have asked him for.

In addition, I would like to thank all the faculty members of the BT/BI Department of JUIT, who have helped me whenever I needed and also would like to thank all the lab engineers and specially **Ms. Somlata Sharma** for providing me with a workplace and for always motivating me. I would also like to appreciate the part that my friends have played in shaping this project work. They have been my constant support and cheered me up at hard times. They helped me whenever I had any doubts. Thanks a lot!I would like to thank the almighty God for his grace throughout my life. Last but not the least I would like to thank my Mother and Father who have always supported me through thick and thin and have been a constant source of encouragement and support; also, who has never given up on me and always motivated me.

Raj laxmi singh
M.Sc. biotechnology
JUIT, SOLAN

TABLE OF CONTENTS

S.No.	**Name**	**Page no.**
1.	Abstract	IX
2.	Chapter 1: Introduction	1-6
3.	Chapter 2: Review of Literature	7-17
4.	Chapter 3 : Materials & Methods	18-25
5.	Chapter 4: Results & Discussion	26-32
6.	Chapter 5: Conclusion	33
7.	Future perspectives	34
8.	References	35-41

List of Figures

Figure	Title	Page no.
Figure 2.2.1	Human Corona Virus	9
Figure 2.3.1	Genetic materials within a virus genome	10
Figure 2.4.1	SARS-CoV-2 life cycle	12
Figure 2.8.1	Main genes involved in SARS-CoV-2	14
Figure 2.9.1	Main SNPs involved in SARS-CoV-2	14
Figure 3.1.1.1	Out of 4082 viruses selected 30 covid-19 cases	19
Figure 3.2.1.1	Flowchart representation methodology of the analysis	20
Figure 3.2.1.2	Sequences were aligned using MAFT	22
Figure 3.2.1.3	Aligned sequences trimmed at 3’ & 5’ end to obtain homology	23
Figure 3.2.1.4	Sequences re-aligned using MAFT	24
Figure 3.2.1.5	SNP & variation analysis was done by DnaSP	25
Figure 3.2.1.6	Phylogeetic re-construction of SARS-CoV-2 genome sequences	26
Figure 4.1.3.1	SNPs in linkage diseuillibrium	30
Figure 4.1.5.1	Phylogenetic trees of SARS-CoV-2 genomes	34

List of tables

Table	Title	Page number
2.9.1	List of genes found to be involved in SARS-CoV-2 & their functions	15
3.1.1.1	Have taken sequences from different places	19
4.1.1.1	COVID-19 cases, deaths till 03rd Feb 2023	27
4.1.1.2	The sequence id's of all genome sequences	28
4.1.2.1	Large conserved genomic regions	29
4.1.3.1	Haplotypes number obtained	30
4.1.4.1	Lineage analysis list of all 30 SARS-CoV-2 genome sequences	31-33

List of Abbreviations

SNP: Single Nucleotide Polyorphism

GSTO2: Glutathione-S-transferase Omega 2

COPD: Chronic Obstructive Pulmonary Disease

FEVR: Familial Exudative Viteroretinopathy

GRMs: Metabotropic Glutamate Receptors

SSAHA: Sequence Search and Alignment by Hashing Algorithm

PCR: Polymerase Chain Reactions

ACE-2 : Angiotensin-converting enzyme 2

TMPRSS2 : Transmembrane protease serine 2

SARS-CoV-2: Severe Acute Respiratory Syndrome- Corona Virus Disease-2

OM: Omicron

ABSTRACT

The brand-new coronavirus SARS-CoV-2 is what caused the COVID-19 pandemic. SARS-CoV-2 accesses host cells via the Angiotensin-converting enzyme 2 (ACE2), which is also a functional receptor on cell surfaces. ACE2 is abundantly expressed in the heart, kidneys, and lungs and is released into the plasma. The rennin angiotensin aldosterone system's main regulator is ACE2 (RAAS). Specifically in individuals with comorbidities such hypertension, Diabetes mellitus, and cardiovascular illness, SARS-CoV-2 promotes ACE/ACE2 balance disturbance and RAAS activation, which ultimately leads to COVID-19 development. As a Result, ACE2 expression may have contradictory effects, promoting SARS-CoV-2 pathogenicity while inhibiting viral infection. In reviewing the present research and understanding of ACE2 in the milieu of COVID-19, Assessing the burgeoning and broaden of infections has been aided by phyloepidemiological techniques. Awareness of the pandemic and dissemination of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in COVID patients isolated from various locations will aid in the establishment of preventative strategies for reducing infection amongst vulnerable groups. The goal of the research was to look at the development of SARS-CoV-2 in Asia, Europe, and North America. On February 3, 2023, 30 full genomes of SARS-CoV-2 were obtained from the GISAID database in order to analyse its evolution across Asia, Europe, and North America. The sequences were selected based on the person's travel history and the date of collection. Other sequences were not chosen since they were too short, featured artefacts that were not from a firsthand source, or had insufficient data.

CHAPTER-1

1.1 INTRODUCTION

Whenever one nucleotide in the sequencing of the gene is altered, genetic changes such as SNIP in a DNA sequence occur. SNIP can appear in both the coding (gene) and non-coding area of the genomic sequence. Numerous SNPs have little impact on cellular activity but they can increase an organism's susceptibility to disease or influence how it responds to treatments [1].

This means that studying SNPs in the *Homo sapiens* material can offer a basic knowledge of several caused by mutations disease hence presenting novel therapy attacks and also SNIP have the capacity to cause diseases but they can also identify who can contract certain illness. One such potential genes for causing COPD. It was the nsSNIP engender the GSTO2 haplotype ASN 142 ASP polymorphism [2].

SNPs and INDELS (insertion/deletion) markers are suitable for examining the genome of both animals and plants by supporting chromosome modeling biological variation and serving as a very important part of genetically improvement initiatives.

Molecular markers are useful equipments of analyzing of genome sequences plus their affiliation of inherited characteristics. In the company of cardinal inherent variations the development of genetic techniques markers has improved fast after a considerable ten years escorted by introduction able to increased genomics techniques. Due to the development of high throughput technologies allowing their identification [3].

The use of SNPs and minor INDELs as molecular markers has undergone transformation and also the proportional contributions of single nucleotide polymorphisms & INDELS (insertions/deletions) to the probability of complicated illness development in *Homo sapiens* are unknown. A SNP affects only one nucleotide but in the case of indels (deletions and insertions of nucleotides) happens [4].

Single nucleotide polymosrphisms (SNPs) and INDELS (insertions/deletions) alterations as well as their factors leading to such likelihood for different clinical development in *Homo sapiens* are not known. Whereas, INDELS add or remove one or more nucleotides from the DNA sequence. A SNP only alters only one nucleotide [5].

Additionally in frame INDELS (coding area insertions or deletions of 3 or more base pairs) can lead to transformed proteins. Like SNPs, INDELS can also have an impact on chromatin structure the affinity of a binding site for a regulatory element or the transcriptional elements in non-coding regions.

The proportion seen between frequency the alteration and also the chosen restriction for INDELS should be maintained by maintaining reading frames with in coding sequence in order to keep the cellular functions and generally speaking regulating choice is applied to encoding INDELS as opposed to SNP’s [6].

1.2 Importance of SNP

On locating disease causing genes SNPs are advantageous in 2 aspects, first reason is that some SNP alleles are real DNA sequence changes which changes how genes were activated and regulated resulting in a significant affect here on onset major conditions such as Norrie sickness Retinopathy of prematurity and familial Exudative Viteroretinopathy (FEVR).

Cystic fibrosis and schizophrenia are caused by a similar SNP in metabotropic glutamate receptors (GRMs). As opposed to the first case the second one involves SNP alleles that indirectly and most likely insignificantly contribute to diseases like diabetes cancer & alzheimers [7].

By enabling drug action by allowing it to pass through the blood-brain barrier. P-glycoprtein is the subset of ABC transporter proteins and it has synonymous SNP which can alter the transportes behavior and where it works once drugs are absorbed.

These have advantages become recognizable biomarkers can b used to locate the functional SNIPs because there are connections with respect to the functional SNPIs and marker SNIPs [8].

A proteins structure regulation and expression could all be changed by various kinds of SNPs & the much more common type of SNPs were non synonymous SNPs in which the allelomorph differ in the amino acid which the protein product contains.

Some SNPs have been shown to change how proteins are produced as well as regulated which has an immediate impact on how well the proteins work. With the promoter regions several SNPs can be detected. SNP in coding areass and that are important genetic markers are required to find causal changes throughout potential genomes that livestock that are considerable [9].

Non-sense SNPs can swap out its anticodon sequence of proteins seem to be within foremost cause for the phenotypical variance within those SNPs essential attributes. Countless SNP within those domains might well be connected to a disease or other phenotype when they are associated with one another through genetic linkage [10].

1.3. Computational elements of SNIP findings: SNIP extracting

The two types of data that can be used for SNP mining are de novo and reference sequencing information. SNIP extraction of the different sequencing information involves within account of bearing footsteps:

(a) Sequencing reading is initially into groups based on how similar their sequences are in order to find reads that cover within unvaried region of the genomes or have within unvaried transcripts emergence.
(b) After within reds have been aligned
(c) Sequence variants were discovered and categorized for candidate **conglomeration.**

The collected information regarding completely sequencing organisms make up within group of referenced sequencing information. The sequence data denoting equivalence *Homo sapiens* genome as well as the genomes of other microbiological Organisms occur were two blueprint denoting equivalence constantly growing referenced groups [11].

The accessibility for cutting-edge equipment facilitates their arrangement procedure wherein their reference group’s information increases rapidly. A sequence data that matches to an imperfect genome may also be utilized in conjunction with the reference sets.

Sequence data collected any organisms where a reference genome are accessible must be append ahead of referenced set using some homology search gadgets. To perform this function a local or global alignment tool like BLAST if not the sequence analysis and alignment using hashing algorithm (SSAHA) could be used [12].

Using Polymerase Chain Reaction (PCR) products conversely the primers were created on the side of some specific sequenced area a separate set of reference data might well be generated. Short Oligo-nucleotide Alignment Program (SOAP) Mapping and Assembly with Qualities (MAQ) are two distinct tools worn the same as mapped the referenced information [13].

Technology programmes like as Phrap and CAP3 are routinely used to assemble the sequences into contigs. Within sequenced changes for every location being portrayed by many more readings. When a species has accessible additional sequencing reads for any particular genome area basics likelihood for discovering any polymorphic increases. Additionally whenever the sequencing mistake is detected a sequenced variation (allele) could be identified backed owing to a large number of reads. The more readings per allele the more likely it is that an allele is a true polymorphism [14].

The process requires more reads thus it takes longer. Specialized methods like d2cluster and TGICL the past created towards achieving a starting separation groupings of same sequenced segments that have existence then furthermore divided different groups with separate origination. Clustering results within every cluster must owing to be treated in order properly synchronize each and every readings contained therein. It is easy to compare the nucleotides from numerous readings that almost all align at the same place on the gene or genome. The fragments cannot be adequately aligned [15].

They are divided into two clusters because they are not all part of one. Individual readings must first be sorted within synchronized same group before the polymorphic similarity process can detect changes in the alignment and apply any matrix system. The design up to modern SNIP finding techniques frequently enables their integration with currently available genome analysis Programmes like the PHRED/PHRAP/CONSED [16].

The level of complexity of the procedure heavily influences the kind of technology required for SNP mining. In just a few hundred sequences a normal workstation is perfectly capable of looking about SNIPs in specific small areas of the genome (up to 100–150 kb). Extracting SNIPs throughout the genomes initiatives frequently necessary server-class processors and accessibility with or hundreds of cloud hosting in megabytes of information particularly unless extraction process's intermediary phases were recorded & the outcomes were logged in a database. Woefully there is no expound pennant data exchange format for either sequence multiple alignments or SNP markup details [17].

Various SNIP extraction equipments now in use require accretion and deliverables in certain file formats. These situations make use of custom scripts to translate data across different tools.

1.3. OBJECTIVES :

(i) Sequence interpretation of different COVID population samples for SNP discovery.

(ii) SARS-CoV-2 phylogenetic reconstruction from various COVID-19 patients.

CHAPTER-2

2.1 REVIEW OF LITERATURE

The deadly corona viruses disease epidemic 2019 (COVID-19) created a big risk to public health worldwide in 2019. On December 31 2019 at Wuhan china revealed the virus's initial prevalence. According to the news report from the "Johns Hopkins University" (University Corona virus Resource Centre) there had been over 144 million COVID-19 disease cases internationally and over 3 millions mortality as of January 2021. The COVD-19 pandemic has been researched from a number of angles & medical professionals are working hard to contain it. Given the possible consequences for COVID-19 infection's consequences severe acute respiratory symptoms should be avoided. The SARS-CoV-2 coronavirus is significant. It has been discovered that some gene expressions are very strongly associated until SNIP linked with coronavirus transmission prevalence & presence differ in the midst of expressions may be more susceptible to COVID-19 infection [18].

However it is probable that ethnicity plays a role in how serious SARS-CoV-2 infection's are. Although this same pathogen initially appeared in East Asia populations in Europe have been found to have considerably greater rates of morbidity and mortality. Therefore it's critical to understand the process underlying potential connection among both harshness as well as race and COVID-19 harshness [19].

Bats organically host and mould coronaviruses. In fact it is being proposed the fact that the majority of coronaviruses in humans originate through the bat reservoir. A number of experimenters have recently demonstrated an evolutionary resemblance between SARS-CoV-2 and a bat betacoronavirus of the subgenus Sarbecovirus. The new pathogen's entire-genome sequence is ninety six percentage indistinguishable to that of a bat SARS-related coronavirus (SARSr-CoV RaTG13) obtained in Yunnan province China but has little resemblance to SARS-CoV (approximately seventy nine percentage) or MERS-CoV (around fifty percent). This has additionally been proven how the SARS-CoV-2 virus makes use of utilises the SARS-CoV uses the indistinguishable receptor is the angiotensin converting enzyme II (ACE2) [20].

Although the precise path of spread from natural cenote to people to humans is unknown Multiple research investigators have demonstrated that pangolins may have SARS-CoV-2 a partial spike gene the essential functional regions in SARS-CoV2 spike proteins discovered in a virus obtained from a pangolin are very similar.

Despite these new breakthroughs, a number of basic difficulties Concerns about the evolutionary trends and causes underlying the SARS-CoV-2 pandemic remained unsolved. Researchers investigated the extent of genetic difference among SARS-CoV-2 and different corona viruses and did population genetic analyses on 30 SARS-CoV-2 -sequencing genomes [21].

2.2 Human Corona Virus

Seven human coronaviruses (HCoVs) have so far been discovered **(fig 2.2.1)**. Some of those are fewer prevalent and generate relatively minor respiratory tract infections in individuals who are healthy. They do-however-account for 1/3rd of typical viral infections and in high-threat individuals in the company of weakened immune systems are capable to result in long lived-deadly diseases. The remaining three viruses (the above mentioned responsible for MERS-SARS and COVID-19 cases) have been shown to result in higher serious disease-including a lack of breathing and mortality are both possible outcomes. COVID-19 sickness is less severe than SARS and MERS yet more lethal than Ebola which is caused by the four most common coronavirus. Because this virus is brand novel nobody is immune to it. As a result- it has the possibility of contaminate an important amount of individuals. Despite the fact that the percentage of severely severe serious incidents is low- a tiny proportion of an extremely large number count up to a large number of people suffering from an acute illness. Each of the 7 human coronaviruses is known to have been disseminating to humans from other animals [22].

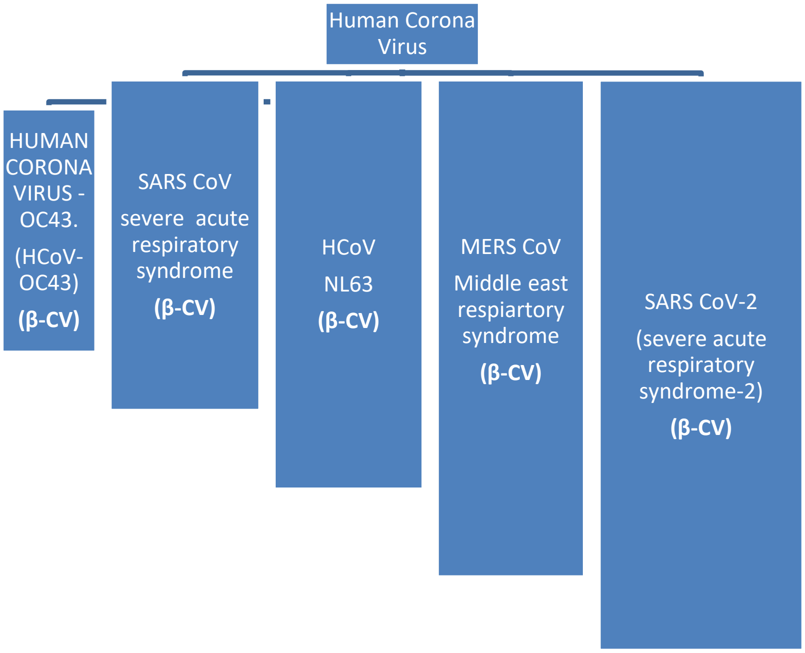

Human Corona Virus
HUMAN CORONA VIRUS - OC43.
(HCoV-OC43)
(β-CV)
SARS CoV
severe acute respiratory syndrome
(β-CV)
HCoV
NL63
(β-CV)
MERS CoV
Middle east respiartory syndrome
(β-CV)
SARS CoV-2
(severe acute respiratory syndrome-2)
(β-CV)

Figure 2.2.1 Human Corona Virus

Cross-species jump: SARS CoV-2(2002-2004) in 2002, virus-carrying horseshoe bats leaped on people, causing us to contract SARS for the first time and it was first reported in Netherland in 2002. That's why when during 2019 SARS CoV-2 were discovered at Wuhan China and had been given that designation.

2.3 What do coronaviruses look like?

Coronaviruses have basic structures that assist in helping us to comprehend how they act. They're round and counterbalance in protein spikes. These spikes help the infectious agent attach and then enter cells that are healthy. The similar spikes, however, are what allow the immune system to 'see' the infection. To promote the body's creation of antibodies against the newly discovered virus, fragments of the spike might be included in future coronavirus vaccinations. Whenever observed with a strong microscope, their spikes resemble a crown [23].

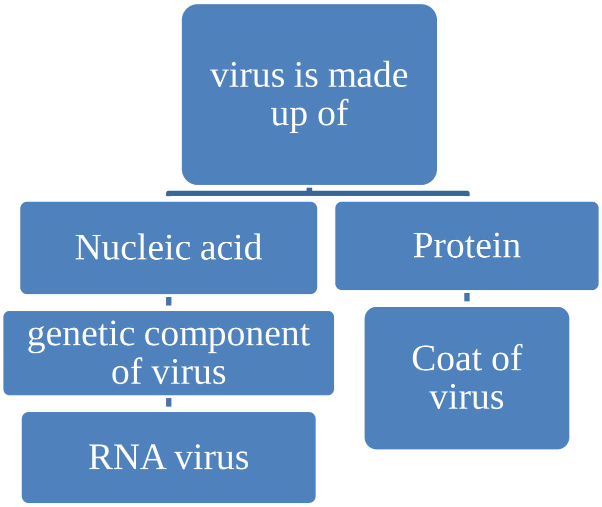


Figure 2.3.1: Genetic material within a virus genome

2.4 Biochemistry of COVID-19

The attachment of CoV-2's viral spike protein(s) to cellular receptors and priming by host cell proteases are key factors influencing the virus entrance into the host cell. According to SARS-biology numerous studies have been identified transmembrane protease serine 2 (TMPRSS2) & Angiotensin converting enzyme-2 (ACE2) as an important participants during this process. SARS CoV-2's interacts within amongst cellular receptor ACE 2 in order to enter the host cell [24].

ACE2 takes role in regulating systems within human bodies. Furthermore, ACE2 serves like a regulatory receptor for the coronavirus that causes severe acute respiratory syndrome (SARS-CoV's). Because of commencement high extent of ACE 2 expression amongst heart and lungs, patients with COVID-19 may have problems with their hearts or lungs. TMPRSS2 degrades the SARS-CoV2 spike protein that activates this same virus and opens its cellular membrane. This correspondence allying race & illness consequences might bring on by single-nucleotide polymorphisms (SNIPs) in the linked genomes until those proteins gain their ingress of SARS-CoV-2's infection of enterocytes [25].

SNPs provide information about folk's sensitivity to environmental influences along with their probable reactivity towards different treatments and drugs. The analysis for SNPs which influence SARSCoV-2's vulnerability potential harshness could therefore be useful for developing personalized coronavirus treatments. Patient-specific drugs & therapies promote quicker recovery through eliminating unnecessary treatments. Additionally, this would minimise and eliminate those adverse effects that particular medications that specific individuals have. Determining the SNIPs for SARS-CoV-2’s pathogenicity that were shared on account of all of SNIPs reported in the numerous studies [26].

SARS-CoV-2 is an enveloped virus with a 29.9 kb positive-strand RNA genome. This disease is mediated by the ACE-2 enzyme. The SARS-CoV2 & SARS-CoV which were 80% interchangeable use the angiotensin-converting enzyme 2 (ACE2) as a cellular entrance receptor. These major protein molecules present in Coronaviruses are indeed the spike (S) membrane (M) Nucleocapsid (N) and an envelope (E) proteins. The spike protein, which makes the Coronavirus's exterior protrude inside a noticeable way gives this disease their term [27].

Membrane merging & adherence were handled, separately, by components S1 and S2, which make up the S proteins as **(fig 2.4.1).** Its S1 subunit of a spiking receptor ties with *Homo sapiens* ACE2’s (hACE2’s) in the biological membranes via its receptor-binding region (RBD). It was found the ACE2 is more affine towards SARS-CoV-2 RBD than to SARS-CoV RBD, by a factor of 10–20. Furthermore, SARS-CoV-2 RBD can withstand soluble hACE2 greater effectively that SARS-CoV. The hACE2's increased propensity might assist toward explain SARS-heightened CoV-2's infectivity, given that COVID-19 is endemic in many areas because new instances are being reported more often [28].

Both transmembrane protease serine protease-2 (TMPRSS-2) and ADAM17 metallopeptidase domains of a human host were required in preparing the S protein so that the S2 subunit may facilitate that merging of either the viral and host membranes. Following internalization of SARS-CoV2 via endocytosis, viral RNA is freed to be used by the host cell's machinery in viral translation and replication as well as in the assembly and exocytosis of many more viral proteins [29].

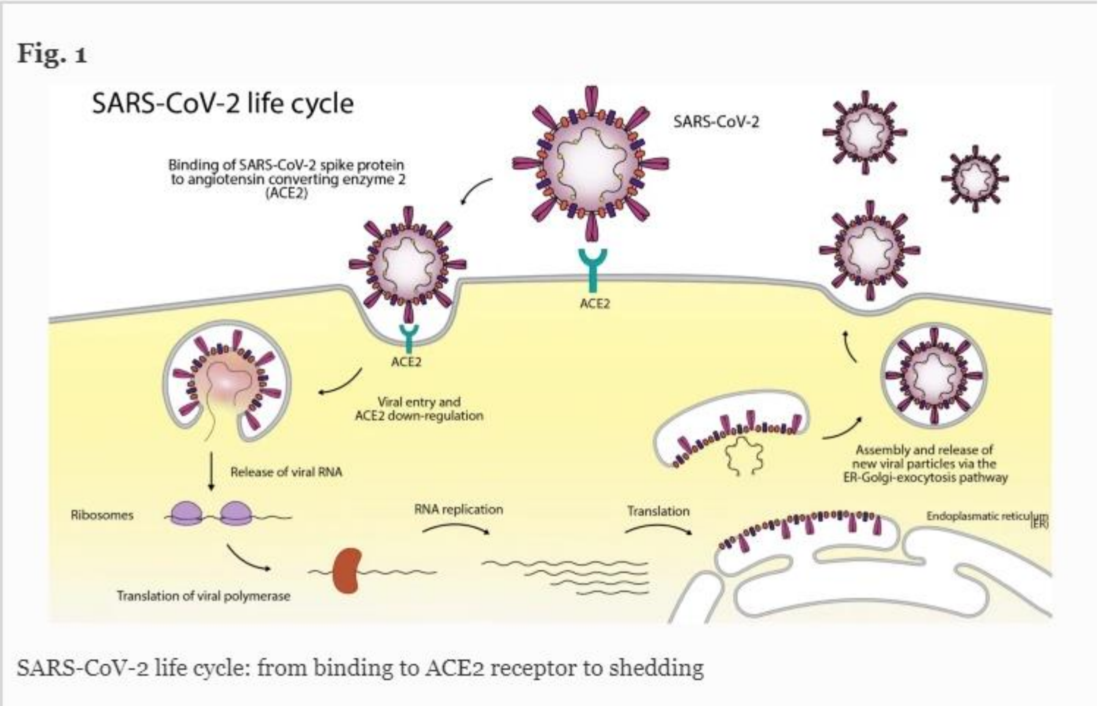


Fig. 1
SARS-CoV-2 life cycle
Binding of SARS-CoV-2 spike protein to angiotensin converting enzyme 2 (ACE2)
SARS-CoV-2
ACE2
ACE2
Viral entry and ACE2 down-regulation
Release of viral RNA
Ribosomes
Translation of viral polymerase
RNA replication
Translation
Assembly and release of new viral particles via the ER-Golgi-exocytosis pathway
Endoplasmatic reticulum (ER)

SARS-CoV-2 life cycle: from binding to ACE2 receptor to shedding

Figure 2.4.1 SARS-C0V-2 Life Cycle

2.5 COVID-19 common symptoms

The much more acute manifestations of COVID-19 include breathlessness, muscle aches, fatigue, or a chest infection. As furthermore, reports of sensory loss, excessive and prolonged, or impaired liver performance were also made. These indications include sputum production, headache, abdominal pain, diarrhoea, nausea, and nausea. The SARS-CoV-2 targets a number of organs which produce ACE2, which could explain all aforementioned symptoms. Any pathogen for one or more unique mutations is known as that of the unique virus variety; these changes can be one or more point mutations [30].

Because changes occur often, alternative forms would unavoidably emerge throughout an epidemic. This D614G variant, that first appeared in the early COVID-19 pandemic and has since emerged as the most common variety circulating worldwide, is present in all of the SARS-CoV2 variants that have already been found to exist. Alpha, Beta, and Gamma—which correspond to Pangolineages B.1.1.7 B.1.351 and P.1 respectively three kinds of particular significance, however, quickly took over in a number of countries as the epidemic spread and raised specific issues. Describing the different of interest, including Zeta have gained popularity as well [31].

2.6 Publications opted for

All of the outcomes from the aforementioned Medline expression search were included in the first 2956 papers. Furthermore, papers that weren't acceptable for this retrospective study were filtered out using the following exclusion criteria.

The ensuing appropriate standards have been used:

1. Analysis demonstrating employ *Homo sapiens* volunteers who have been infected with the coronavirus.

2. Investigations of the COVID-19 emergency (survey released in December 2019 or later).

3. COVID-19’s research which focuses on the ancestry or mode of intercellular contamination.

4. Investigation of referencing genes and SNPs specifically linked to COVID-19 [32].

2.7 The ensuing exemption standards have been used:

1. Investigations into many other Corona viruses, such as the bovine and delta Corona viruses, in both animals and people.

2. Samples of such types of research include editorial characters, symbols, mark, type, figures case studies, technical notes, reviews, and systematic reviews.

3. Research that are immaterial, like those on porcine diarrhoea.

4. Research of COVID-19 which ignored genetics or the manner wherein cells became infected [33].

2.8 Main genes involved are:

This gene ACE-2 was named highest, while in other articles, TMRSS2 and IFITM3, CD147 IFIH1**(fig2.8.1)** have also been highlighted. As per data from various studies, a number of SNPs were generally connected that how terrible COVID-19 and SARS-CoV-2 transmission.

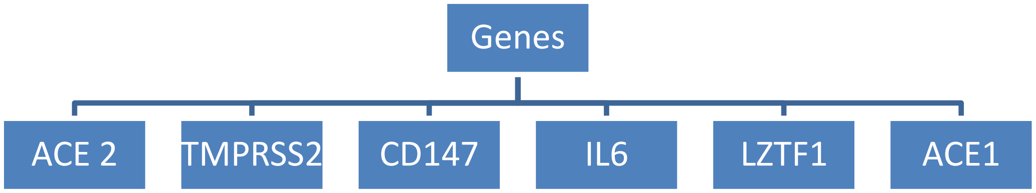


Figure 2.8.1 Main Genes name involved in SARS-CoV-2

2.9 The associated SNPs and genes

The following Snps are:

It is generally known that among the China inhabitants, the rs12252-C alternative is substantially associated with influenza infection. Furthermore, given that Spanish databases regularly identify it as a risk factor, rs12252 C,rs14393628 **(in fig 2.9.1).**

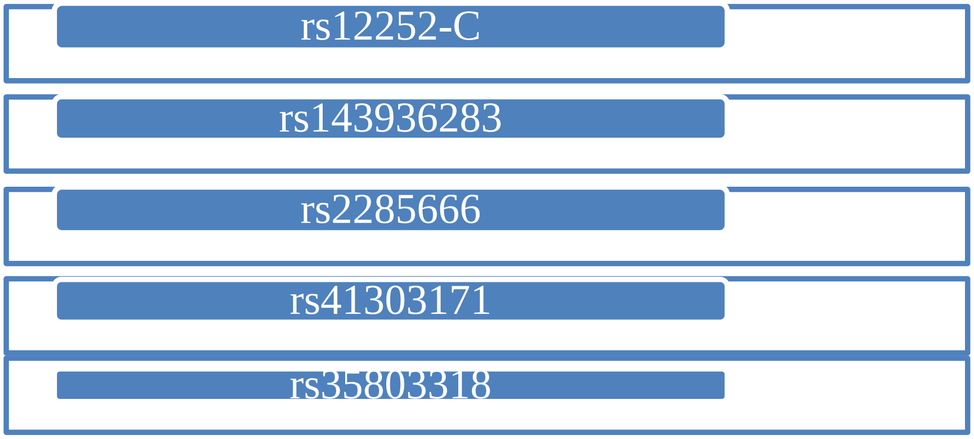


Figure 2.9.1: Main SNPs involved in SARS-Cov-2

It is generally known that among the China inhabitants, the rs12252-C alternative is substantially associated with influenza infection. Furthermore, given that Spanish databases regularly identify it as a risk factor, rs12252 C may affect SARS-CoV-2 infection in all populations, including those in Europe. The studies we analysed identified 2 related SNPs for IFITM3: rs12252-C and rs6598045 **(table 2.9.1).** The most relevant SNPs were found in ACE2 and IFITM3, followed by TMPRSS2 [34].

Table 2.9.1 List of genes found to be involved in SARS-Cov-2 & their functions

S. No.	**Genes**	**SNPs**	**Function**
1	ACE2(angiotensin1-converting enzyme 2)	rs75603675rs2285666 rs879922rs73635825, rs4646114 rs464611	SARS-CoV-2 spike protein entry receptor
2	IFITM3 (interferon-induced transmembrane protein 3)	rs12252-C rs6598045	IFITM3 gene variations have been linked to pneumonia and viral infection. IFITM3 plays an important role in antiviral activities..

When working with transcript data mapping the data to a group of unigenes is the easiest option because it results in an ungapped alignment. In the absence of such a dataset a dataset could delineate to genomic data using a spliced alignment technique. At the mapped data, the novel sequenced scrutinize on the reference resides in its aligned place [35].

Technology programmes like as Phrap and CAP3 are routinely used to assemble the sequences into contigs. Within sequenced changes for every location being portrayed by many more readings. When a species has accessible additional sequencing reads for any particular genome area basics likelihood for discovering any polymorphic increases. Additionally whenever the sequencing mistake is detected a sequenced variation (allele) could be identified backed owing to a large number of reads. The more readings per allele the more likely it is that an allele is a true polymorphism [36].

Specialized assembly technologies are used to segregate the accretion datasets aren’t assembled as contigs in the example of de novo sequence information where what groups sequencing information of the unvaried area for the genomic [37].

The process requires more reads thus it takes longer. Specialized methods like d2cluster and TGICL the past created towards achieving a starting separation groupings of same sequenced segments that have existence then furthermore divided different groups with separate origination. Clustering results within every cluster must owing to be treated in order properly synchronize each and every readings contained therein. It is easy to compare the nucleotides from numerous readings that almost all align at the same place on the gene or genome. The fragments cannot be adequately aligned [38].

They are divided into two clusters because they are not all part of one. Individual readings must first be sorted within synchronized same group before the polymorphic similarity process can detect changes in the alignment and apply any matrix system. The design up to modern SNIP finding techniques frequently enables their integration with currently available genome analysis Programmes like DnaSP [39].

The level of complexity of the procedure heavily influences the kind of technology required for SNP mining. In just a few hundred sequences a normal workstation is perfectly capable of looking about SNIPs in specific small areas of the genome (up to 100–150 kb). Extracting SNIPs throughout the genomes initiatives frequently necessary server-class processors and accessibility with or hundreds of cloud hosting in megabytes of information particularly unless extraction process's intermediary phases were recorded & the outcomes were logged in a database [40].

2.10. Phylogeny Reconstruction

A phylogeny is the evolutionary background is a set of items considering that this is only possible to determine in rare cases the primary goal of phylogeny rebuilding is to define evolutionary connections with regard to of the relative recency of common ancestry. All of these connections are depicted as a branching diagram or tree with branches linked by nodes and ultimately to terminals at the tree's points . The three major forms of relationships are monophyly paraphyly and polyphyly. The evolution of monophyletic and paraphyletic groupings is the same. Monophyletic groupings comprise all offspring of a single ancestor as well as that ancestor. If one lineage from a monophyletic group is removed, a paraphyletic group remains. Polyphyletic groupings, on the other hand, arise as a result of convergent evolution and the individuals who promote the collective are missing from the most recently prevalent progenitor. These concepts are similar to orthology and paralogy in biological family. Orthology is the term for clusters of genomes that show biological ancestry. As a consequence inside each different gene group every organism is portrayed via just one orthologue. whereas paralogues represent the history of a gene family. As a consequence of this inside a gene's group every organism might possess several paralogues [41].

2.11. Overview of phylogenetic analysis

Selecting a Research Groups Prior beginning phylogenetic reconstruction, consider the particular biology issue to be addressed. To minimise artefactual linkages among terminals, sample as densely as feasible. If the goal of the remodeling is to determine when imitating happened among a family of genes from just one species it is relevant to sample the gene family from that species extensively [42].

Nevertheless if the goal is to acknowledge how a gene family progress it is critical to sample as many times as feasible not just inside species but also across species. A excellent place to start is to go through the literature on the topic of concern. This will influence the kind of organism and genes chosen incorporated into the study and will determine groups whose links will probably to be clarified and quantitatively validated in the resultant phylogeny. This will also indicate groups that need further testing or care in alignments [43].

CHAPTER-3

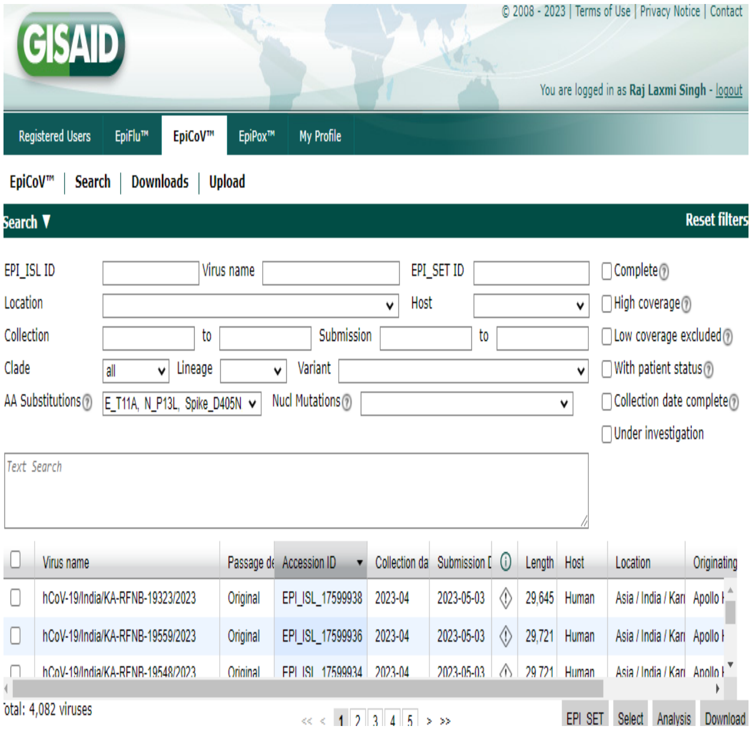
3.1. Materials & Methods

3.1.1. Materials

The data sets were extracted from the GISAID database in FASTA format for further processing. Furthermore, the whole genome sequences of SARS-CoV2 from different COVID-19 cases and the reference genome (NC_045512.2) have been retrieved via the GISAID and NCBI GenBank records, as well. Have taken sequences from three different places: Asia, Europe & North America (**table 3.1.1.1).**

Table 3.1.1.1: Have taken sequences from three different places: Asia, Europe, North America

Asia	**Sequences**	**Europe**	**Sequences**	**North America**	**Sequences**
Singapore	3 sequences	Netherlands	2 sequences	USA- New York	17 sequences
Mumbai	2 sequences	Germany	1 sequences		
		Denmark	5 sequences		


© 2008 - 2023 | Terms of Use | Privacy Notice | Contact
GISAID
You are logged in as Raj Laxmi Singh - logout
Registered Users
EpiFlu™
EpiCoV™
EpiPox™
My Profile
EpiCoV™ | Search | Downloads | Upload
Search ▼
Reset filters
Text Search
<< < 1 2 3 4 5 > >>
EPI SET
Select
Analysis
Download

EPI_ISL ID
Virus name
EPI_SET ID
□ Complete
Location
Host
□ High coverage
Collection to
Submission to
□ Low coverage excluded
Clade: all
Lineage
Variant
□ With patient status
AA Substitutions: E_T11A, N_P13L, Spike_D405N
Nucl Mutations
□ Collection date complete
□ Under investigation

□	Virus name	Passage de	Accession ID	Collection da	Submission D	ⓘ	Length	Host	Location	Originating
□	hCoV-19/India/KA-RFNB-19323/2023	Original	EPI_ISL_17599938	2023-04	2023-05-03	⚠	29,645	Human	Asia / India / Kar	Apollo H
□	hCoV-19/India/KA-RFNB-19559/2023	Original	EPI_ISL_17599936	2023-04	2023-05-03	⚠	29,721	Human	Asia / India / Kar	Apollo H
□	hCoV-19/India/KA-RFNB-19548/2023	Original	EPI_ISL_17599934	2023-04	2023-05-03	⚠	29,721	Human	Asia / India / Kar	Apollo H

otal: 4,082 viruses

Figure 3.1.1.1: Out of 4082 virus selected 30 cases by applying complete & high coverage filter

3.2.1. Methodology

Whole genome sequence datasets of SARS-CoV-2 secluded amidst distinctive COVID-19 cases were repossessed by downloading from GISAID database. A total of 30 sequences that triumphant quality assurance (length 29,700 nts) were used for the study in the (**fig 3.1.1.1)**. In amenity SARS-CoV2 genome sequences gathered amidst distinctive COVID19 patients as well as the reference genome (Accession NC_ 045512.2) were repossess amidst the GISAID and GenBank databases correspondingly. MAFFT (Version 7.471) was used for multiple sequence alignment (MSA) while DnaSP (Version 6.12.03) was used for SNP calling, which was subsequently visualized in Jalview (Version 2.11.1.0). MEGA X software was used for the phylogenetic analysis [44].

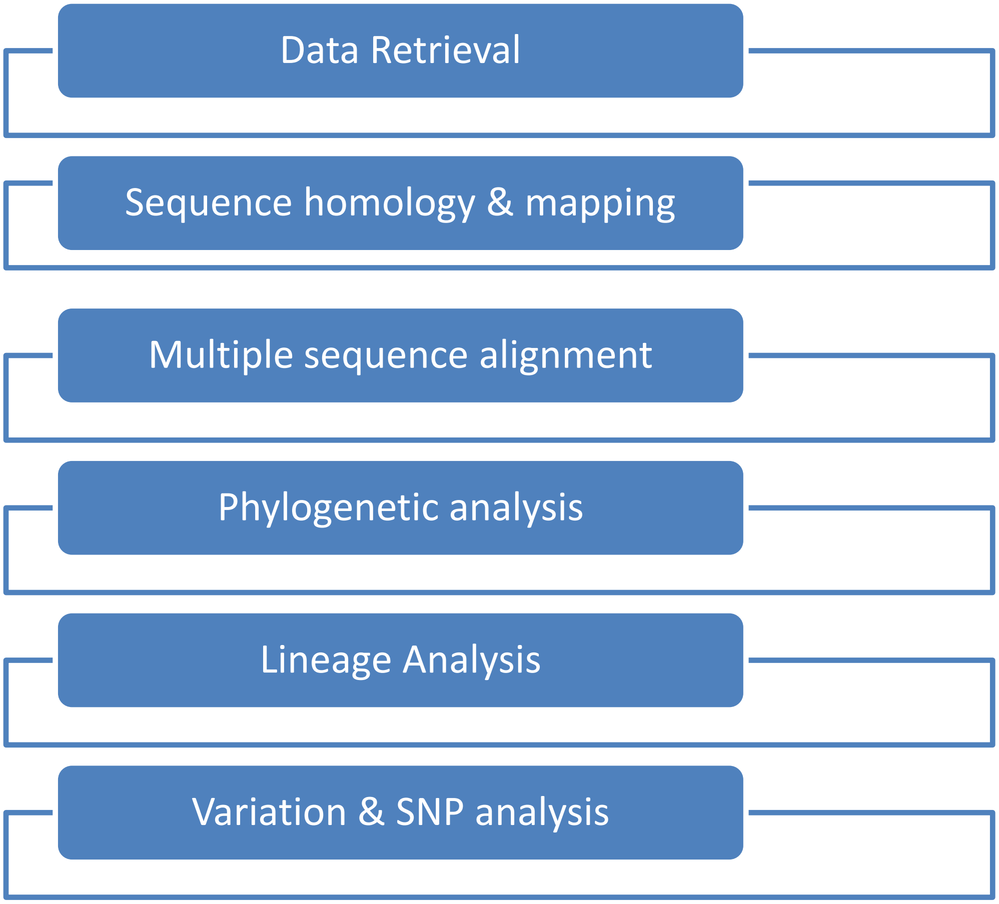


Figure 3.2.1.1: Flowchart representing methodology of the analysis

Softwares used to analyse different COVID-19 cases in the following steps:

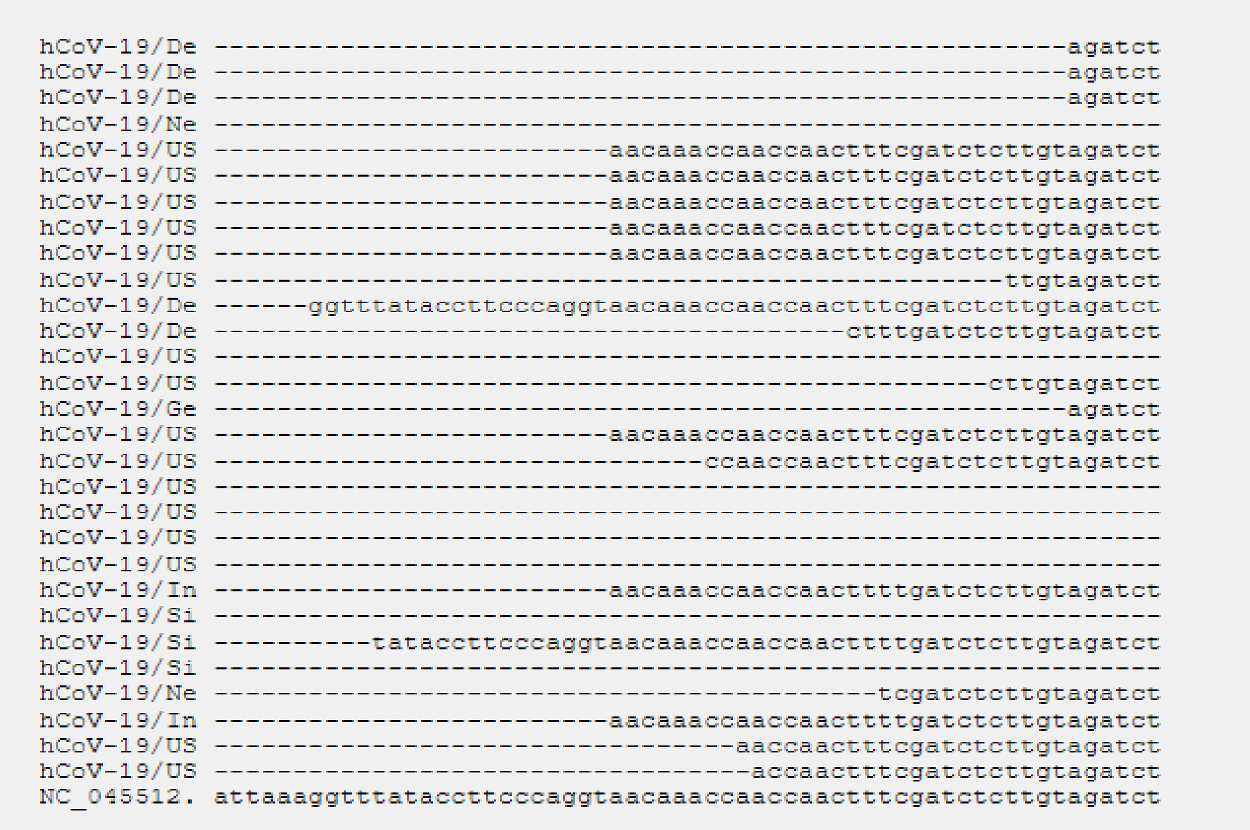
1. Data retrieval: Collection of different COVID-19 cases were downloaded from **GISAID databases (https://gisaid.org/).** The GISAID Initiative encourages the swift transfer of information regarding COVID-19-causing coronaviruses as well as various influenza virus strains. In order to better comprehend how viruses evolve and spread during pandemics and outbreaks researchers can use the sequence of genes, relevant clinically and epidemiological information along with geographic & species-specific data linked with avian and other animal viruses as well as data associated with human viruses. By removing obstacles and constraints that discouraged or hindered the exchange of virological data prior to official publication GISAID is able to achieve its goals. The Initiative makes sure that everyone has unrestricted access to GISAID data for no cost. SARSCoV2 complete genome sequences extracted among different COVID19 patients has been downloaded via the GISAID (Global Initiative for Sharing All Influenza Data) website. Till 03rd Feb 2023, a total of thousands of sequences has been submitted. Once 30 of these sequences were percolate by using the "complete" filter option on the GISAID database page, it implied that a some sequence was incomplete (**Fig 3.1.1.1).** The "enormous range" sorting button was also selected to guarantee appropriate size and quality of the SARSCoV2 gene sequence considering the purpose of the investigation. The data sets were extracted from the GISAID database in FASTA format for further processing. Furthermore, the whole genome sequences of SARS-CoV2 from different COVID-19 cases and the reference genome (NC_045512.2) have been retrieved via the GISAID and NCBI GenBank records, as well [45]. Have taken sequences from three different places: Asia, Europe & North America (**refer to table 3.1.1.1).**

2. Sequencing homologous & mappings

Sequence aligned in **(MAFT version 7.471 https://mafft.cbrc.jp/alignment/software/):** A multiple sequence alignment programme for Unix-like operating systems is called MAFFT. It provides a variety of multiple alignment techniques including L-INS-i (accurate for alignment of 200–200) and FFT-NS-2 (rapid for alignment of 30000–30000) sequences) & trimmed in (**MEGA-X https://www.megasoftware.net/).** The Molecular Evolutionary Genetics Analysis (Mega) software executes a lot analytical techniques and applications for phylogenomics and phylomedicine.

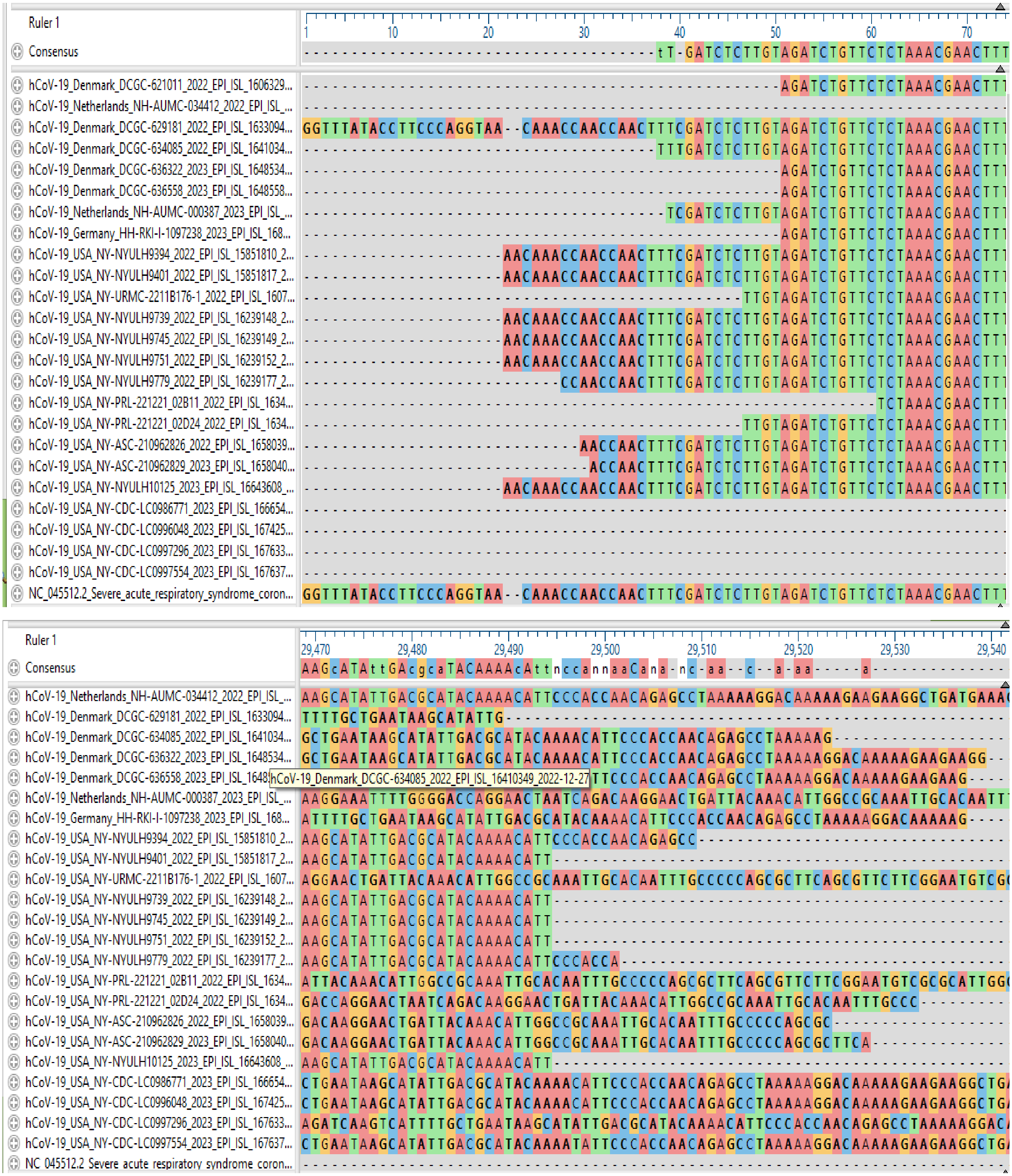
To verify accurate similarities, non-biological distinctive (i.e. changes owing to technological variants) were eliminated amidst the recovered sequences. MAFFT (**Fig 3.2.1.2)** was used for aligning the regions of DNA, and MEGA X was used for cutting the 5'and 3'ends (**Fig 3.2.1.3)** to yield sequences that are identical of 29,787 nts piece. Aligning indicated that 80 nucleotide ought to be deleted from the 5’ end and 200 nucleotides ought to be eliminated at the 3’ end to produce the 29787 nts for the respective sequences that are homo utilized in the study.

The trimmed sequences have been plotted in contact with a reference genome sequence of SARSCoV2 acquired amidst NCBI GenBank to identify the precise location of places on the chosen genomic sequences (Accession No: NC_045512.2) [46].



```
hCoV-19/De ------------------------------------------------------agatct
hCoV-19/De ------------------------------------------------------agatct
hCoV-19/De ------------------------------------------------------agatct
hCoV-19/Ne ------------------------------------------------------------
hCoV-19/US -------------------------aacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/US -------------------------aacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/US -------------------------aacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/US -------------------------aacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/US -------------------------aacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/US --------------------------------------------------ttgtagatct
hCoV-19/De ------ggtttataccttcccaggtaacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/De ----------------------------------------ctttgatctcttgtagatct
hCoV-19/US ------------------------------------------------------------
hCoV-19/US -------------------------------------------------cttgtagatct
hCoV-19/Ge ------------------------------------------------------agatct
hCoV-19/US -------------------------aacaaaccaaccaactttcgatctcttgtagatct
hCoV-19/US -------------------------------ccaaccaactttcgatctcttgtagatct
hCoV-19/US ------------------------------------------------------------
hCoV-19/US ------------------------------------------------------------
hCoV-19/US ------------------------------------------------------------
hCoV-19/US ------------------------------------------------------------
hCoV-19/In -------------------------aacaaaccaaccaacttttgatctcttgtagatct
hCoV-19/Si ------------------------------------------------------------
hCoV-19/Si ----------tataccttcccaggtaacaaaccaaccaacttttgatctcttgtagatct
hCoV-19/Si ------------------------------------------------------------
hCoV-19/Ne ------------------------------------------tcgatctcttgtagatct
hCoV-19/In -------------------------aacaaaccaaccaacttttgatctcttgtagatct
hCoV-19/US ---------------------------------aaccaactttcgatctcttgtagatct
hCoV-19/US ----------------------------------accaactttcgatctcttgtagatct
NC_045512. attaaaggtttataccttcccaggtaacaaaccaaccaactttcgatctcttgtagatct
```

Figure 3.2.1.2: Sequences were aligned by using MAFT version 7.47

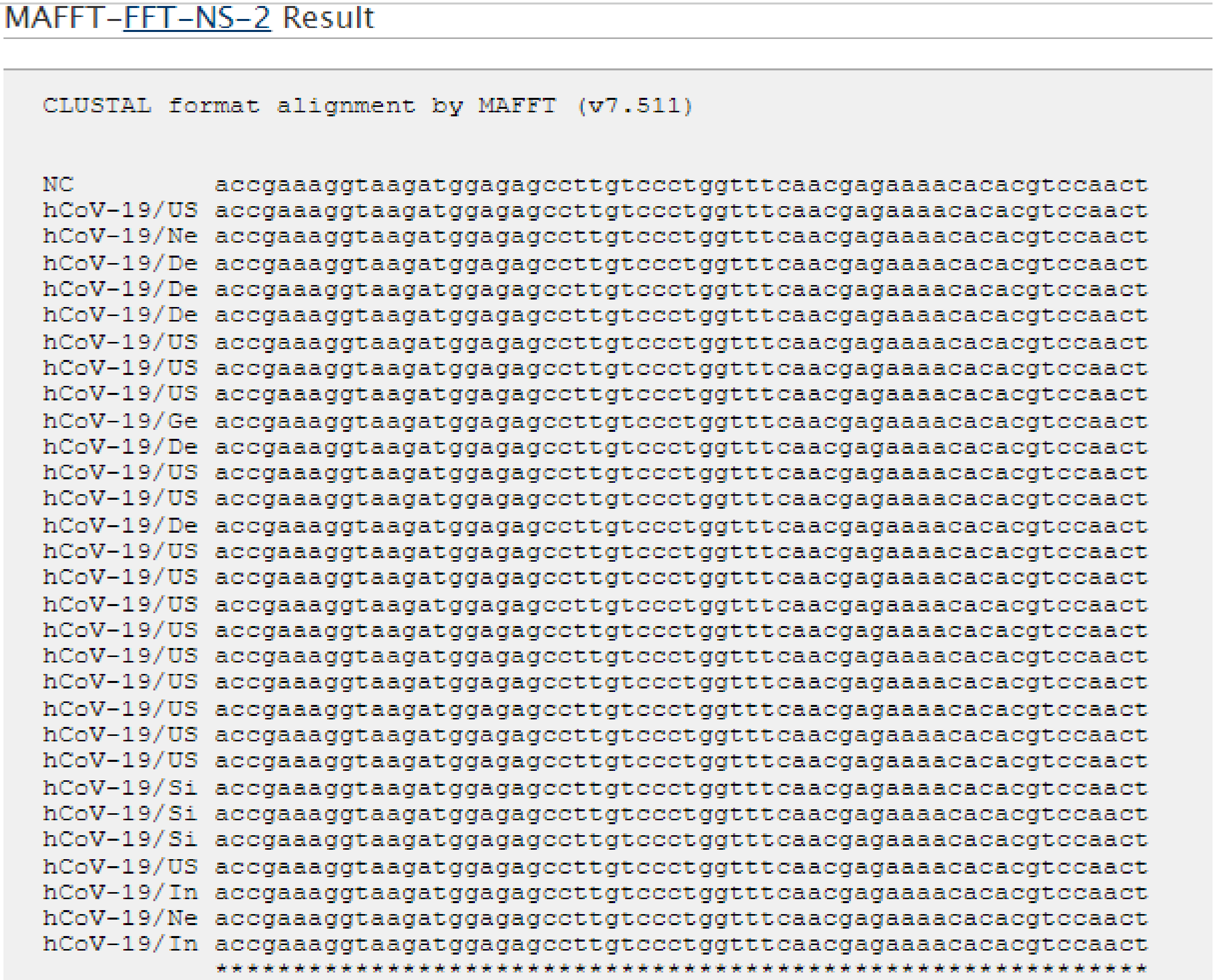


```
Ruler 1      1        10        20        30        40        50        60        70
Consensus    -------------------------------------tT-GATCTCTTGTAGATCTGTTCTCTAAACGAACTTT

hCoV-19_Denmark_DCGC-621011_2022_EPI_ISL_1606329...     --------------------------------------------------AGATCTGTTCTCTAAACGAACTTT
hCoV-19_Netherlands_NH-AUMC-034412_2022_EPI_ISL_...     --------------------------------------------------------------------------
hCoV-19_Denmark_DCGC-629181_2022_EPI_ISL_1633094...     GGTTTATACCTTCCCAGGTAA--CAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_Denmark_DCGC-634085_2022_EPI_ISL_1641034...     -------------------------------------TTTGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_Denmark_DCGC-636322_2023_EPI_ISL_1648534...     --------------------------------------------------AGATCTGTTCTCTAAACGAACTTT
hCoV-19_Denmark_DCGC-636558_2023_EPI_ISL_1648558...     --------------------------------------------------AGATCTGTTCTCTAAACGAACTTT
hCoV-19_Netherlands_NH-AUMC-000387_2023_EPI_ISL_...     --------------------------------------TCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_Germany_HH-RKI-I-1097238_2023_EPI_ISL_168...    --------------------------------------------------AGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH9394_2022_EPI_ISL_15851810_2...     ---------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH9401_2022_EPI_ISL_15851817_2...     ---------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-URMC-2211B176-1_2022_EPI_ISL_1607...     ----------------------------------------------TTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH9739_2022_EPI_ISL_16239148_2...     ---------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH9745_2022_EPI_ISL_16239149_2...     ---------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH9751_2022_EPI_ISL_16239152_2...     ---------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH9779_2022_EPI_ISL_16239177_2...     ---------------------------CCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-PRL-221221_02B11_2022_EPI_ISL_1634...    ------------------------------------------------------------TCTAAACGAACTTT
hCoV-19_USA_NY-PRL-221221_02D24_2022_EPI_ISL_1634...    ----------------------------------------------TTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-ASC-210962826_2022_EPI_ISL_1658039...    -----------------------------AACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-ASC-210962829_2023_EPI_ISL_1658040...    ------------------------------ACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-NYULH10125_2023_EPI_ISL_16643608_...     ---------------------AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT
hCoV-19_USA_NY-CDC-LC0986771_2023_EPI_ISL_166654...     --------------------------------------------------------------------------
hCoV-19_USA_NY-CDC-LC0996048_2023_EPI_ISL_167425...     --------------------------------------------------------------------------
hCoV-19_USA_NY-CDC-LC0997296_2023_EPI_ISL_167633...     --------------------------------------------------------------------------
hCoV-19_USA_NY-CDC-LC0997554_2023_EPI_ISL_167637...     --------------------------------------------------------------------------
NC_045512.2_Severe_acute_respiratory_syndrome_coron...  GGTTTATACCTTCCCAGGTAA--CAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTT

Ruler 1      29,470    29,480    29,490    29,500    29,510    29,520    29,530    29,540
Consensus    AAGcATAttGAcgcaTACAAAAcAttnccannaaCana-nc-aa--c--a-aa-----a---------------

hCoV-19_Netherlands_NH-AUMC-034412_2022_EPI_ISL_...     AAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAA
hCoV-19_Denmark_DCGC-629181_2022_EPI_ISL_1633094...     TTTTGCTGAATAAGCATATTG-----------------------------------------------------
hCoV-19_Denmark_DCGC-634085_2022_EPI_ISL_1641034...     GCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAG-------------------
hCoV-19_Denmark_DCGC-636322_2023_EPI_ISL_1648534...     GCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGG---
hCoV-19_Denmark_DCGC-636558_2023_EPI_ISL_1648 hCoV-19_Denmark_DCGC-634085_2022_EPI_ISL_16410349_2022-12-27 TTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAG----
hCoV-19_Netherlands_NH-AUMC-000387_2023_EPI_ISL_...     AAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATT
hCoV-19_Germany_HH-RKI-I-1097238_2023_EPI_ISL_168...    ATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAG----
hCoV-19_USA_NY-NYULH9394_2022_EPI_ISL_15851810_2...     AAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCC---------------------------------
hCoV-19_USA_NY-NYULH9401_2022_EPI_ISL_15851817_2...     AAGCATATTGACGCATACAAAACATT------------------------------------------------
hCoV-19_USA_NY-URMC-2211B176-1_2022_EPI_ISL_1607...     AGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATGTCG
hCoV-19_USA_NY-NYULH9739_2022_EPI_ISL_16239148_2...     AAGCATATTGACGCATACAAAACATT------------------------------------------------
hCoV-19_USA_NY-NYULH9745_2022_EPI_ISL_16239149_2...     AAGCATATTGACGCATACAAAACATT------------------------------------------------
hCoV-19_USA_NY-NYULH9751_2022_EPI_ISL_16239152_2...     AAGCATATTGACGCATACAAAACATT------------------------------------------------
hCoV-19_USA_NY-NYULH9779_2022_EPI_ISL_16239177_2...     AAGCATATTGACGCATACAAAACATTCCCACCA-----------------------------------------
hCoV-19_USA_NY-PRL-221221_02B11_2022_EPI_ISL_1634...    ATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATGTCGCGCATTGG
hCoV-19_USA_NY-PRL-221221_02D24_2022_EPI_ISL_1634...    GACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCC----------
hCoV-19_USA_NY-ASC-210962826_2022_EPI_ISL_1658039...    GACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGCGC-------------------
hCoV-19_USA_NY-ASC-210962829_2023_EPI_ISL_1658040...    GACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGCGCTTCA---------------
hCoV-19_USA_NY-NYULH10125_2023_EPI_ISL_16643608_...     AAGCATATTGACGCATACAAAACATT------------------------------------------------
hCoV-19_USA_NY-CDC-LC0986771_2023_EPI_ISL_166654...     CTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGA
hCoV-19_USA_NY-CDC-LC0996048_2023_EPI_ISL_167425...     CTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGA
hCoV-19_USA_NY-CDC-LC0997296_2023_EPI_ISL_167633...     AGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACA
hCoV-19_USA_NY-CDC-LC0997554_2023_EPI_ISL_167637...     CTGAATAAGCATATTGACGCATACAAAATATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGA
NC_045512.2_Severe_acute_respiratory_syndrome_coron...  --------------------------------------------------------------------------
```

Figure 3.2.1.3: Aligned Sequences Trimmed at 5’ and 3’ ends to Obtain True Homology

3. Muliple sequence alignment

MAFFT Version 7.471 was used for multiple sequence alignment (MSA), and MEGA X was used for phylogenetic reconstruction using Pdistance (in units of number of base differences per site). The entire genomes were subsequently matched in the company of MAUVE to look considering major-scurf genomic alterations such as significant eliminations gene inversions & genome rearrangements. The resulting sequences come about then realigned in MAFFT (**fig 3.2.1.4**) to yield aligned sequences which take place towards the DnaSP for SNP and haplotype evaluation then imported into Jalview 2.11.1.0 for visualisation & automated allelic frequency calculation of SNPs [47].



MAFFT-FFT-NS-2 Result

```
CLUSTAL format alignment by MAFFT (v7.511)


NC         accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/Ne accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/De accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/De accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/De accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/Ge accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/De accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/De accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/Si accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/Si accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/Si accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/US accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/In accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/Ne accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
hCoV-19/In accgaaaggtaagatggagagccttgtccctggtttcaacgagaaaacacacgtccaact
           ************************************************************
```

Figure 3.2.1.4: Sequences were re- aligned by using MAFT version 7.47

4. **SNP and Variation Analysis: detection of SNPs was done by DnaSPv6.12.03 http://www.ub.edu/dnasp/):** DNA Sequence Polymorphism (DnaSP) is a bioinformatics application that uses a nice Graphic User Interface to analyse DNA sequence data variance. The programme enables extensive characterisation of the extent and swatch of DNA sequence discrepancy at various time scales using the polymorphic variations (intraspecific data) bifurcation data (interspecific or interpopulation data) or a mixture of either. Version 6 has new characteristics that are especially suited for analysing hundreds of DNA sequence areas in a single pass a feature that is increasingly in demand for RADseq-based research along with countless disciplines such as population genomics molecular ecology and clinical virology. In addition DnaSP6 involves additional features for running coalescent reconstruction below a variety of demographic information scenarios.

   Snps genetic sites have been discovered with regard via the NCBI GenBank SARS-CoV-2 reference genome sequence (Accession No: NC_045512.2). A polymorphism position is one where the prevalence of the next most common allele is more than 1%; otherwise, they are monomorphic. Uncommon alleles have a minor allele frequency (MAF) of fewer than one percent & are carried by uncommon versions of viruses. DnaSPv6.12.03 (**Fig 3.2.1.5)** was used for SNP detection linkage disequilibrium and haplotype analysis [48].

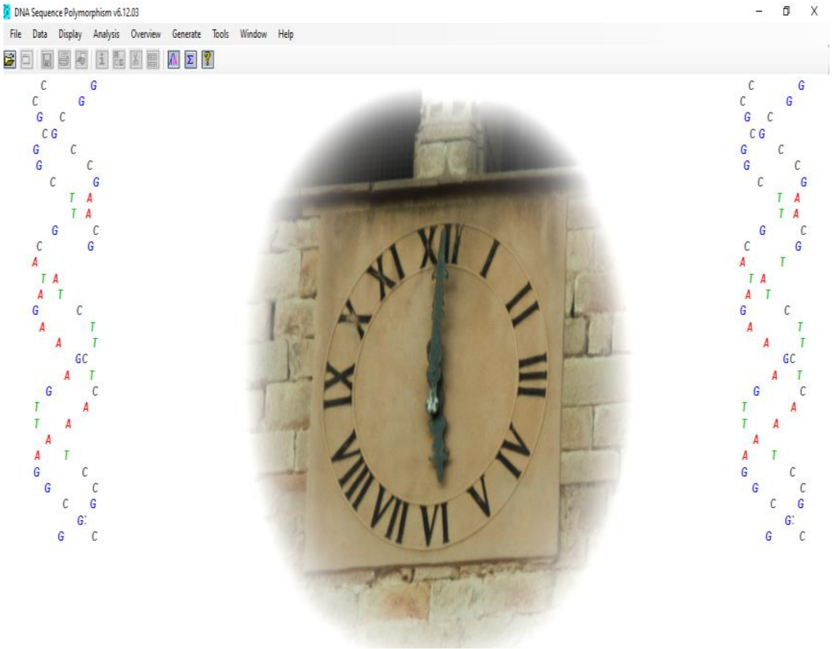


Figure 3.2.1.5: SNP and variation analysis was done by DnaSP version 6.12.0

5.Lineage Analysis

It was done by **pangolinv2.4.2 package** **https://github.com/cov-lineages/pangolin** **:** This enables an individual to attribute the highest common lineage (Pango lineage) to SARS-CoV-2 query sequence. The pangolin v2.4.2 software was used to conduct lineage analysis on the recovered SARS-CoV-2 genomic sequences [49].

6.Phylogenetic Analysis

MEGA X was utilised to build the greatest-likelihood trees of phylogeny **(Fig 3.2.1.6)** from data matched by MAFFT using the Tamura Nei evolutionary model with the persumptin of constant nucleotide replacement. Using the neighbor joining (NJ) & bioNJ algorithms is a type of heuristic searches the tree with the better log likelihood value was chosen through the starting trees. Topology and clustering pattern analysis were used to analyze with the bootstrapped value of 1000 replicates.

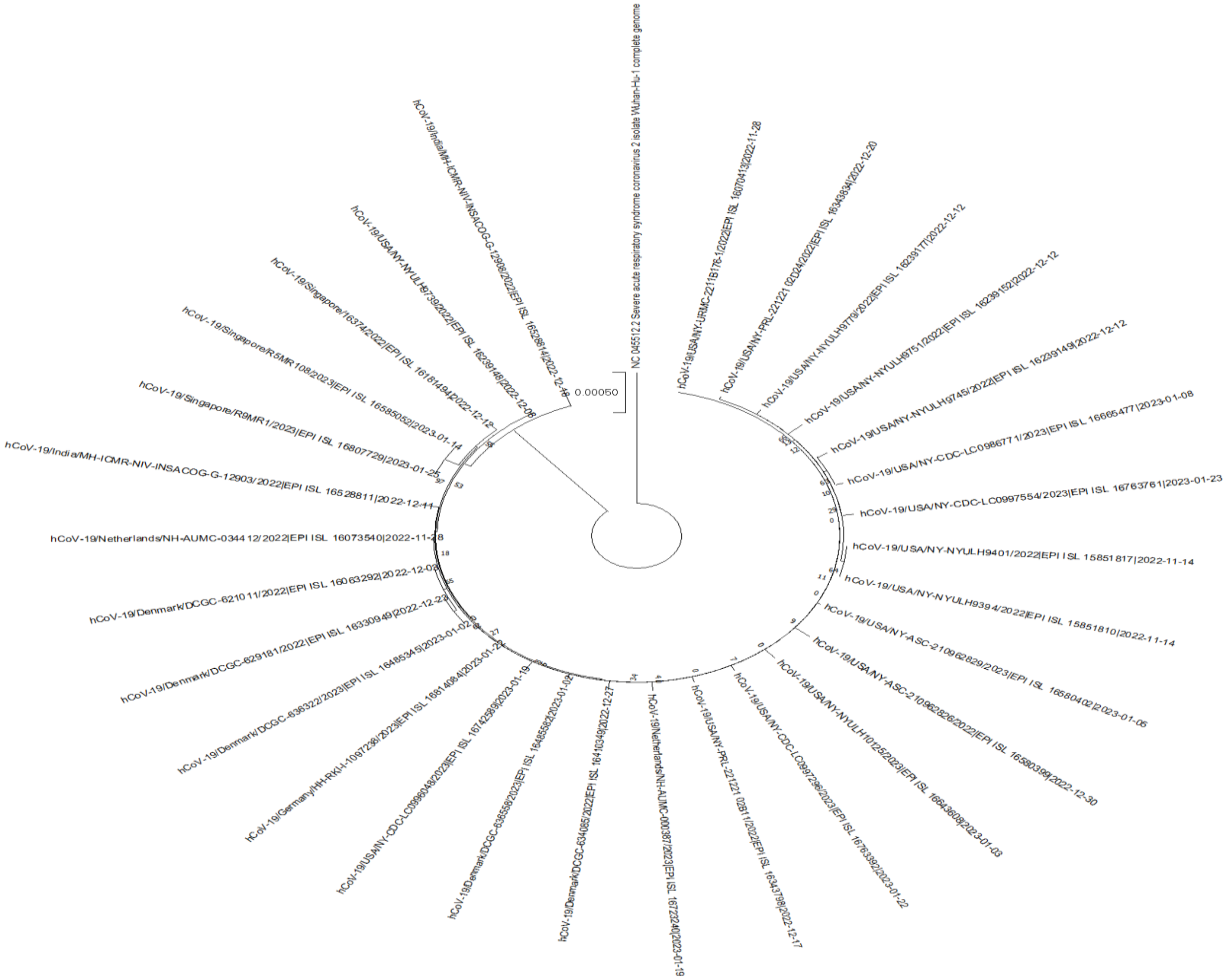

NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome
0.00050
hCoV-19/USA/NY-URMC-2211B176-1/2022|EPI ISL 16070413|2022-11-28
hCoV-19/USA/NY-PRL-221221 02D24/2022|EPI ISL 16343834|2022-12-20
hCoV-19/USA/NY-NYULH9779/2022|EPI ISL 16239177|2022-12-12
hCoV-19/USA/NY-NYULH9751/2022|EPI ISL 16239152|2022-12-12
hCoV-19/USA/NY-NYULH9745/2022|EPI ISL 16239149|2022-12-12
hCoV-19/USA/NY-CDC-LC0986771/2023|EPI ISL 16665477|2023-01-08
hCoV-19/USA/NY-CDC-LC0997554/2023|EPI ISL 16763761|2023-01-23
hCoV-19/USA/NY-NYULH9401/2022|EPI ISL 15851817|2022-11-14
hCoV-19/USA/NY-NYULH9394/2022|EPI ISL 15851810|2022-11-14
hCoV-19/USA/NY-ASC-210962829/2023|EPI ISL 16580402|2023-01-06
hCoV-19/USA/NY-ASC-210962826/2022|EPI ISL 16580399|2022-12-30
hCoV-19/USA/NY-NYULH10125/2023|EPI ISL 16643608|2023-01-03
hCoV-19/USA/NY-CDC-LC0997296/2023|EPI ISL 16763392|2023-01-22
hCoV-19/USA/NY-PRL-221221 02B11/2022|EPI ISL 16343798|2022-12-17
hCoV-19/Netherlands/NH-AUMC-000387/2023|EPI ISL 16723240|2023-01-19
hCoV-19/Denmark/DCGC-634085/2022|EPI ISL 16410349|2022-12-27
hCoV-19/Denmark/DCGC-638558/2023|EPI ISL 16485582|2023-01-02
hCoV-19/USA/NY-CDC-LC0996048/2023|EPI ISL 16742589|2023-01-19
hCoV-19/Germany/HH-RKI-I-097238/2023|EPI ISL 16814084|2023-01-26
hCoV-19/Denmark/DCGC-636322/2023|EPI ISL 16485345|2023-01-02
hCoV-19/Denmark/DCGC-629181/2022|EPI ISL 16330949|2022-12-23
hCoV-19/Denmark/DCGC-621011/2022|EPI ISL 16063292|2022-12-08
hCoV-19/Netherlands/NH-AUMC-034412/2022|EPI ISL 16073540|2022-11-28
hCoV-19/India/MH-ICMR-NIV-INSACOG-G-12903/2022|EPI ISL 16528811|2022-12-11
hCoV-19/Singapore/R9MR1/2023|EPI ISL 16807729|2023-01-25
hCoV-19/Singapore/R5MR108/2023|EPI ISL 16585052|2023-01-14
hCoV-19/Singapore/16374/2022|EPI ISL 16181494|2022-12-12
hCoV-19/USA/NY-NYULH9739/2022|EPI ISL 16239148|2022-12-08
hCoV-19/India/MH-ICMR-NIV-INSACOG-G-12908/2022|EPI ISL 16528814|2022-12-16
97
53
18
27
34
4
0
7
0
9
0
11
0
12

Figure 3.2.1.6: phylogenetic reconstruction of 30 SARS-CoV-2 genome sequences

CHAPTER 4

4.1.RESULTS & DISCUSSION

4.1.1.An overview regarding the obtained sequence

SARS-CoV-2 was isolated from 30 individuals shown in (3M: 11F: 16Unknown Sex) from three different places: Asia, Europe & north America and also no of COVID cases, deaths, and testing of 6 countries shown in (**Table 4.1.1.1)**

COVID-19 cases, deaths, and tests in 6 countries as of 03,Feb 2023

Table 4.1.1.1: COVID-19 statistics such as overall scenarios overall fatalities overall cases and deaths per million people and tests per million cases were collected from the Worldometer website for all nations and regions throughout the world**.**

Countries	Total cases	Total death	Total case/1M population	Countries	Total cases
India	4,46,96,338	5,30,806	31,775	India	4,46,96,338
Singapore	22,34,996	1,722	3,76,037	Singapore	22,34,996
Denmark	31,76,785	8337	5,44,441	Denmark	31,76,785
Germany	3,82,97,037	169661	4,56,550	Germany	3,82,97,037
Netherlands	86,05,996	22992	5,00,016	Netherlands	86,05,996
USA	105,972,038	1151642	3,16,518	USA	3,49,8142

The 30 SARS-CoV-2 genomic sequences utilised in the present study have the following parameters.

Table 4.1.1.2: The sequence IDs (GISAID) of the 30 genomes, as well as the acknowledgement list of all 30 isolates' sequence submitters

Asia	Accession I'D	Europe	Accession I'D	North America	Accession I'D
Singapore	epi_ISL_16807729	Netherlands	epi_ISL_16723240	USA-New York	epi_ISL_16763761
Singapore	epi_ISL_16585052	Netherlands	epi_ISL_16073540	USA-New York	epi_ISL_16763392
Singapore	epi_ISL_16181494	Denmark	epi_ISL_16485582	USA-New York	epi_ISL_16742589
Mumbai	epi_ISL_16528814	Denmark	epi_ISL_16485345	USA New York	epi_ISL_16665477
Mumbai	epi_ISL_16528811	Denmark	epi_ISL_16410349	USA New York	epi_ISL_16643608
		Denmark	epi_ISL_16410949	USA New York	epi_ISL_16580402
		Denmark	epi_ISL_16063292	USA New York	epi_ISL_16580399
		Germany	epi_ISL_16814084	USA New York	epi_ISL_16343834
				USA New York	epi_ISL_16343798
				USA New York	epi_ISL_16239177
				USA New York	epi_ISL_16239152
				USA New Yok	epi_ISL_16239149
				USA New York	epi_ISL_16239148
				USA New York	epi_ISL_16070413
				USA New York	epi_ISL_15851817
				USA New York	epi_ISL_15851810

4.1.2.Results of SNP analysis

The different place cases of SARSCoV2 utilised for this investigation shows ninety nine percent genetic similarity with 15 large conserved genomic regions (**Table 4.1.2.1)**. In the SARSCoV2 genomes that were utilized in this investigation 60 SNPs were found in which 30 snps are synonymous snps and the other 30 snps are non-synonymous snps. Apart for a triallelic SNP at 29791 nt all of the SNPs were diallelic. Pi = 0.00042 was the total nucleotide diversity across the SARS-CoV-2 genomes studied.

Table 4.1.2.1: large conserved genomic regions

Regions	Start-End	Conservation	Homozygosity	P-value
1.	419-1299	1	1	0.0058
2.	1301-1994	1	1	0.0175
3.	1996-2537	1	1	0.0429
4.	3067-3650	1	1	0.0335
5.	4282-4930	1	1	0.0228
6.	5633-6223	1	1	0.0321
7.	7604-8638	1	1	0.0023
8.	11093-11703	1	1	0.0286
9.	11705-12627	1	1	0.0045
10.	12629-13412	1	1	0.0103
11.	14515-15198	1	1	0.0186
12.	16091-16805	1	1	0.0155
13.	17996-19073	1	1	0.0018
14.	19075-19702	1	1	0.0259
15.	20226-21365	1	1	0.0012

4.1.3 Haplotype analysis for delineation of L and S lineages

The persual of linkages using pairwise SNP comparisons revealed that numerous SNPs in the SARSCoV2 genome are in 2locus linkage disequilibrium (LD) with the vastness of SNPs which are above the noteworthy (P 0.05) horizontal threshold line (**fig 4.1.3.1).** Fisher's exact test found 91 significant pairwise haplotypes 14 of which befall extremely noteworthy (P 0.001) using Bonferoni.

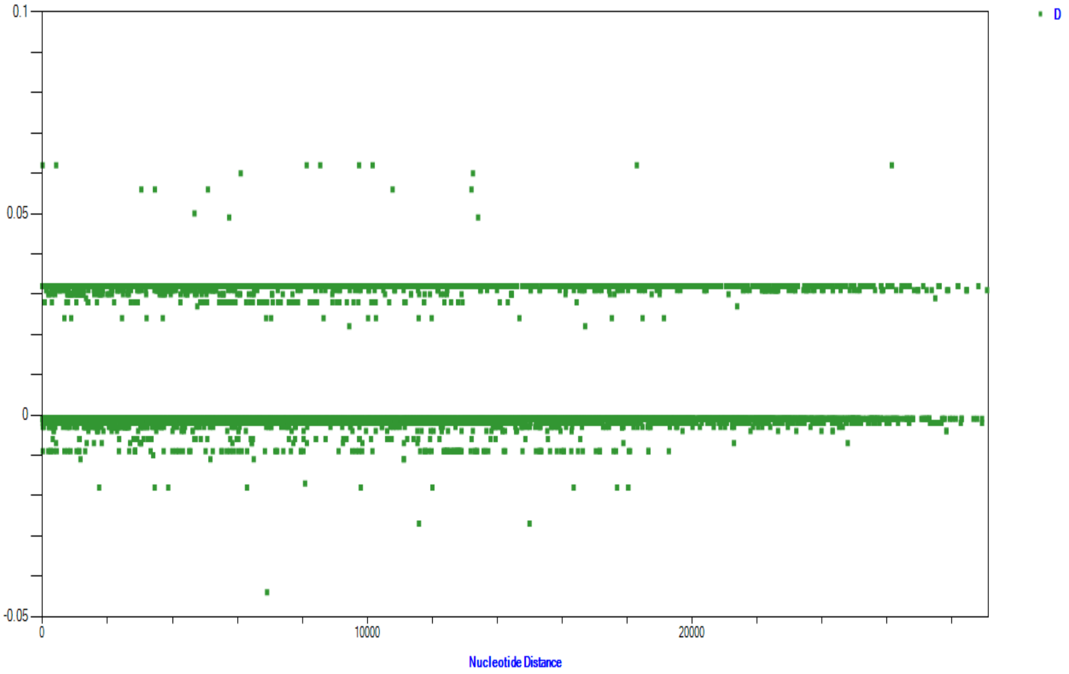


Figure 4.1.3.1: SNPs in linkage disequilibrium in different SARSCoV2 genome which were used for this study

Table 4.1.3.1 : Haplotypes obtained from sites in highly significant linkage disequilibrium

Haplotypes number	Site(I)	Site(II)	Distance(nt.)	Linkage disequilibrium (D)
1	153	26325		
2	3656	13413	9746	0.062
3	3656	13836	10163	0.062
4	3656	16872	13205	0.056
5	3656	21974	18296	0.062
6	11704	24966	13251	0.06
7	13413	13830	417	0.062
8	13413	16872	3459	0.056
9	13413	21974	8550	0.062
10	13836	16872	3042	0.056
11	13830	21974	8133	0.062
12	16872	21974	5091	0.056
13	16872	27663	10779	0.056

4.1.4.. SARSC0V2 lineages

Out of all, 30 sequences which were used in this study of B lineages in (**table 4.1.4.1).**

Table 4.1.4.1: lineages list of SARSCoV2 genomes used in this study

Sequence name	**Lineage**	**Scorpio call**
hCoV19-Denmark-DCGC-621011/2022|EPI_ISL_16063292(2022-12-03)	XBB.1.5	0M(BA.2-like)
hCoV19-Netherlands-NH-AUMC-034412/2022|EPI_ISL_16073540(2022-11-28)	XBB.1.5	0M (BA.2-like)
hCoV19-Denmark-DCGC-629181/2022|EPI_ISL_16330949(2022-12-23)	XBB.1.5	0M (BA.2-like)
hCoV19-Denmark-DCGC-634085/2022|EPI_ISL_16410349(2022-12-27)	XBB.1.5	0M (BA.2-like)
hCoV19-Denmark-DCGC-636322/2023|EPI_ISL_16485345(2023-01-02)	XBB.1.5	0M (BA.2-like)
hCoV19-Denmark-DCGC-636558/2023|EPI_ISL_16485582(2023-01-02)	XBB.1.5.7	0M (BA.2-like)
hCoV19-Netherlands-NH-AUMC-000387/2023|EPI_ISL_16723240(2023-01-19)	XBB.1.5	0M (BA.2-like)
hCoV19-Germany-HH-RKI-I-1097238/2023|EPI_ISL_16814084(2023-01-22)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-NYULH9394/2022|EPI_ISL_15851810(2022-11-14)	XBB.1.5.15	0M (BA.2-like)
hCoV19-USA-NY-NYULH9401/2022|EPI_ISL_15851817(2022-11-14)	XBB.1.5.15	0M (BA.2-like)
hCoV19-USA-NY-URMC-2211B176-1/2022|EPI_ISL_16070413(2022-11-28)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-NYULH9739/2022|EPI_ISL_16239148(2022-12-08)	XBB.1.5	0M (BA.2-like)

hCoV19-USA-NY-NYULH9745/2022|EPI_ISL_16239149(2022-12-12)	XBB.1.5.17	0M (BA.2-like)
hCoV19-USA-NY-NYULH9751/2022|EPI_ISL_16239152(2022-12-12)	XBB.1.5.20	0M (BA.2-like)
hCoV19-USA-NY-NYULH9779/2022|EPI_ISL_16239177(2022-12-12)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-PRL-221221_02B11/2022|EPI_ISL_16343798(2022-12-17)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-PRL-221221_02D24/2022|EPI_ISL_16343834(2022-12-20)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-ASC-210962826/2022|EPI_ISL_16580399(2022-12-30)	XBB.1.5.16	0M (BA.2-like)
hCoV19-USA-NY-ASC-210962829/2023|EPI_ISL_16580402(2023-01-05)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-NYULH10125/2023|EPI_ISL_16643608(2023-01-03)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-CDC-LC0986771/2023|EPI_ISL_16665477(2023-01-08)	XBB.1.5.17	0M (BA.2-like)
hCoV19-USA-NY-CDC-LC0996048/2023|EPI_ISL_16742589(2023-01-19)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-CDC-LC0997296/2023|EPI_ISL_16763392(2023-01-22)	XBB.1.5	0M (BA.2-like)
hCoV19-USA-NY-CDC-LC0997554/2023|EPI_ISL_16763761(2023-01	XBB.1.5	0M (BA.2-like)
hCoV19-Singapore-16374-2022-EPI_ISL_16181494(2022-12-12)	XBB.1.5	OM (BA.2-like)
hCoV19-India-MH-ICMR-NIV-INSACOG-G-12903-2022-EPI_ISL_16528811(2022-12-11)	XBB.1.5	0M (BA.2-like)
hCoV19-india-MH-ICMR-NIV-INSACOG-G-12908-2022-EPI_ISL_16528814(2022-12-18)	XBB.2.7	0M (BA.2-like)

hCoV19-Singapore-R5MR108-2023-EPI_ISL_16585052(2023-01-14)	XBB.1	0M (BA.2-like)
hCoV19-Singapore-R9MR1-2023-EPI_ISL_16807729(2023-01-25	XBB.1.5	0M (BA.2-like)
NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genom	B	

4.1.5. Phylogenetic Analysis

Clustering analysis of the maximum likelihood phylogenetic tree gave 3 major clades **(fig 4.1.5).**

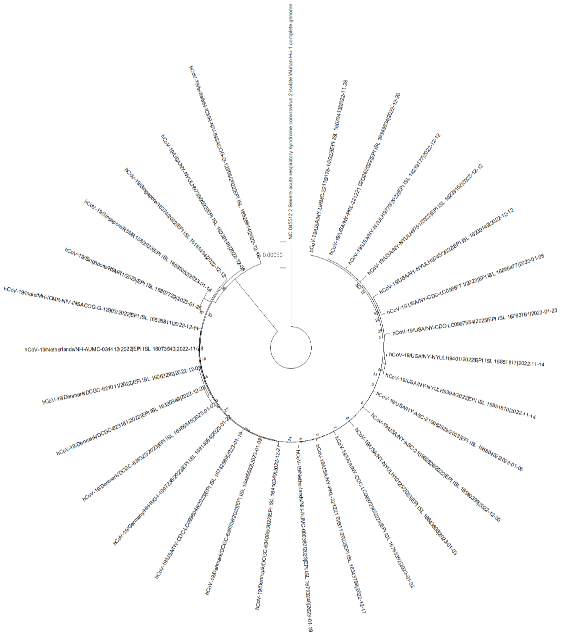

NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome
hCoV-19/USA/NY-URMC-2211B176-1/2022|EPI ISL 16070413|2022-11-28
hCoV-19/USA/NY-PRL-221221 02D24/2022|EPI ISL 16343834|2022-12-20
hCoV-19/USA/NY-NYULH9779/2022|EPI ISL 16239177|2022-12-12
hCoV-19/USA/NY-NYULH9751/2022|EPI ISL 16239152|2022-12-12
hCoV-19/USA/NY-NYULH9745/2022|EPI ISL 16239149|2022-12-12
hCoV-19/USA/NY-CDC-LC0986771/2023|EPI ISL 16665477|2023-01-08
hCoV-19/USA/NY-CDC-LC0997554/2023|EPI ISL 16763761|2023-01-23
hCoV-19/USA/NY-NYULH9401/2022|EPI ISL 15851817|2022-11-14
hCoV-19/USA/NY-NYULH9394/2022|EPI ISL 15851810|2022-11-14
hCoV-19/USA/NY-ASC-210962829/2023|EPI ISL 16680402|2023-01-05
hCoV-19/USA/NY-ASC-210962826/2022|EPI ISL 16580399|2022-12-30
hCoV-19/USA/NY-NYULH10125/2023|EPI ISL 16643608|2023-01-03
hCoV-19/USA/NY-CDC-LC0997296/2023|EPI ISL 16763392|2023-01-22
hCoV-19/USA/NY-PRL-221221 02B11/2022|EPI ISL 16343798|2022-12-17
hCoV-19/Netherlands/NH-AUMC-000387/2023|EPI ISL 16723240|2023-01-19
hCoV-19/Denmark/DCGC-634085/2022|EPI ISL 16410349|2022-12-27
hCoV-19/Denmark/DCGC-636558/2023|EPI ISL 16485582|2023-01-02
hCoV-19/USA/NY-CDC-LC0996048/2023|EPI ISL 16742589|2023-01-19
hCoV-19/Germany/HH-RKI-I-1097238/2023|EPI ISL 16814084|2023-01-24
hCoV-19/Denmark/DCGC-636322/2023|EPI ISL 16485345|2023-01-02
hCoV-19/Denmark/DCGC-629181/2022|EPI ISL 16330949|2022-12-23
hCoV-19/Denmark/DCGC-621011/2022|EPI ISL 16063292|2022-12-02
hCoV-19/Netherlands/NH-AUMC-034412/2022|EPI ISL 16073540|2022-11-28
hCoV-19/India/MH-ICMR-NIV-INSACOG-G-12903/2022|EPI ISL 16528811|2022-12-11
hCoV-19/Singapore/R9MR1/2023|EPI ISL 16807729|2023-01-25
hCoV-19/Singapore/R5MR108/2023|EPI ISL 16585052|2023-01-14
hCoV-19/Singapore/16374/2022|EPI ISL 16181494|2022-12-12
hCoV-19/USA/NY-NYULH9739/2022|EPI ISL 16239148|2022-12-08
hCoV-19/India/MH-ICMR-NIV-INSACOG-G-12908/2022|EPI ISL 16528814|2022-12-18
0.00050

Figure 4.1.5: Phylogenetic trees of 30 SARS-CoV-2 genomes

CHAPTER-5

Conclusion

The current work examined the complete genome sequences characterisation and phylogenetic reconstruction of SARSCoV2 isolates from three distinct locations, namely Asia, Europe, and North America. COVID-19 participants were examined. Only 30 of the hundreds of complete genome sequences from Asia, Europe, and North America in the GISAID database met the requirements for the purpose of the research. Given the importance of data quality for result validity we believe it is preferable to utilise 30 sequences of sufficient accuracy instead of several hundred sequences of low or doubtful integrity. The various SARS-CoV-2 data utilised for this investigation exhibited 99.9% resemblance meaning 0.01% difference to the reference genome sequence, thus being consistent with an overall worldwide trend. It's hardly unexpected that this analysis discovered 15 significant conserved genetic areas. This finding lends credence to the widely held belief that the new virus is of recent development, with an approximated origin date of between 6-oct-2019 to 11-dec-2019.A total of 60 SNPs were identified out of which 30 were synonymous SNPs and 30 were nonsynonymous SNPs. All SNPs were diallelic. The overall nucleotide diversity among the SARSCoV2 genomes analyzed was Pi=0.00042. The retrieved sequences reveales 3 major clades on a neighbor joining phylogenetic tree.

Future perspectives

It allows for the recognition of SNPs in diverse animals. The significant number of happening, fewer expenses of creating tests, and flexibility of such assays between different research facilities were considerations in support of employing SNPs for researching genetic changes among particular creatures, people, plants, or even microorganisms among a group of people. As a result, SNP has a comprehensive ambit uses & the execution of personalised medicines.

SNPs are not only accountable for changes in fundamental physical features across people in general, but they also impact variations in illness susceptibility and treatment response between individuals. Such SNPs are important in viral illnesses as well as metabolic ailments, like the Covid-19 pandemic. As a result, it is critical to take these SNPs into consideration in order to provide personalised diagnosis and treatment choices to tackle illnesses.

It is important to emphasise that none of these numerous uses of SNIPs would be conceivable without technological breakthroughs that help in discovery prophecy & testimony of SNPs. Without a doubt, advances in bioinformatics are essential for studying SNPs.

REFERENCES

[1] L. Statello, C.-J. Guo, L.-L. Chen, and M. Huarte, “Gene regulation by long non-coding RNAs and its biological functions,” *Nat. Rev. Mol. Cell Biol.*, vol. 22, no. 2, pp. 96–118, 2021.

[2] H. Hofmann and S. Pöhlmann, “Cellular entry of the SARS coronavirus,” *Trends Microbiol.*, vol. 12, no. 10, pp. 466–472, 2004.

[3] T. S. Pillay, “Gene of the month: the 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein,” *J. Clin. Pathol.*, vol. 73, no. 7, pp. 366–369, 2020.

[4] V. Mollica, A. Rizzo, and F. Massari, “The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer,” *Future Oncol.*, vol. 16, no. 27, pp. 2029–2033, 2020.

[5] A. Stang, “Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses,” *Eur. J. Epidemiol.*, vol. 25, no. 9, pp. 603–605, 2010.

[6] G. A. Wells, B. Shea, D. O’connell, J. Peterson, V. Welch, and M. Losos, *The Ottawa Hospital Research Institute*. .

[7] L. Torre-Fuentes *et al.*, “ACE2, TMPRSS2, and Furin variants and SARS-CoV-2 infection in Madrid, Spain,” *J. Med. Virol.*, vol. 93, no. 2, pp. 863–869, 2021.

[8] J. Gómez *et al.*, “The Interferon-induced transmembrane protein 3 gene (IFITM3) rs12252 C variant is associated with COVID-19,” *Cytokine*, vol. 137, no. 155354, p. 155354, 2021.

[9] J. Gómez *et al.*, “Angiotensin-converting enzymes (ACE, ACE2) gene variants and COVID-19 outcome,” *Gene*, vol. 762, no. 145102, p. 145102, 2020.

[10] T. T. Nguyen *et al.*, “Genetic diversity of SARS-CoV-2 and clinical, epidemiological characteristics of COVID-19 patients in Hanoi, Vietnam,” *PLoS One*, vol. 15, no. 11, p. e0242537, 2020.

[11] M. Hussain *et al.*, “Structural variations in human ACE2 may influence its binding with SARS-CoV-2 spike protein,” *J. Med. Virol.*, vol. 92, no. 9, pp. 1580–1586, 2020.

[12] J. Wang *et al.*, “Molecular simulation of SARS-CoV-2 spike protein binding to pangolin ACE2 or human ACE2 natural variants reveals altered susceptibility to infection,” *J. Gen. Virol.*, vol. 101, no. 9, pp. 921–924, 2020.

[13] C. Strafella *et al.*, “Analysis of ACE2 genetic variability among populations highlights a possible link with COVID-19-related neurological complications,” *Genes (Basel)*, vol. 11, no. 7, p. 741, 2020.

[14] A. Srivastava *et al.*, “Genetic association of ACE2 rs2285666 polymorphism with COVID-19 spatial distribution in India,” *Front. Genet.*, vol. 11, p. 564741, 2020.

[15] K. Fujikura and K. Uesaka, “Genetic variations in the human severe acute respiratory syndrome coronavirus receptor ACE2 and serine protease TMPRSS2,” *J. Clin. Pathol.*, vol. 74, no. 5, pp. 307–313, 2021.

[16] J. Sieńko *et al.*, “COVID-19: The influence of ACE genotype and ACE-I and ARBs on the course of SARS-CoV-2 infection in elderly patients,” *Clin. Interv. Aging*, vol. 15, pp. 1231–1240, 2020.

[17] A. Paniri, M. M. Hosseini, M. Moballegh-Eslam, and H. Akhavan-Niaki, “Comprehensive in silico identification of impacts of ACE2 SNPs on COVID-19 susceptibility in different populations,” *Gene Rep.*, vol. 22, no. 100979, p. 100979, 2021.

[18] S. Senapati, S. Kumar, A. K. Singh, P. Banerjee, and S. Bhagavatula, "Assessment of risk conferred by coding and regulatory variations from TMPRSS2 and CD26 in susceptibility of SARS-CoV-2 infection in human," 2020.

[19] A. Novelli *et al.*, "Analysis of ACE2 genetic variants in 131 Italian SARS-CoV-2-positive patients," *Hum. Genomics*, vol. 14, no. 1, p. 29, 2020.

[20] G. Vargas-Alarcón, R. Posadas-Sánchez, and J. Ramírez-Bello, "Variability in genes related to SARS-CoV-2 entry into host cells (ACE2, TMPRSS2, TMPRSS11A, ELANE, and CTSL) and its potential use in association studies," *Life Sci.*, vol. 260, no. 118313, p. 118313, 2020.

[21] E. Benetti *et al.*, "ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population," *Eur. J. Hum. Genet.*, vol. 28, no. 11, pp. 1602–1614, 2020.

[22] A. E. Shikov *et al.*, "Analysis of the spectrum of ACE2 variation suggests a possible influence of rare and common variants on susceptibility to COVID-19 and severity of outcome," *Front. Genet.*, vol. 11, p. 551220, 2020.

[23] Y.-C. Kim and B.-H. Jeong, "Strong correlation between the case fatality rate of COVID-19 and the rs6598045 single nucleotide polymorphism (SNP) of the interferon-induced transmembrane protein 3 (IFITM3) gene at the population-level," *Genes (Basel)*, vol. 12, no. 1, p. 42, 2020.

[24] A. K. Maiti, "The African-American population with a low allele frequency of SNP rs1990760 (T allele) in IFIH1 predicts less IFN-beta expression and potential vulnerability to COVID-19 infection," *Immunogenetics*, vol. 72, no. 6–7, pp. 387–391, 2020.

[25] H.-C. Yang *et al.*, "Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 48, pp. 30679–30686, 2020.

[26] L. M. Irham, W.-H. Chou, M. J. Calkins, W. Adikusuma, S.-L. Hsieh, and W.-C. Chang, "Genetic variants that influence SARS-CoV-2 receptor TMPRSS2 expression among population cohorts from multiple continents," *Biochem. Biophys. Res. Commu.*

[27] Abdel-Moneim, A. S., & Abdelwhab, E. M. (2020). Evidence for SARS-CoV-2 infection of animal hosts. *Pathogens*, *9*(7), 529.

[28] Alotaibi, F., Alharbi, N. K., Rosen, L. B., Asiri, A. Y., Assiri, A. M., Balkhy, H. H., Al Jeraisy, M., Mandourah, Y., AlJohani, S., Al Harbi, S., Jokhdar, H. A. A., Deeb, A. M., Memish, Z. A., Jose, J., Ghazal, S., Al Faraj, S., Al Mekhlafi, G. A., Sherbeeni, N. M., Elzein, F. E., … Saudi Critical Care Trials Group. (2023). Type I interferon autoantibodies in hospitalized patients with Middle East respiratory syndrome and association with outcomes and treatment effect of interferon beta-1b in MIRACLE clinical trial. *Influenza and Other Respiratory Viruses*, *17*(3), e13116

[29] Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, *26*(4), 450–452.

[30] Baddal, B., & Cakir, N. (2020). Co-infection of MERS-CoV and SARS-CoV-2 in the same host: A silent threat. *Journal of Infection and Public Health*, *13*(9), 1251–1252.

[31] Chen, J., Bai, H., Liu, J., Chen, G., Liao, Q., Yang, J., Wu, P., Wei, J., Ma, D., Chen, G., Ai, J., & Li, K. (2020). Distinct clinical characteristics and risk factors for mortality in female inpatients with Coronavirus disease 2019 (COVID-19): A sex-stratified, large-scale cohort study in Wuhan, China. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, *71*(12), 3188–3195.

[32] Cheng, V. C. C., Lau, S. K. P., Woo, P. C. Y., & Yuen, K. Y. (2007). Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clinical Microbiology Reviews*, *20*(4), 660–694.

[33] Cui, J., Li, F., & Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature Reviews. Microbiology*, *17*(3), 181–192.

[34] Fouchier, R. A. M., Kuiken, T., Schutten, M., van Amerongen, G., van Doornum, G. J. J., van den Hoogen, B. G., Peiris, M., Lim, W., Stöhr, K., & Osterhaus, A. D. M. E. (2003). Aetiology: Koch's postulates fulfilled for SARS virus: Aetiology. *Nature*, *423*(6937), 240.

[34] Happi, C., Ihekweazu, C., Oluniyi, P. E., & Olawoye, I. (2020). SARS-CoV-2 Genomes from Nigeria Reveal Community Transmission, Multiple Virus Lineages and Spike Protein Mutation Associated with Higher Transmission and Pathogenicity. *Genome Rep*.

[35] Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., Hastie, K. M., Parker, M. D., Partridge, D. G., Evans, C. M., Freeman, T. M., de Silva, T. I., Sheffield COVID-19 Genomics Group, McDanal, C., Perez, L. G., … Montefiori, D. C. (2020). Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, *182*(4), 812-827.e19.

[36] Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549.

[37] Li, D., Zhang, J., & Li, J. (2020). Primer design for quantitative real-time PCR for the emerging Coronavirus SARS-CoV-2. *Theranostics*, *10*(16), 7150–7162.

[38] Paraskevis, D., Kostaki, E. G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., & Tsiodras, S. (2020). Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, *79*(104212), 104212.

[39] Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407.

[40] Sah, R., Rodriguez-Morales, A. J., Jha, R., Chu, D., Gu, H., & Peiris, M. (2020). Complete genome sequence of a 2019 novel coronavirus (SARS-CoV2) strain isolated in Nepal. *Microbiol Res Announc*, *9*(1).

[41] Somasundaram, N. P., Ranathunga, I., Ratnasamy, V., Wijewickrama, P. S. A., Dissanayake, H. A., Yogendranathan, N., Gamage, K. K. K., de Silva, N. L., Sumanatilleke, M., Katulanda, P., & Grossman, A. B. (2020). The impact of SARS-CoV-2 virus infection on the endocrine system. *Journal of the Endocrine Society*, *4*(8), bvaa082. https://doi.org/10.1210/jendso/bvaa082

[42] Sternberg, A., & Naujokat, C. (2020). Structural features of coronavirus SARS-CoV-2 spike protein: Targets for vaccination. *Life Sciences*, *257*(118056), 118056.

[43] Su, W., Choy, K. T., Gu, H., Sia, S. F., Cheng, K. M., Nizami, S. I. N., Krishnan, P., Ng, Y. M., Chang, L. D. J., Liu, Y., Cheng, S. M. S., Peiris, M., Poon, L. L. M., Nicholls, J. M., & Yen, H.-L. (2023). Reduced pathogenicity and transmission potential of Omicron BA.1 and BA.2 sublineages compared with the early severe acute respiratory syndrome Coronavirus 2 D614G variant in Syrian hamsters. *The Journal of Infectious Diseases*, *227*(10), 1143–1152.

[43] Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512–526.

[44] Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., Cui, J., & Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, *7*(6), 1012–1023.

[45] Van Blerkom, L. M. (2003). Role of viruses in human evolution. *American Journal of Physical Anthropology*, *Suppl 37*(S37), 14–46.

[46] Van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., & Ormond, L. (2020). Emergence of genomic diversity and recurrent mutations in SARSCoV-2. Infection, genetics and evolution. Epub 05/05. *J Mol Epidemiol Evol Genet Infect Dise*, *83*.

[47] Vankadari, N. (2020). Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. *Gene*, *752*(144792), 144792.

[48] 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. *Journal of Medical Virology*, *92*(6), 667–674.

[49] Woo, P. C. Y., Huang, Y., Lau, S. K. P., & Yuen, K.-Y. (2010). Coronavirus genomics and bioinformatics analysis. *Viruses*, *2*(8), 1804–1820.

[50] N. Vankadari, “Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution,” *Gene*, vol. 752, no. 144792, p. 144792, 2020.