SEQUENCE ANALYSIS OF SNPS AND PHYLOGENETIC ANALYSIS FOR COVID-19 AMONG INDIAN POPULATION

PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY

in

BIOTECHNOLOGY / BIOINFORMATICS

By

DEEVYANI KAPLAS (201804)

VAIBHAV SHARMA (201907)

Under the guidance & supervision of

DR. SHIKHA MITTAL

to

Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

CERTIFICATE

This is to certify that the project report entitled "**Sequence Analysis of SNPs and Phylogenetic Analysis for COVID-19 Among Indian Population**" submitted to the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, in partial fulfilment of the requirements for the award of the degree of B.Tech in Biotechnology and Bioinformatics, is an authentic record of the work completed by Deevyani Kaplas (201804) and Vaibhav Sharma (201907) during the period from August 2023 to May 2024 under the supervision of Dr. Shikha Mittal, Assistant Professor, Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat.

----------	**----------**
(STUDENT SIGNATURE)	(STUDENT SIGNATURE)
Student Name: **Deevyani Kaplas**	Student Name: **Vaibhav Sharma**
Roll No. **201804**	Roll No. **201907**
DATE:	DATE:

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled **" Sequence Analysis of SNPs and Phylogenetic Analysis for COVID-19 Among Indian Population"** in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology** in **Biotechnology / Bioinformatics** submitted in the Department of Biotechnology & Bioinformatics**,** Jaypee University of Information Technology Waknaghat is an authentic record of my work carried out throughout August 2023 to May 2024 under the supervision of **Dr Shikha Mittal** (Assistant Professor, Department of Biotechnology &Bioinformatics)

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

----------	**----------**
(STUDENT SIGNATURE)	(STUDENT SIGNATURE)
Student Name: **Deevyani Kaplas**	Student Name: **Vaibhav Sharma**
Roll No. **201804**	Roll No. **201907**
DATE:	DATE:

SUPERVISOR'S DECLARATION

This is to certify that the project report entitled "**Sequence Analysis of SNPs and Phylogenetic Analysis for COVID-19 Among Indian Population**" submitted to the Department of Biotechnology & Bioinformatics, Jaypee University of Information Technology, Waknaghat, in partial fulfilment of the requirements for the award of the degree of B.Tech in Biotechnology and Bioinformatics, is an authentic record of the work completed by Deevyani Kaplas (201804) and Vaibhav Sharma (201907) during the period from August 2023 to May 2024 under my supervision and guidance. To the best of my knowledge, the candidate's declaration is true, and the work has not been submitted elsewhere for a degree.

SIGNATURE: _____

DATE:

Dr. Shikha Mittal
Assistant Professor
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology, Waknaghat

ACKNOWLEDGEMENT

"The only way to do great work is to love what you do." – Steve Jobs

Every project big or small is successful largely due to the effort of several wonderful people who have always given their valuable advice or lent a helping hand. We sincerely appreciate the inspiration; support and guidance of all those people who have been instrumental in making this project a success.

First and foremost, our sincere appreciation goes to **Dr. Shikha Mittal**, whose invaluable guidance, continuous support, and insightful feedback were instrumental in shaping the direction of our research and ensuring its successful completion. Her encouragement and expertise provided us with the necessary tools and motivation to achieve our goals.

We are also immensely grateful to the **Department of Biotechnology and Bioinformatics** at **Jaypee University of Information Technology** (JUIT), Waknaghat, for providing us with the resources and facilities required for this project. The support from the faculty and staff has been crucial in allowing us to pursue our research with confidence and dedication.

Our heartfelt appreciation extends to our family members, whose unwavering support and encouragement have been a constant source of strength throughout our academic journey. Their belief in us and their patience have been invaluable.

Additionally, we would like to express our gratitude to our friends and peers, who have offered their support, advice, and camaraderie, making our journey through this project both productive and enjoyable.

TABLE OF CONTENT

LIST OF ABBREVIATIONS

ABBREVIATIONS	FULL FORM
SNP	Single-nucleotide polymorphisms
MERS	Middle East Respiratory Syndrome
SARS	Severe Acute Respiratory Syndrome
HCoV	Human Coronavirus
ARDS	Acute respiratory distress syndrome
GWAS	Genome-Wide Associated Studies
WHO	World Health Organization
ACE2	Angiotensin converting enzyme 2
ICTV	International Committee on Taxonomy of Viruses
PDB	Protein Data Bank
RAAS	Renin-AngioTensin-Aldosterone System
RNA	Ribonucleic acid
TMPRSS2	Transmembrane Protease Serine 2

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The order Nidovirales contains the family Coronaviridae, which includes the coronaviruses. CoVs can be lethal and are present around the world. They can infect a wide range of animals and people and cause disorders such as gastrointestinal tract infections, encephalitis, and demyelination. The family Coronaviridae includes a wide range of human and animal viruses, each of which has a unique morphology. Since many years ago, coronaviruses have been recognized to infect people. The COVID-19 pandemic was sparked by the recently discovered coronavirus SARS-CoV-2.

A genetic variation at a single base location in the DNA is known as a single nucleotide polymorphism. The influence of SNPs on qualities like health, sickness, treatment responsiveness, and other attributes is a topic of research for scientists. The study that was conducted is assisting in our understanding of SNP in connection to coronavirus illness. Its intricate patterns differ between individuals and even populations. How SNP markers contribute to the various effects of the 2019 Covid study

CHAPTER 1

INTRODUCTION

1.1 **INTRODUCTION**-

The coronaviruses are members of the order Nidovirales' Coronaviridae family. CoVs are widespread and spread disease by infecting a wide range of animals and people. For numerous decades, people have been known to contract coronaviruses. There are four endemic subtypes of the virus: HCoV (human coronavirus) −229E, -NL63, -OC43, and -HKU1 [1].

The world saw its first outbreak of severe acute respiratory syndrome (SARS) on November 16, 2002 [2], when symptoms unique to SARS appeared in Foshan municipality, Guangdong Province, China. By 2003, further case clusters were recorded from Canada, Hong Kong, Taiwan, Singapore, Vietnam, and the Chinese mainland. Eventually, these instances spread to 32 different nations or regions[3][4].

After a decade from this epidemic of SARS (severe acute respiratory syndrome), 2019 witnessed another disease, referred to as novel coronavirus caused by the SARS-CoV virus. This outbreak when studied by scientists and researchers brought in the conclusion to name this virus as SARS-CoV-2 and introduced COVID-19 with the news report given by “Johns Hopkins University” the resource center of Coronavirus at Wuhan, China.
By January 2021 more than 144 million cases were reported for covid-19 internationally with over 3 million mortality rates and it continues [5].

1.2 PROBLEM STATEMENT

Upon analysing the structural makeup and virology of COVID-19, we discovered that SARS-CoV-2 is composed of three proteins: phosphorylated nucleocapsid protein (N) within the viral envelope, transmembrane glycoprotein (M), and glycoprotein S (spike), which assembles into bulky peplomers. A minor

transmembrane protein E is also found in coronaviruses, and some of them have an additional envelope protein called HE that performs both hemagglutination and esterase functions [6]. In "Figure 2.2.1". We also found that the S protein comprises of two subunits S1 & S2 which are the initial contact sites with human cells.

As human cells have different receptors on the surface like CD4, CD46, DC-SIGN receptor, ACE2 etc. they become the initial contact of any virus to enter the human body. The spike protein on the host surface is allowed to bind with the ACE2 receptor and follow its replication cycle. Thus, it is very important to understand how the replication of COVID-19 takes place and how mutation occurs.

As the studies were conducted it was noted that the COVID-19 viral protein targets some genes which are present on the host cell and were associated with SNP and were used to study effects and changes it brought on the host and virus genome.

It is usually observed that whenever any change in nucleotide happens in a sequence, alteration of gene and genetic codes takes place. The most common change is seen in SNP of DNA in both coding and non-coding regions. Single Nucleotide Polymorphism occurs due to variations in base pair of DNA. Adenine [A], Cytosine [C], Guanine [G], and Thymine [T] are the four nucleotides that make up the DNA sequence. SNP can occur when in any sequence A gets replaced with T or C may replace G. Most SNP occurs during the replication cycle of DNA but can be caused due to environmental factors. Sometimes SNP can be inherited from parents resulting in effects on gene function associated with certain diseases. For example, SNPs are linked in development of certain cancers, while others may be associated to increase immunity to infections.

Study of SNP is also important as they act as markers to examine sequences for plants and animals and the changes they bring. The detection of single-nucleotide polymorphisms (SNPs) can provide valuable information on the genetic diversity of the virus and how it has spread across different regions. They identify regions of the genome that are associated with disease or traits of interest. The markers help in studying the risk factors and susceptibility. We can compare the frequency

of specific SNPs, for example, genetic tests through SNP markers can identify genetic disorders like cystic fibrosis or response to certain medications [7][8].

Evaluating the frequency of particular SNPs in COVID-19 patients can aid in the management with their health and identify variations or mutations associated with a higher or lower risk of experiencing severe symptoms or issues from the virus.

1.3 OBJECTIVE

1. The objective of this study was to understand how SNPs can be used in COVID-19 research among genome-wide-associated studies (GWAS). By understanding the genotype of thousands and even millions of SNPs across the genome in a large no. of individuals, and then putting a comparison study of specific SNPs between cases of COVID-19 patients along with healthy individuals.

A recent study under GWAS identified SNPs with ACE2 gene in people with covid symptoms and several SNPs in the ABO blood group loci. [9]

2. SARS-CoV-2 phylogenetic reconstruction from various samples taken for the study which were affected with COVID-19.

CHAPTER 2

REVIEW OF LITERATURE

The coronaviruses are members of the order Nidovirales' Coronaviridae family. Worldwide, CoVs infect a variety of animals, resulting in diseases like encephalitis, demyelination, and gastrointestinal tract infections, all of which can be fatal [10].

Table 2.1: Virus Classification

Organism	virus
Realm	Riboviria
Kingdom	Orthornavirae
Phylum	Pisuviricota
Class	Pisoniviricetes
Order	Nidovirales
Family	Coronaviridae

The history highlights the very first evidence of coronavirus in domestic chicken in North America During the late 1920s which marked the emergence of the coronavirus[11]. After which isolation of the virus took place in 1933 with the cultivation of the virus took place for the very first time in 1937. In the late 1940s, two more coronaviruses were discovered: JHM, which causes mouse encephalitis, and MHV, which causes mouse hepatitis.

Coronaviruses has been recognized to contaminate people for numerous years. There are four endemic subtypes of coronaviruses: HCoV (human coronavirus) −229E, -NL63, -OC43 and -HKU1. Those specifically cause a moderate higher breathing or respiratory disease; nevertheless, in vulnerable people, they could

induce a more severe respiratory disorder and, less frequently, CNS dysfunction. [12].

In Foshan Municipality, Guangdong Province, China, the first case of severe acute respiratory syndrome (SARS) was reported on November 16, 2002. Mainland China, Hong Kong, Taiwan, Singapore, Vietnam, Canada, and 32 other countries or areas were eventually hit by the SARS pandemic[13].

On March 12 of that year, WHO issued the first worldwide notification in response to a cluster of cases of highly uncommon pneumonia in hospitals in Guangdong, Hanoi, and Hong Kong [14]. Once the scientific study of SARS was finished, the researchers discovered that the infectious agent responsible for SARS in April 2003 was a single coronavirus, or the SARS-CoV[2]. During the 2002–2003 SARS pandemic, 8437 potential cases of the disease had been recorded, 813 of which had already passed away.

Similar to COVID-19, SARS causes symptoms such as fever, chills, coughing, headaches, and body pains in addition to coughing as well as feeling short of breath. However, SARS was more severe than COVID-19, with a higher mortality rate. Among 29 countries, 8,000 people had been infected with SARS and 774 had died by the time the outbreak was contained in July 2003 (WHO 2003c) . At the time of the outbreak's containment, over 8,000 people had been infected and 774 died[15].

The SARS outbreak had a vital impression on global public health, leading to the development of new protocols and guidelines for controlling infectious diseases. It also highlighted the importance of global cooperation and information sharing in responding to health emergencies. The experience gained during the SARS outbreak has been used in answering and reacting to the COVID-19 pandemic situation, including the development of diagnostic tests, treatments, and vaccines[16].

Upper respiratory tract infections and gastrointestinal tract infections that self-limit themselves have frequently been linked to human coronaviruses (hCoVs). However, it has recently become evident that hCoVs can cause more serious respiratory tract infections, including bronchitis, pneumonia, and acute respiratory distress syndrome (ARDS), and may even result in death. Currently, seven CoVs

are studied to infect humans, with four "common cold" CoVs circulating worldwide every year. On the other hand, the three remaining strains, on the other hand, are more pathogenic and cause outbreaks with high mortality rates [17].

Numerous human and animal viruses with distinctive characteristics are members of the family Coronaviridae. All viruses are spherical and contained, except for toroviruses, which produce virions that are kidney-, disc-, or rod-shaped. Encircling every particle is a fringe, or "corona," comprised of the bulbous distal ends of encapsulated envelope glycoproteins. Until 2003, it was believed that coronaviruses apart from this one had a predominant connection with the livestock industry, and that members of this family exclusively affected people who had moderate respiratory illnesses.

However, the exposure to the severe acute respiratory syndrome virus (SARS-CoV) that year rekindled interest in these important viruses, leading to the discovery of numerous new coronaviruses, some of which have the potential to spread through zoonotic means and cause serious disease outbreaks in humans. MERS-CoV's more recent emergence serves as an example. Furthermore, coronaviruses are known to possess the biggest genomes of positive-sense RNA. The main mechanism via which coronavirus genes are expressed is complicated and generates nested mRNA transcripts, which are subsequently regulated to control the replication cycle.

It was in 2019 when the deadly coronavirus (COVID-19) or SARS-CoV-2 broke out to become the biggest risk to public health.

On 31 December 2019, the initial knowledge of the virus was introduced to the world when the outbreak news report was given by “Johns Hopkins University” the resource centre of Coronavirus at Wuhan, China.

By January 2021 more than 144 million cases were reported for covid-19 internationally with over 3 million mortality rates and it continued.

2.1 PROPERTIES OF CORONAVIRUSES

2.1.1 CLASSIFICATION

One-third of the families are made up of the Coronaviridae. Bird and insect infections are found in the two additional RNA virus families in the order Nidovirales, Arteriviridae and Roniviridae. Within this family, there are two subfamilies: Coronavirinae and Torovirinae. The latter is mostly responsible for gastrointestinal infections in horses, cattle, pigs, cats, and goats. Members of the subfamily Torovirinae are not further addressed since, despite their economic significance, they have yet to be proven to cause human infection. Coronaviruses are similar in appearance and contain single-stranded RNA genomes up to 30 kb in size.

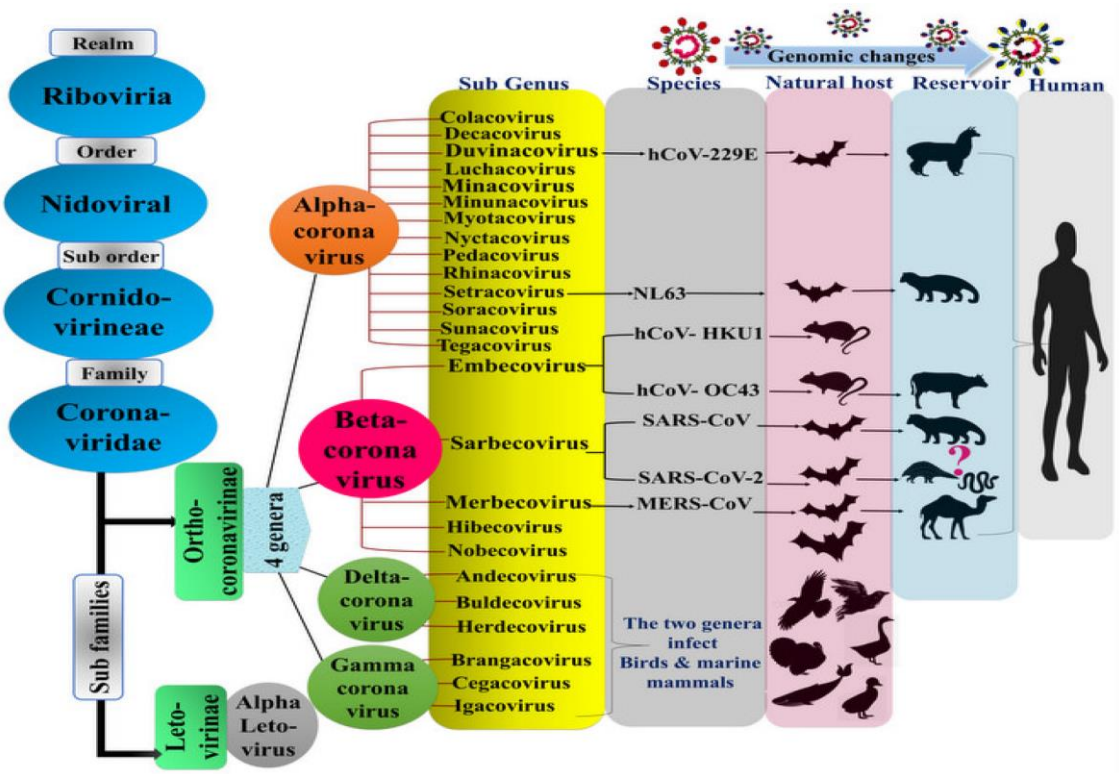


Figure 2.1.1.1 The classification of the SARS-CoV-2 is shown in the coronavirus taxonomy as determined by the International Committee on Taxonomy of Viruses (ICTV). In a tortuous evolutionary pattern, structure, and genome, the virus was able to reemerge and jump from its original host (bats) to an intermediate host, and then finally to humans[18].

2.1.2 FOUR GENERA MAKE UP THE SUBFAMILY CORONAVIRINAE'S MEMBERS.

a) The genus Alpha-coronavirus contains two kind of the human virus HCoV-229E, and other (HCoV-NL63), and many animal viruses. Further it is classified into two subgenera: Tegacovirus and Duvinacovirus. The viruses can cause symptoms of fever, cough, respiratory distress etc. The structure of alphacoronaviruses is similar to that of other coronaviruses, with a distinctive "spike" protein on their surfaces that allows them to enter human cells [19].
It is also possible for alphacoronaviruses to mutate, resulting in new strains with different characteristics as with other coronaviruses.

The Middle Eastern respiratory syndrome (MERS), the mouse hepatitis virus prototype, three human viruses, SARS-HCoV, HCoV-HKU1, the SARS related coronavirus, and a group of animal coronaviruses are all included in the Betacoronaviruses.

There are four subgenera for betacoronaviruses: Embecovirus, Hibecovirus, Merbecovirus, and Nobecovirus. Mild to severe symptoms, such as fever, coughing, and shortness of breath, can be brought on by viral respiratory illnesses.
In addition to their characteristic spike protein on their surface, betacoronaviruses have receptors on the surface of human cells that enable them to attach to them and facilitate their entry into the body. A vaccine and some therapeutics have been developed based on this spike protein. In addition to their impact on human health, they also have significant economic and social impacts. These impacts include travel restrictions, business shutdowns, school closures etc.
The emergence of new betacoronavirus strains has increased the importance of ongoing surveillance and research into these viruses which is to prevent and control future outbreaks.

b) The genus of Gamma-coronavirus are said to have set of viruses for cetaceans (whales) and birds. It also has its sub genera: colecovirus and Igacovirus. This virus mostly causes disease in birds by infecting their respiratory and digestive tract with (IBV) infectious bronchitis virus. Like other strains of coronavirus, the

"spike" protein can bind to the surface of human cell and mutate into new strains with different characteristics. In humans, some gammacoronaviruses have been detected, such as the human coronavirus HKU1 (HCoV-HKU1). HCoV-HKU1 can cause mild respiratory illnesses.

c) The genus Delta-coronavirus contains viruses isolated from pigs and birds. Aside from the word "delta" being described as a variation of the latest SARS-CoV-2 virus, which caused COVID-19, no such description of the delta coronavirus is currently referenced. This variant is also known as B.1617.2 and was identified in India December 2020 and spread to most of the countries. The delta variant has several mutations in the spike protein of the virus, which may make it more transmissible and potentially more resistant to some treatments and vaccines. According to preliminary research, compared to other variants, the delta variant may be linked to a higher risk of hospitalisation and mortality.

Table 2.1.2.1: Human Corona Virus Strains

Structure	**Human Coronavirus Strain**	**Disease**
Alphacoronavirus	HCoV-229E	Usually, mild respiratory illness
	HCoV-NL63	-NA-
Betacoronavirus	HCoV-OC43	-NA-
	HCoV-HKU1	-NA-
	MERS- CoV	Middle East Respiratory Syndrome (MERS)
	SARS-CoV	Severe Acute Respiratory Syndrome (SARS)
	SARS-CoV-2	Covid-19

Since 2005, a large number of novel coronaviruses have been identified from bats, therefore there is evidence that coronaviruses that cause MERS and SARS in humans, as well as other coronaviruses, could have originated from progenitor bat viruses.

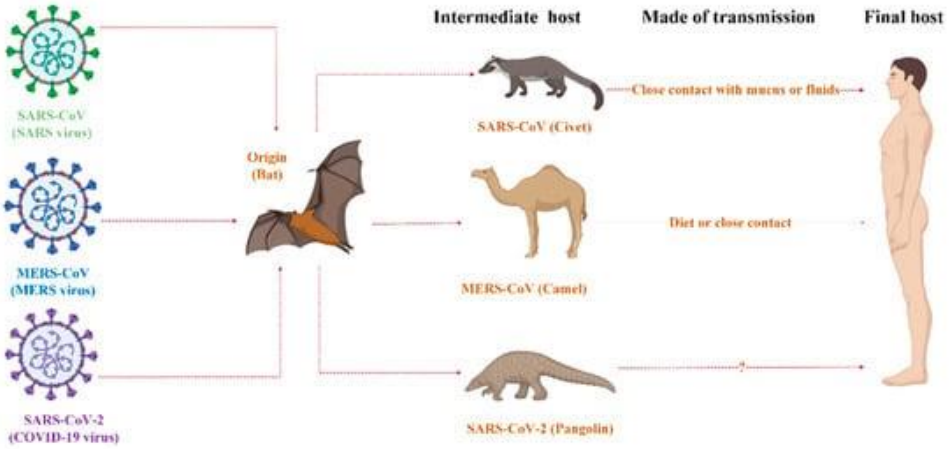


Figure 2.1.2.1: Human coronaviruses' animal sources (Compiled using BioRender.com; retrieved on August 1, 2020)

2.2 VIROLOGY

The three primary structural proteins found in coronavirus virions are the large (200 K) glycoprotein S (for spike), which forms the bulky (15 to 20 nm) peplomers visible in the viral envelope; a distinct transmembrane glycoprotein (M); and the inner phosphorylated nucleocapsid protein (N). Certain coronaviruses additionally include an extra envelope protein (HE) with esterase and hemagglutination activity, as well as a small transmembrane protein E [19].

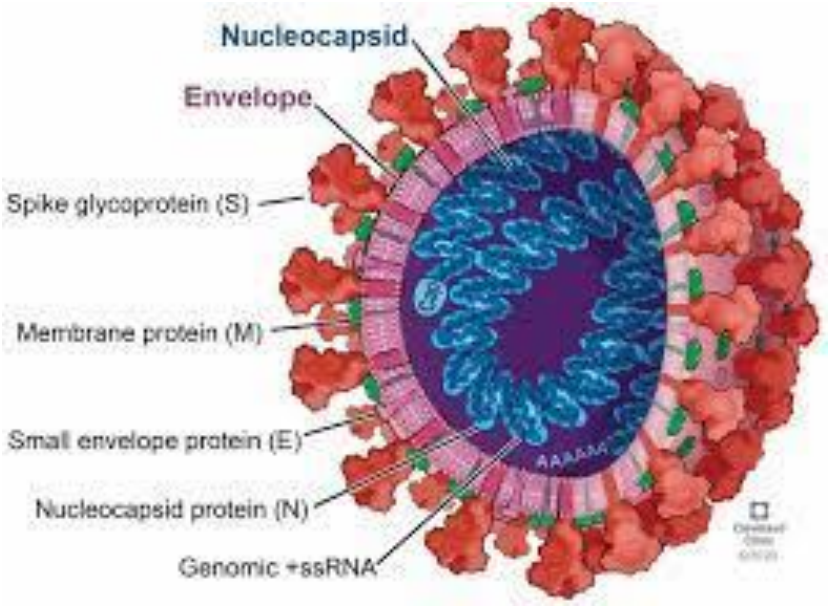


Figure 2.2.1: Structure of coronavirus

Structure: The S glycoprotein from the recently identified SARS-CoV-2 is composed of two subunits, S1 and S2, which are depicted by a sword-like spike in "Figure 2.2.2." Crystallography is another method that may be used to discern the structure of the (S) protein. [6][27]

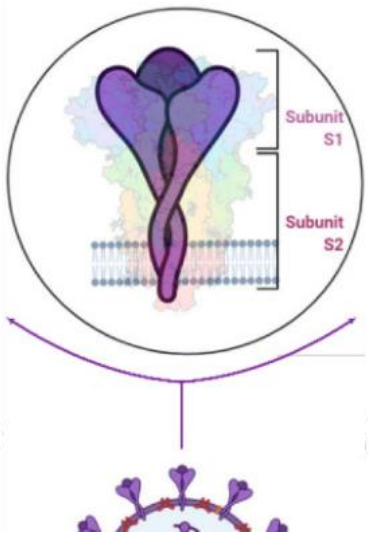


Figure 2.2.2: SARS-CoV-2 S protein (Reprinted from "An In-depth Look into the Structure of the SARS-CoV-2 Spike Glycoprotein", by BioRender.com, accessed on 1 August 2020)[27]

This glycoprotein's Protein Data Bank (PDB) model (PDB ID: 6VXX-PDB) demonstrates that the subunits are composed of numerous regions that are critical to the infection process. "Figure 2.2.3" indicates this. S1 and S2 are linked together by a polybasic amino acid bridge, which may be essential for comprehending when exploring viral targeting. [19]

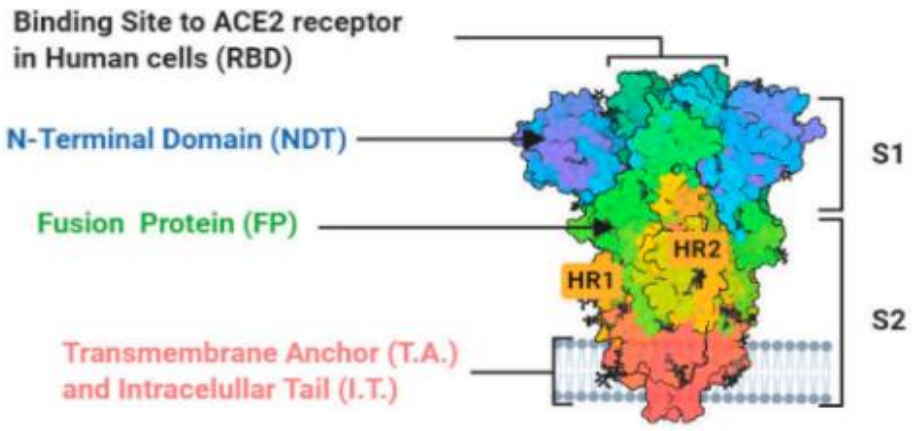


Figure 2.2.3: Virus spike protein, crystallographic structure under electron microscope

This virus is said to be an enveloped virus with +ve sense single-strand RNA genome. The spike protein on its surface allows it to bind to ACE2 receptor on human cells and the follow its replication cycle. [28]

The largest known viral RNA genome is of 30 kb of +ve sense, SS RNA. They have a 3′-terminal polyadenylation and a 5′-end capping, and they are contagious. Due to the size, sets of fixed mRNAs with the same 5′-end sequence are produced, which can result in a complex process for the expression of a single gene. [20] The result of heterologous RNA recombination may lead to large-scale rearrangements. At the 5′-End genome is an untranslated (UTR) sequence of 65-98 nucleotides, termed the leader RNA, which are said to be available at the 5′-Ends of all sub-genomic. There is one more untranslated sequence of 200–500 nucleotides at the 3′ end of the RNA genome, followed by a poly(A) tail. In order to replicate and transcribe RNA, both untranslated regions are crucial.

There are 7 to 14 open reading frames (ORFs) in the genome of Coronavirus . The 5′-end for Gene 1 is made up by 2/3 of the genome with approximately 20 to 22 kb long. It has two overlapping ORFs (1a and 1b), which together function as viral RNA polymerase (Pol), are present. The following order describes the other structural protein genes to be: 5′-S-E-M-N (spike, envelope, membrane, nucleocapsid). These genes contain a number of ORFs that, when present, encode non-structural proteins and HE glycoproteins.

Although they are conserved within the same serogroup, the number, nucleotide sequence, gene order, and mode of expression of each gene vary between coronaviruses. Other coronaviruses lack a few minor ORFs seen in the 3′ region of the SARS-CoV genome. These ORFs are expected to encode eight new auxiliary proteins. "Serum isolated from SARS patients has shown antibodies reactive against all SARS-CoV proteins, indicating that these proteins are expressed by the virus in vivo" [20].

2.3 EPIDEMIOLOGY

The epidemiological characteristics for covid-19 includes the rate of transmission, its incubation, risk factors and its fatality rate.

a) Transmission- Infected persons spread Covid-19 through respiratory droplets when talking, coughing, or sneezing. A virus is seen to be spread by touching the infected or contaminated surface and then coming in contact with the mouth, nose, or eyes.

b) Incubation Period - When seen for COVID-19 its around 5-6 days but can range from 2-14 days. During this time, any infected person may not have any symptoms but can still transmit the virus to others.

c) Risk factor- Advanced age, underlying medical disorders including diabetes, cardiovascular disease, obesity, etc., and certain ethnic and racial groupings are risk factors for severe illness and fatality from COVID-19.

d) Asymptomatic cases- which were transmission by individuals who did not exhibit any symptoms which created challenge to manage and control transmission.

e) Clinical Feature- the clinical aspect with covid-19 varied in 2019 situation as different individuals experienced mild to no symptoms while other developed sever distress or multi-organ failure.

Around the world, researchers have studied different aspects of viruses, including their clinical characteristics, which range from asymptomatic infection to severe respiratory distress and multi-organ failure.

The most reoccurring symptoms of COVID-19 included fever along coughing with shortness of breath. Other symptoms noted in course of virus attacking the host mechanism of the immune system included fatigue, muscle & body aches, headaches, sore throat, loss of taste and smell. Many people explained their experienced gastrointestinal, neurological, and dermatological symptoms. Severity of covid-19 ranged from mild to severe & required hospitalization and intensive care. Risk factors at severe disease and mortality include advanced age, underlying medical conditions, and certain ethnic and racial groups.

The diagnosis of covid-19 started with laboratory testing, which includes PCR and antigen tests. Chest X-Rays and CT scans were also considered to evaluate the severity of respiratory illness. The treatment varied according to the severity of disease. Mild cased required supportive care, which severe cases required oxygen therapy, mechanical ventilation, and other interventions.

People with COVID-19 long-term effects reported feeling exhausted and had trouble breathing while going about their regular lives. These symptoms persisted for several months beyond the initial infection. These long-term effects are also being referred to as " Long COVID" and the research is still ongoing on the treatment for the same.

Vaccines are another important aspect for treatment of covid or any viral infection[21]. Vaccines against covid-19 has developed to effectively prevent

severe disease and hospitalization. Continuous campaigns for vaccination were implemented to control spread of virus worldwide.

The COVID-19 pandemic has been a significant global health crisis since its emergence in late 2019 as given researchers around the world to study various aspects of the virus from different angles and through medical professionals working to contain it. It was seen that the SARS-CoV-2 coronavirus was significant as discovery of some gene expression of viral protein on host cell were associated with SNP helped to understand transmission, prevalence and help to study mutation and expression of infection.

Sequencing and phylogenetic analysis for SARS-CoV-2 virus are essential for understanding its evolution, transmission, and pathogenesis. The detection of single-nucleotide polymorphisms (SNPs) can provide valuable information on the genetic diversity of the virus and how it has spread across different regions.

2.4 MAIN GENE INVOLVED

As the study continues the biochemistry and interaction of SARS-CoV-2 virus with host cells explains the characteristics of "spike" protein on the host surface allowing binding to the ACE2 receptor on human cells, initiating its entry and replicate in the body[28]. The RNA genome is released by the virus as it penetrates human cells, where it multiplies and produces viral protein.

Angiotensin-converting enzyme 2 (ACE2) is expressed on a variety of body cell types, including those found in the kidney, intestines, heart, and lungs. Many researchers explain that ACE2 receptor may contribute to pathogenesis of COVID-19, as the virus infects and damages multiple organs. As ACE2 receptors control the renin-angiotensin-aldosterone system (RAAS), which is crucial to maintaining blood pressure and fluid balance, they may also play an important role in the immune response to COVID-19.

Further, another receptor notes were TMPRSS2 (Transmembrane Protease Serine 2) which also plays an important role in pathogenesis of Covid-19 as it is

described as a human cell surface protease. The SARS-CoV-2 virus has a "spike" protein which comes in contact with the host cell "Figure 2.3.1" and TMPRSS2 is an essential host protease that helps in viral entry [22]. With TMPRSS2, SARS-CoV-2 virus spike proteins can be cleaved (S1/S2 and S2 sites) and entered into human cells. TMPRSS2 is expressed in many cell types, including cells in lungs, prostate, and intestines. The expression of TMPRSS2 may contribute to the tropism of virus for these tissues & also contribute to the severity of respiratory disease caused by the virus [23].

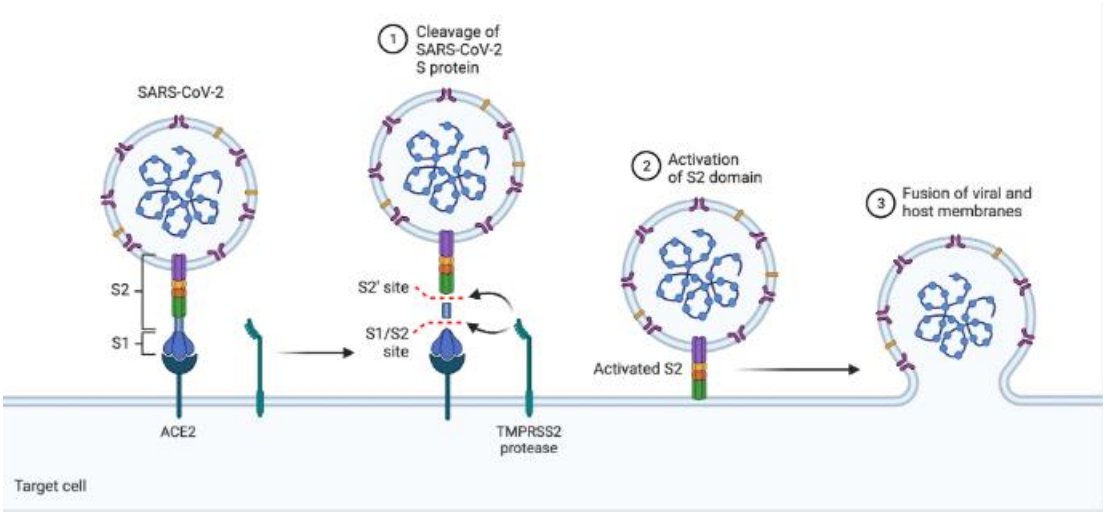


Figure 2.3.1: Mechanism of SARS-CoV-2 viral entry. "The TMPRSS2 protease cleaves the SARS-CoV-2 S protein at the S1/S2 and S2′ locations subsequent to its interaction with the host ACE2 receptor. Consequently, the S2 domain is activated and the viral and host membranes fuse" [24].

Apart from ACE-2 and TMRSS2 which has been discussed in various articles, there are other articles which highlights IFITM3, CD147 and IFIH1.

As the study continuous we see that there are reports from researches on [25] developed a bioinformatics pipeline for the detection of SNPs in viral genomes. They used this pipeline to analyze 1,776 SARS-CoV-2 genomes from the GISAID database to identify several SNPs that were associated with different

viral lineages. They also used these SNPs to construct a phylogenetic tree of the virus.

Another study from Hadfield et al. (2018) [7] used common approach to analyze 1,610 SARS-CoV-2 genomes taken into study via GISAID database. They identified several SNPs that were associated through different viral lineages & found evidence of multiple introductions of virus into Europe. There has been an increased interest in SNP detection and phylogenetic analysis due to the emergence of new SARS-CoV-2 strains like B.1.1.7 in the UK and B.1.351 in South Africa in recent times [20].

2.5 ASSOCIATED SNPs AND GENES

From the early studies of COVID-19 in China (2019), we found that the inhabitants of China have an alternative SNP of rs12252-C which is substantially associated with influenza infection. Further, by studying various databases & prominent papers we identified risk factors for rs12252-C, rs14393628, rs2285666, rs41303171, and rs35803318 are these are main SNPs involved in SARS-CoV-2.

It was also noted that rs12252-C which was involved in influenza infection may affect all populations with SARS-CoV-2 infection. There are noted to be other SNPs for IFITM3: rs12252-c and rs6598045. The most researched SNPs were found in ACE2 and IFITM3, followed by TMPRESS2 [26].

Overall, sequencing, and phylogenetic analysis have played a critical role in understanding the evolution and spread of SARS-CoV-2. The detection of SNPs could provide valuable information on the genetic diversity of Coronavirus and how it can spread across different regions. The studies also highlight the importance in global collaboration and data sharing in tracking the spread of infectious diseases [30][31].

CHAPTER 3

MATERIAL AND METHODS

OUTLINE

- The Genome sequence data for SARS-CoV-2 is isolated from India COVID cases which were downloaded & retrieved from the GISAID database.
- The study used 18 sequences which met quality assurance measures (length 29,910 nts and number of unknown bases 5-8%).
- The reference genome (Accession NC_ 045512.2) and the genome sequence of SARS-CoV-2, which was retrieved from the Indian COVID-19 index, were identified using the GISAID and GenBank databases.
- In MAFFT, multiple sequence alignment (MSA) was carried out (Version 7.471) along with alignment followed in DNA Star while SNP calling was implemented in DnaSP (Version 6.12.03, 64-bit environment), respectively and then visualised.
- MEGA X was used for the phylogenetic analysis.
- Haplotype were found from DnaSP.
- Further, the Quality of sequence, Conservation of seq. with respect to reference sequence and Consensus was conducted, identified & noted from Jalview Software and then viewed on SRplot software

3.1 MATERIAL USED:

3.1.1) Through GISAID databases (Global Initiative on Sharing All Influenza Data), whole genome sequences for Indian cases of SARS-CoV-2 were retrieved[25].

A global science initiative known as GISAID (Global Initiative on Sharing Avian Influenza Data) promotes the sharing of genomic information about influenza viruses, also the influenza virus that causes COVID-19.

GISAID was established in 2006 to enhance global preparedness and response to influenza pandemics by encouraging rapid and open sharing of data among scientists and public health officials worldwide.

GISAID has been crucial in facilitating the swift exchange of genetic sequencing data of SARS-CoV-2, the virus responsible for COVID-19, amongst researchers and public health agencies worldwide within the epidemic. This has allowed scientists to track the spread of the virus, understand its genetic mutations and evolution, and develop diagnostics, treatments, and vaccines.

GISAID provides a platform for researchers on share and access data on SARS-CoV-2 and other influenza viruses. The platform allows users to upload, download, and analyse genomic data, including sequences, metadata, and associated information. However, to access data, researchers must agree to GISAID's terms and conditions, which require data sharing and acknowledgment of the data providers. This ensures that data is shared responsibly and with appropriate credit given to the researchers who generated it.

Table 3.1.1.1: Genomic Sequences utilized for the study.

Country – INDIA		
State	**Variant**	**Accession ID**
Delhi	GRA	EPI_ISL_17048988
		EPI_ISL_16549308
		EPI_ISL_16549280
Gujarat	GRA	EPI_ISL_17190105
		EPI_ISL_17190102
		EPI_ISL_17190087
Kerala	GK	EPI_ISL_16608014
		EPI_ISL_16639467
		EPI_ISL_16639466
Karnataka	GRA	EPI_ISL_16818447
		EPI_ISL_16818435

		EPI_ISL_15973546
Odisha	GRA	EPI_ISL_16487018
		EPI_ISL_16487030
		EPI_ISL_16504207
Himachal Pradesh	GRA	EPI_ISL_17096480
		EPI_ISL_16395186
		EPI_ISL_17257455

3.1.2) The National Library of Medicine (NLM) is home to the National Centre for Biotechnology Information (NCBI), a branch of the National Institutes of Health (NIH). It was founded in 1988 with the goal of creating molecular biology information systems. It offers access to genomic and biomedical information resources, including PubMed, BLAST, and GenBank.

NCBI maintains and updates several databases, including GenBank, which is a comprehensive public database of nucleotide sequences for over 400,000 organisms. It also maintains databases for protein sequences, gene expression, genetic variation, and clinical studies.

In addition, NCBI provides various tools and software for data analysis, such as BLAST, a widely used tool for sequence align Workbench, a platform for visualisation and analysis of genomic data. NCBI also offers educational resources and training programs for researchers, educators, and the public.

3.1.3) GenBank is an NIH genetic sequence database, an annotated collection of all DNA sequences publicly available. [32] The data in GenBank is annotated and curated to provide accurate information about the sequence, its origin, and its biological function. Researchers can use GenBank to search for genetic information related to specific genes, organisms, or diseases, and use the information to study various aspects of genetics, genomics, and evolution. The data in GenBank is freely accessible to the public and is an important resource for biomedical research, genetics, and biodiversity conservation.

3.1.4) MAFFT The software programme known as "Multiple Alignment using Fast Fourier Transform" is widely used for multiple sequence alignment (MSA) of amino acid and nucleotide sequences. It uses fast and accurate algorithm to align multiple sequences based on their pairwise distances, with a choice of several different methods for estimating these distances. MAFFT is widely used in bioinformatics research and is available as a standalone program as well as in the form of a web server. It can handle large datasets with thousands of sequences [31] and has several options for customising the alignment output, such as specifying the gap penalties and alignment scoring matrices. MAFFT also supports the alignment of sequences with different lengths and has options for dealing with gaps and missing data.

Figure 3.1.4.1: Alignment results via MAFFT.

Figure 3.1.4.2: Alignment Sequence Trimmed at 5' End and 3' End to Obtain True Homology

3.1.5) DNAStar software is a suite of bioinformatics software tools designed for DNA and protein sequence analysis, alignment, assembly, and visualisation. It includes various modules such as SeqMan Pro, Lasergene Genomics Suite, Protean, and ArrayStar. These modules provide tools for any DNA and RNA sequence alignment, annotation, & analysis, primer design, gene expression analysis, as well as next-generation sequencing data analysis. The software can also be used for a variety of applications such as molecular biology, genetics,

genomics, and proteomics research, and are commonly used in academic, industrial, and government research laboratories. DNAStar known for its user-friendly interface, quality algorithms, and comprehensive support.

3.1.6) DNASP, or DNA Sequence Polymorphism, is a software program that analyses DNA polymorphisms based on data from an individual locus (MSA data) or a number of loci (Multiple-MSA data generated by some assembler). This software can estimate several DNA sequence variation parameters (such as linkage disequilibrium, recombination, gene flow, and gene conversion) within and between populations.[33] Furthermore, DnaSP can perform neutrality tests and compute their confidence intervals by coalescence. Analyses of results are displayed in tabular and graphic form.

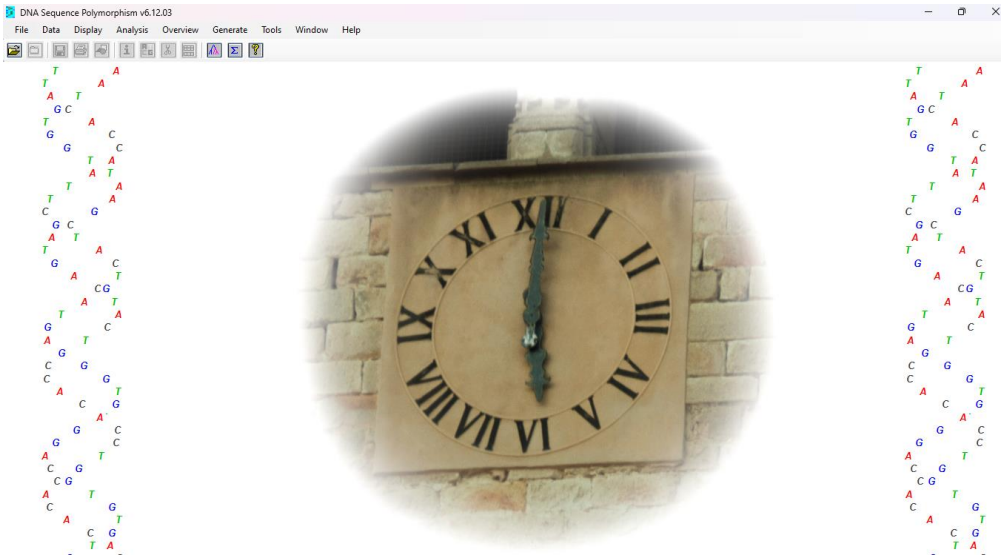


Figure 3.1.6.1: DNAsp to find polymorphisms.

3.1.7) SRplot: SRplot is a web-based application tailored for creating numerous scientific visualizations, such as heatmaps. It is extensively utilized in the fields of genetics and bioinformatics to display single nucleotide polymorphisms (SNPs) and their patterns across various samples or experimental conditions.

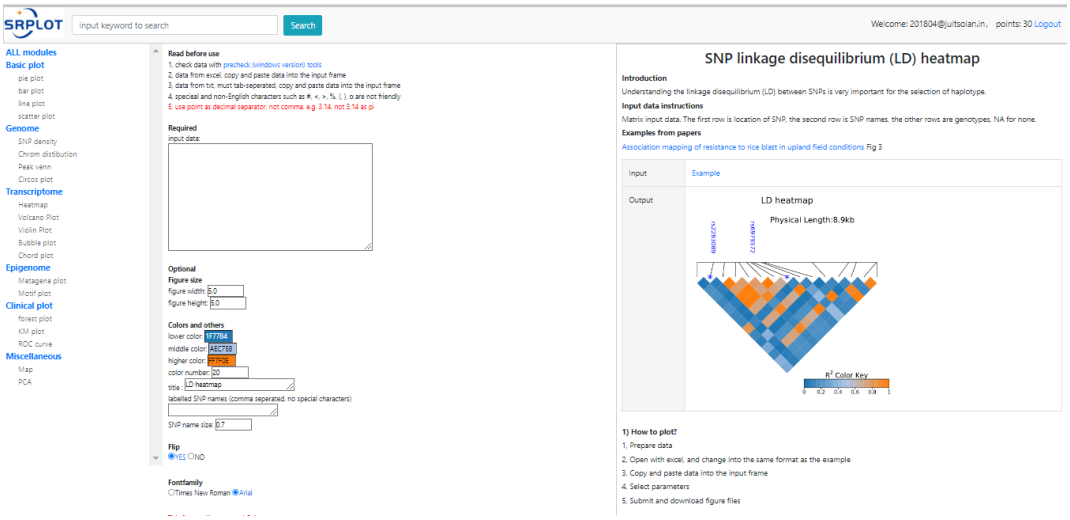


Figure 3.1.7.1 SRplot webpage

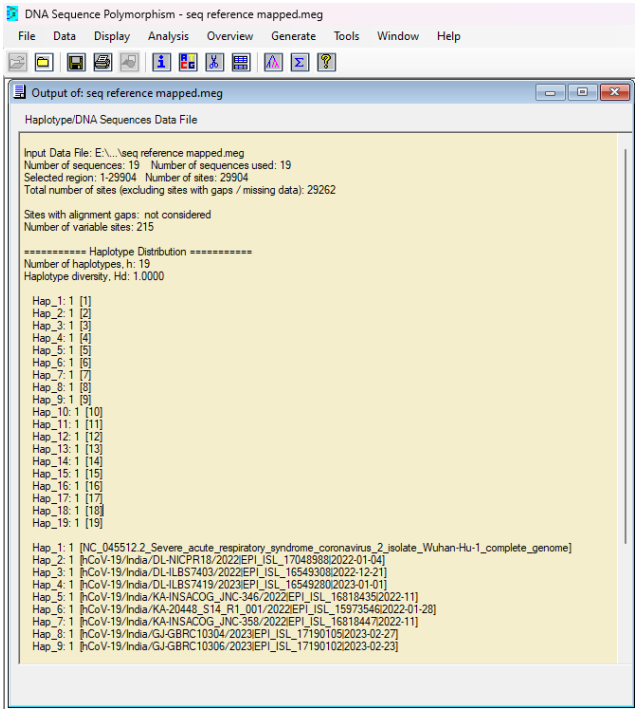
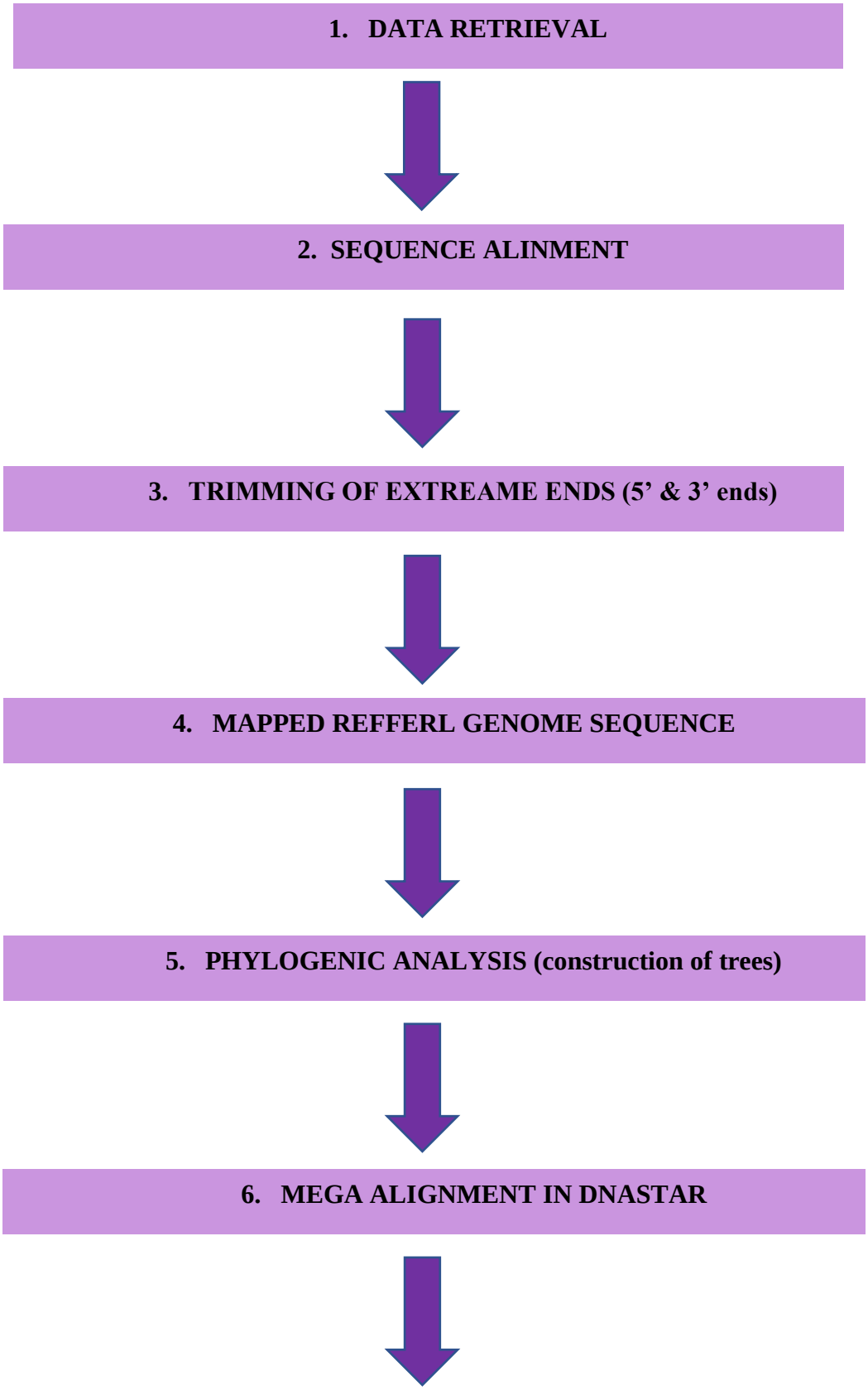


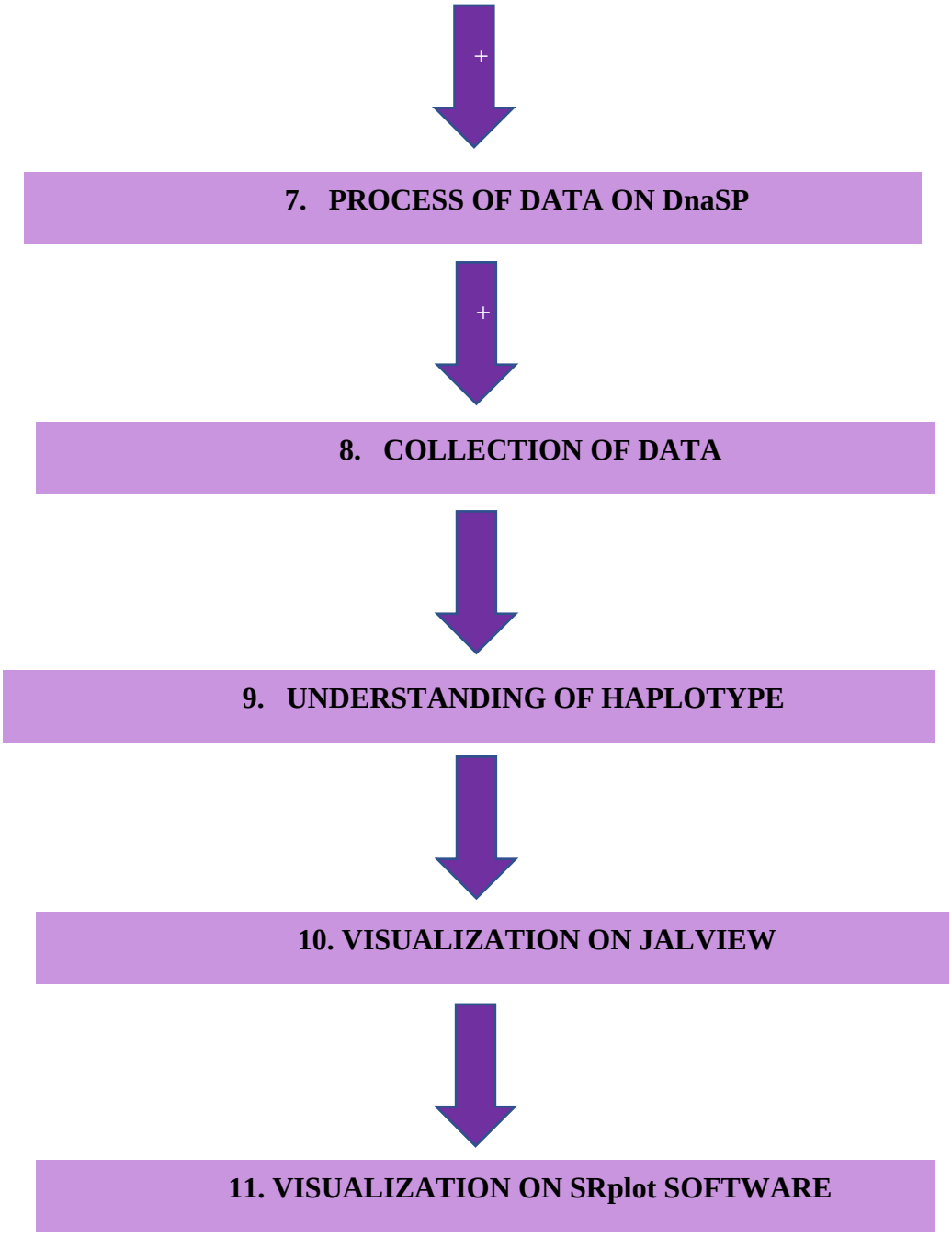
Figure 3.1.7.2: Data set for haplotype findings

3.1.8) Pangolin is a software application developed for the purpose of categorizing SARS-CoV-2 sequences into specific lineages. It was created by the outbreak.info team at the University of California, San Francisco (UCSF), along with collaborators from other institutions. Pangolin's primary function is to assign SARS-CoV-2 sequences to distinct lineages or clades based on their genetic characteristics. These lineages are identified using a combination of phylogenetic

analysis and statistical methods. Pangolin employs a standardized lineage naming system, assigning each lineage a unique alphanumeric identifier. This system facilitates efficient tracking and communication among researchers and public health officials. The software allows users to upload sequence data in various formats and automates the process of lineage assignment. Pangolin allows for the monitoring of recently identified variations of interest and concern by facilitating real-time monitoring of SARS-CoV-2 lineages. It also offers visualization tools for analyzing lineage distribution over time and across geographic regions. Pangolin integrates with public databases of SARS-CoV-2 sequences, providing access to extensive genomic datasets for lineage analysis. Overall, Pangolin plays a crucial role in genomic surveillance efforts, contributing valuable insights into the evolution and transmission dynamics of the virus. Its ability to classify lineages is crucial for guiding public health measures in response to the COVID-19 pandemic.

3.2 METHOD:


1. DATA RETRIEVAL
2. SEQUENCE ALINMENT
3. TRIMMING OF EXTREAME ENDS (5' & 3' ends)
4. MAPPED REFFERL GENOME SEQUENCE
5. PHYLOGENIC ANALYSIS (construction of trees)
6. MEGA ALIGNMENT IN DNASTAR


7. PROCESS OF DATA ON DnaSP
8. COLLECTION OF DATA
9. UNDERSTANDING OF HAPLOTYPE
10. VISUALIZATION ON JALVIEW
11. VISUALIZATION ON SRplot SOFTWARE

CHAPTER 4

RESULTS

4.1 AN OVERVIEW OF THE OBTAINED SEQUENCES

SARS-CoV-2 was isolated from 30 individuals for the study and 18 sequences were finalized from Indian Covid cases from different places: Himachal Pradesh, Delhi, Kerala, Gujarat, Karnataka & Odisha.

4.2 COVID-19 CASES, DEATHS, AND TESTS AS OF 2024 STATISTICS.

Table 4.2.1: 2024 COVID-19 Statistics

Coronavirus Cases	704,753,890
Deaths	7,010,681
Recovered	675,619,811

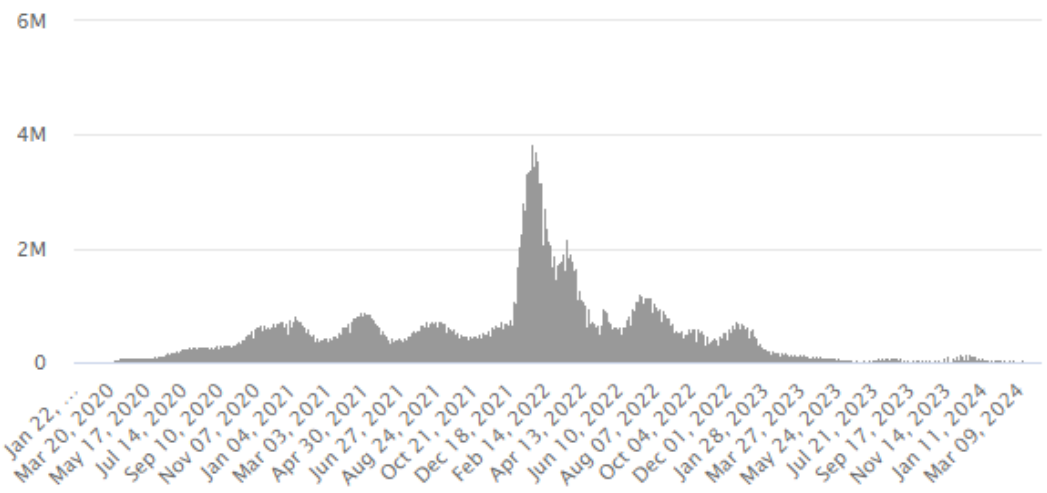


Figure 4.2.1: Daily new cases of COVID-19 from 2019-2024

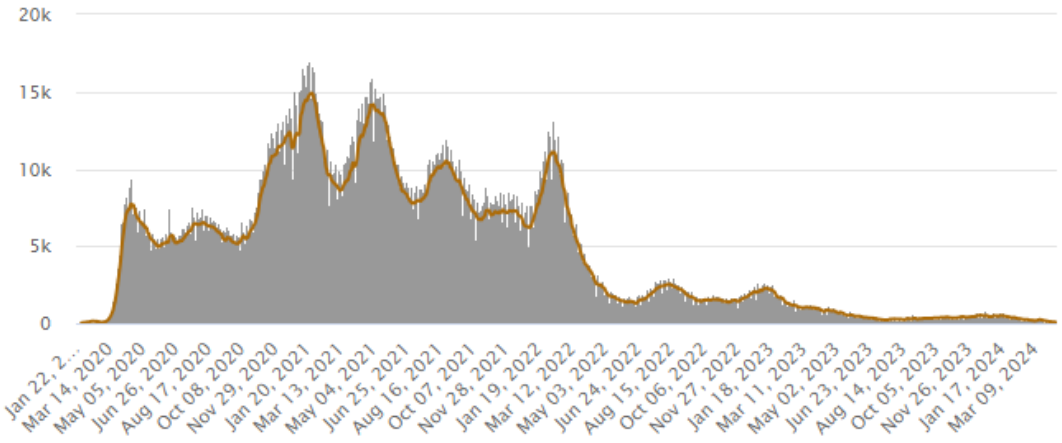


Figure 4.2.2: Death rate from 2019-2024 for COVID-19 cases.

Table 4.2.2: Reported COVID-19 cases country-wise.

Countries	Total Cases	Total Deaths	Total Recovered
Africa	12,860,924	258,892	12,090,808
South America	70,200,879	1,367,332	66,683,585
Asia	221,500,265	1,553,662	205,673,091
North America	131,889,132	1,695,941	127,665,129
Europe	253,406,198	2,101,824	248,754,104
Oceania	14,895,771	33,015	14,752,388

Source: https://www.worldometers.info/coronavirus/#countries

4.3 STATISTICS OF THE RETRIEVED SEQUENCES

Table 4.3.1 displays the summary statistics of 18 complete SARS-CoV-2 sequences that met our criteria. These sequences were sourced from 18 individuals, yielding complete genome sequences with an average size of 29,831, with a standard deviation of 34.51.

Table 4.3.1: The Comprehensive Statistics of All 18 SARS-Cov-2

Statistical parameter	Values
maximum size (nts)	29904
minimum size (nts)	29736
Mean (nts)	29831
SD	34.512994862869
Coeff of variation (%)	0.11886%

4.3.1 HAPLOTYPE ANALYSIS

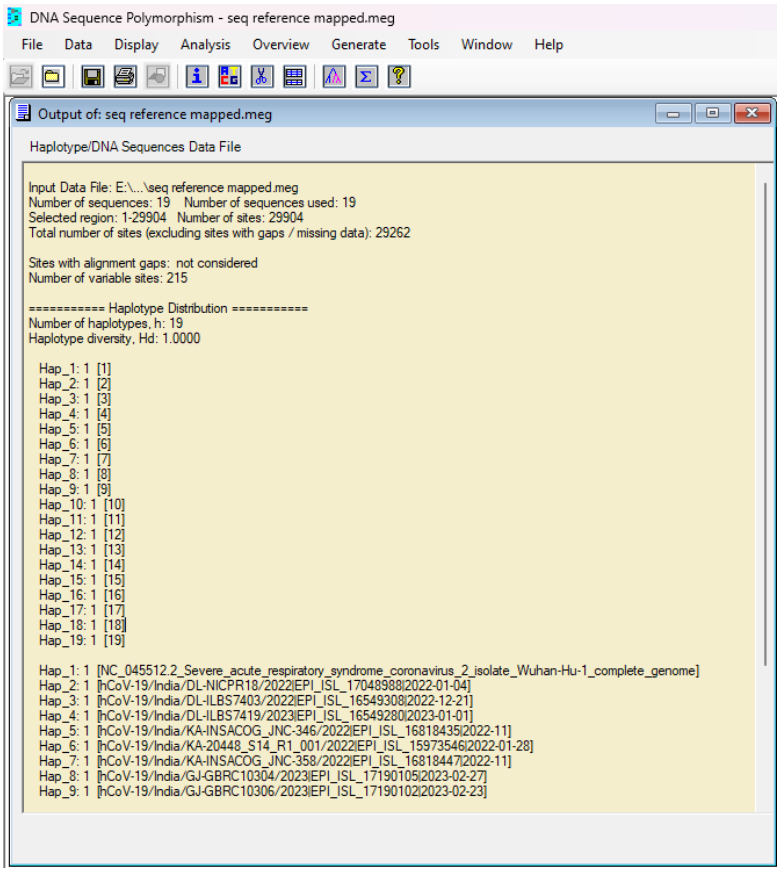


Figure 4.3.1.1 Haplotype Data File

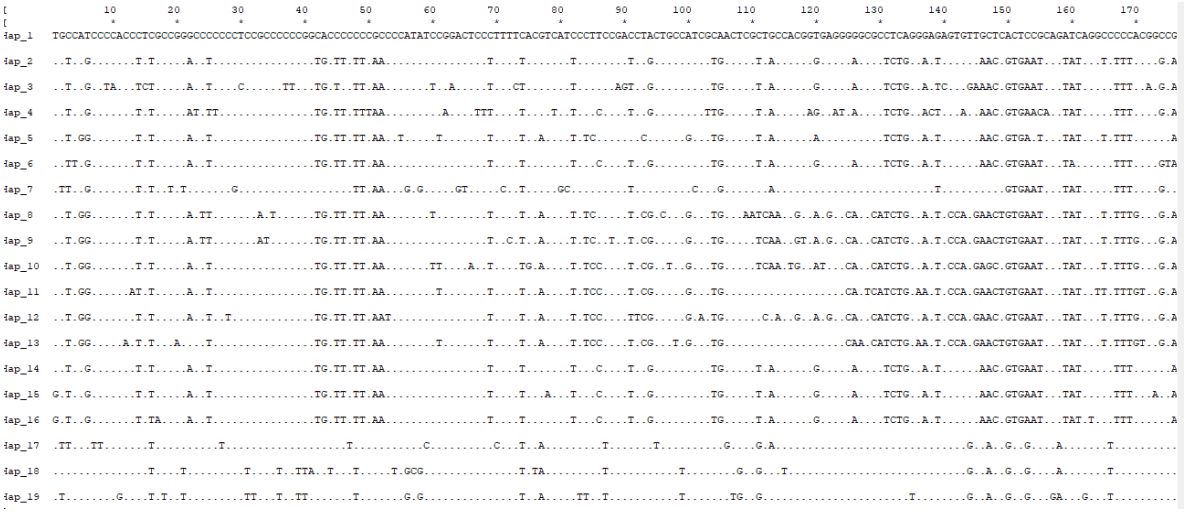


Figure 4.3.1.2: Haplotype Data File Result Via Notepad

The following Figure 4.3.1.1 & Figure 4.3.1.2 explains that there is the presence of haplotype in all sequences of the test samples.

4.3.2 DNAsp RESULTS

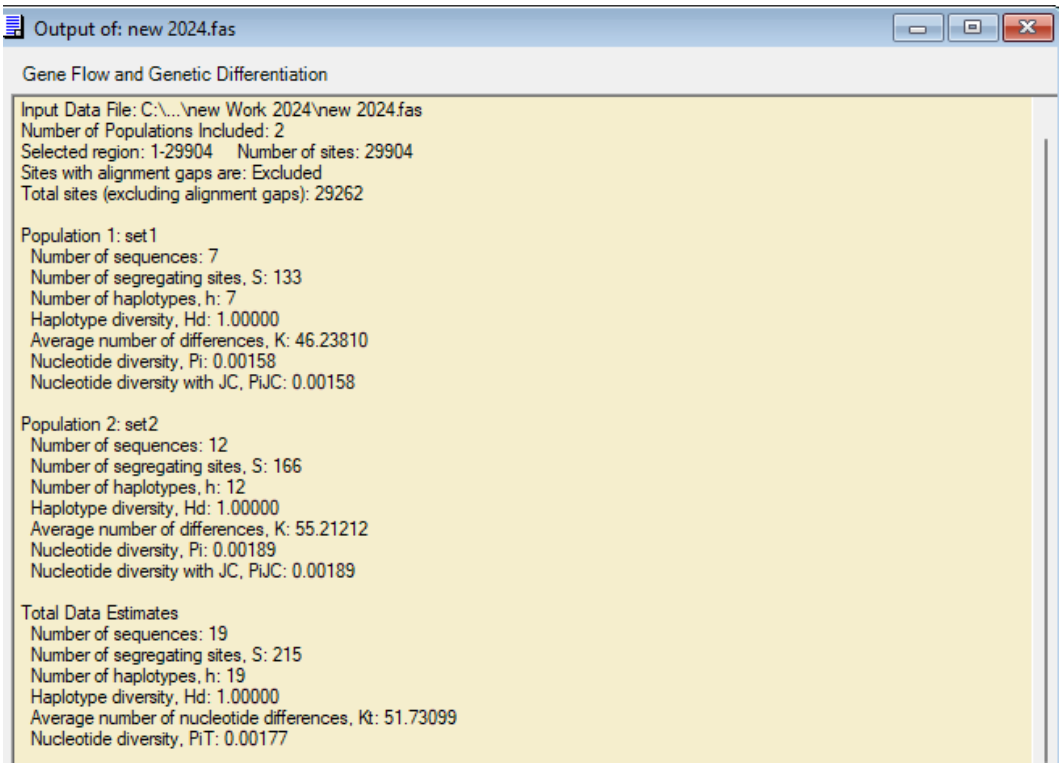


Figure 4.3.2.1 Gene Flow & Genetic Differentiation

Population 1 comprises of set 1 (Reference + Sample from Delhi + Sample from Himachal)

Population 2 comprises of set 2 (Samples from Kerala + Karnataka + Gujarat + Odisha)

4.4 SNP ANALYSIS

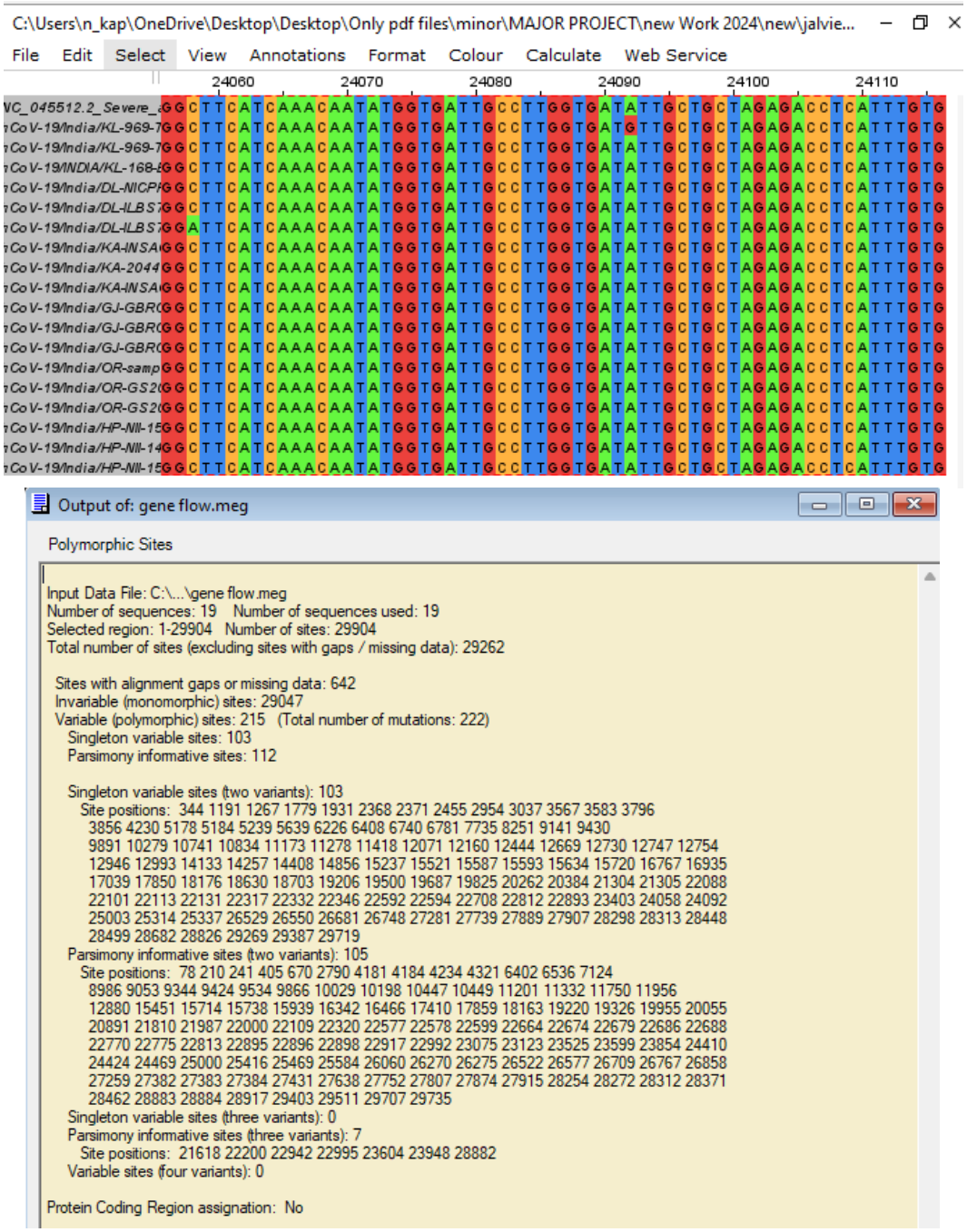


Figure 4.4.1: SNPs sites

In all, 215 SNPs were found; 103 of these were monoallelic, 105 were diallelic, and 7 were triallelic, as shown in "Figure 4.4.1."

We used SRplot to generate the heatmap and the input data was structured into matrix with rows and columns, Figure 4.4.2. An input data file was created keeping in view:

- First-Row: SNP Locations - This row should include the genomic locations of the SNPs, indicating their specific positions on the chromosome.
- First-Column: Name and ID for the sample data set.
- Following-Row: The second row onwards should contain the names or unique samples of the SNPs. These are essential for referencing and differentiating between the SNPs.
- "NA" should be mentioned for missing data.

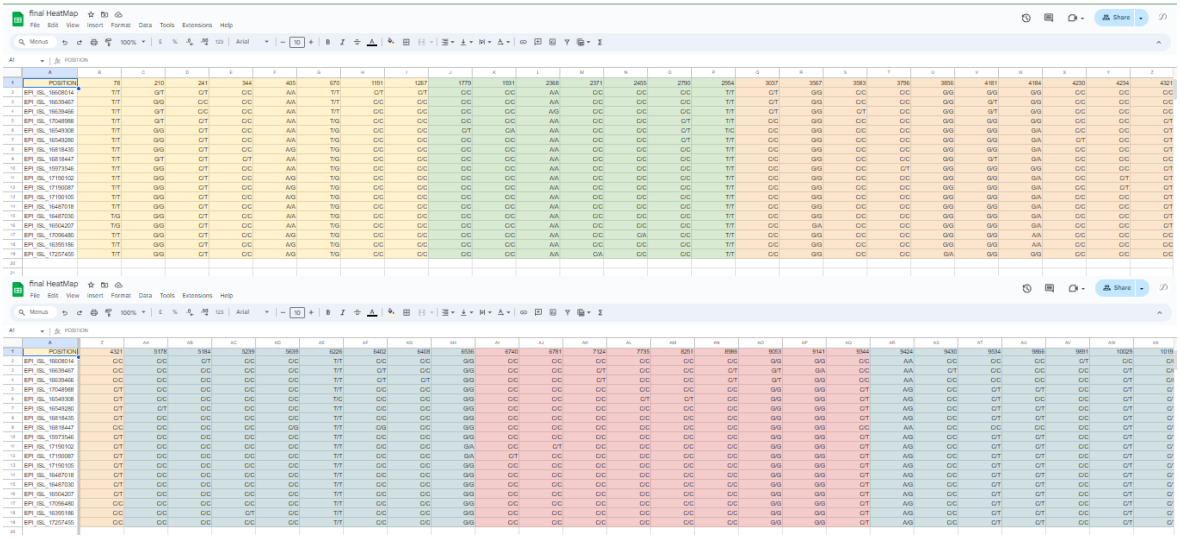


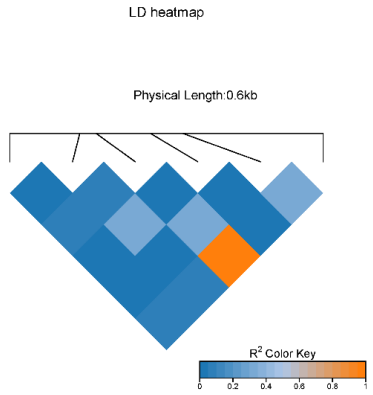
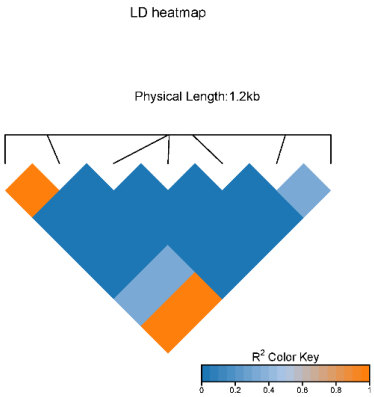
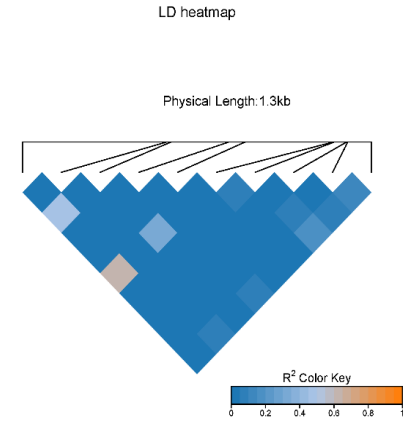
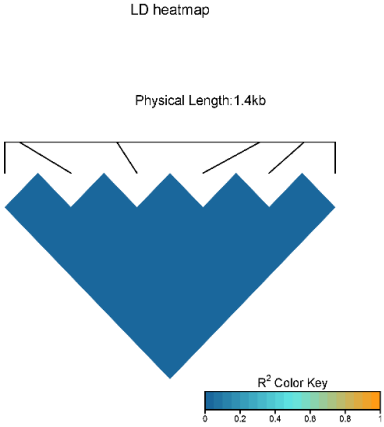
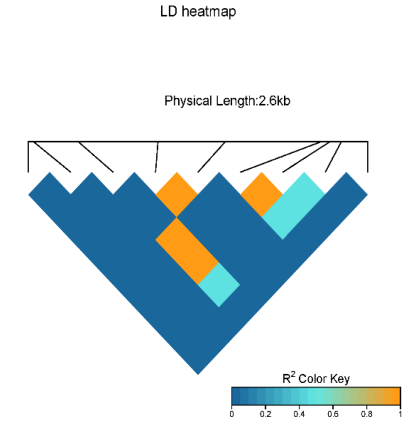
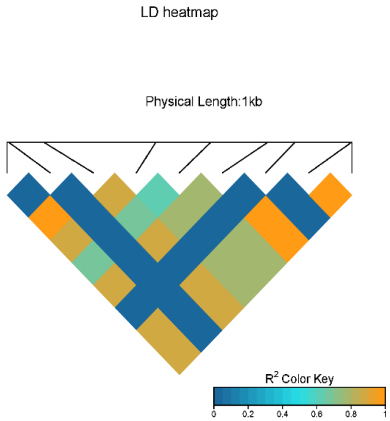
Figure 4.4.2 Data file with SNPs of the sample

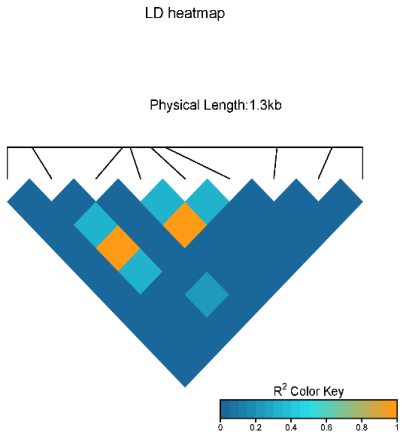
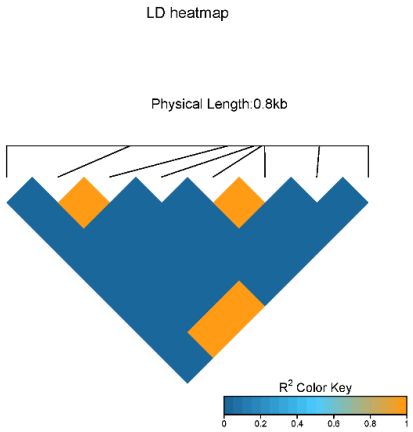
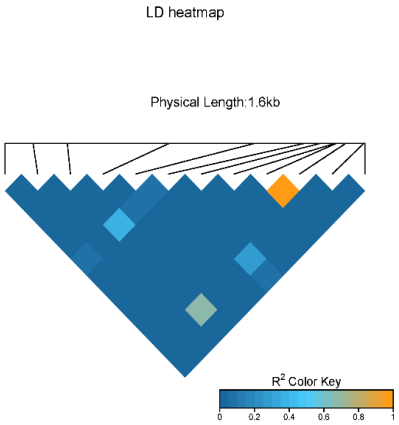
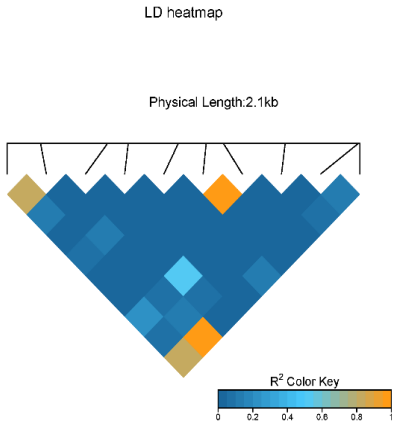
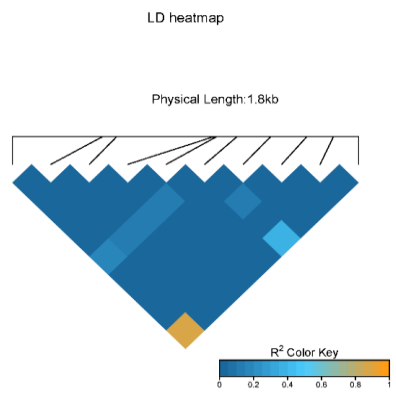
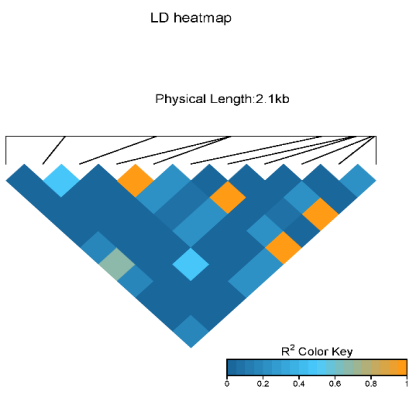
After formatting the matrix data, SRplot processes the information to create a heatmap. This visual tool enables researchers to quickly identify patterns and variations in SNP genotypes across multiple samples.

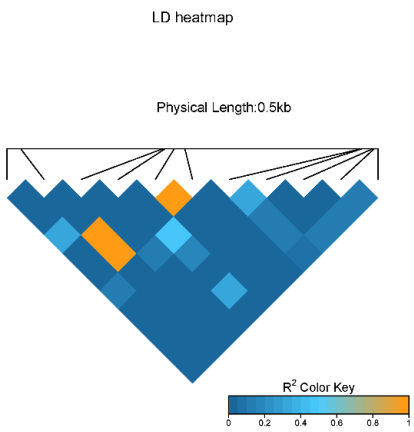
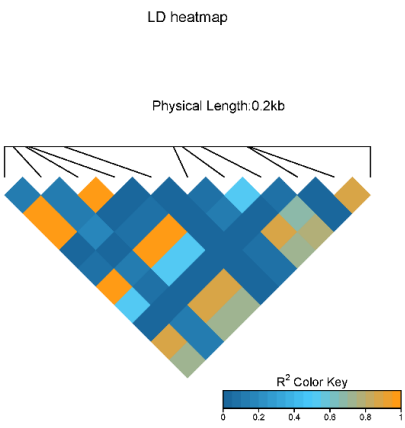
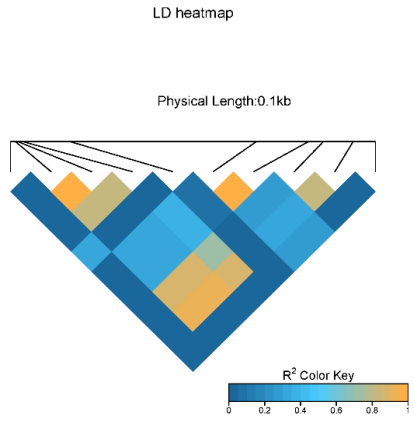
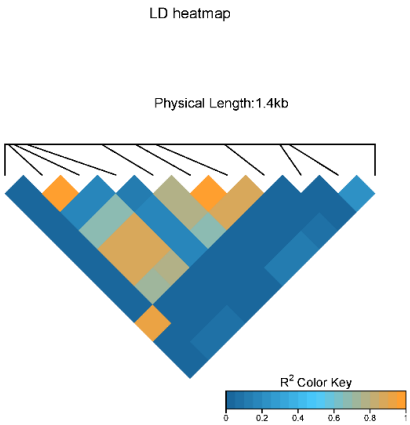
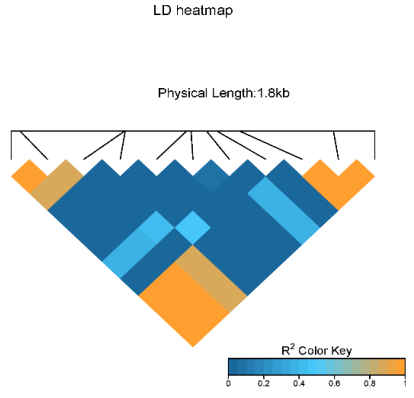
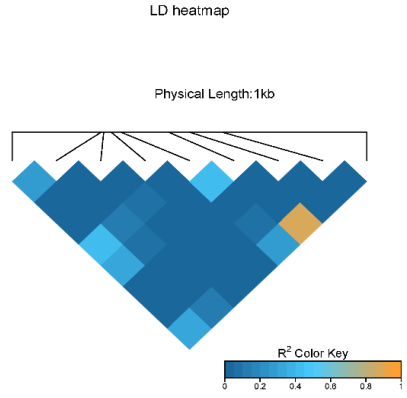
The analysed results can be summarized in the Table 4.4.1, which is a comprehensive overview of the SNPs at different positions, detailing:

- The precise locations and identifiers of the SNPs along with the variation according to the colour scheme.
- The distribution of genotypes across the samples.
- Occurrences of missing genotype data, indicated by "NA".
- $\mathbf{R^2}$ value indicates the relationship between alleles.

Table 4.4.1 LD Heatmap Construction

 SNPs between 78-1267 Position	 SNPs between 1779-2954 Position
 SNPs between 3037-4321 Position	 SNPs between 5178-6536 Position
 SNPs between 6740-9344 Position	 SNPs between 9424-10449 Position

 SNPs between 10741-12071 Position	 SNPs between 12160-12993 Position
 SNPs between 14133-15720 Position	 SNPs between 15738-17859 Position
 SNPs between 18163-19955 Position	 SNPs between 20055-22110 Position

 SNPs between 22113-22599 Position	 SNPs between 22674-22882 Position
 SNPs between 22893-23040 Position	 SNPs between 23042-24410 Position
 SNPs between 24424-26270 Position	 SNPs between 26275-27259 Position

4.5 PHYLOGENETIC ANALYSIS

The MEGA X software was utilized to construct phylogenetic trees from data aligned by MAFFT, employing the Tamura Nei evolutionary model assuming constant nucleotide replacement. The Neighbor Joining (NJ) and BioNJ algorithms were employed for heuristic search, selecting the tree with the highest log likelihood value as the starting point. Topology and clustering pattern analysis were conducted, supported by bootstrap values derived from 1000 replicates.

4.5.1) The evolutionary history was inferred using the Neighbor-Joining technique [36]. The evolutionary connections of the species under investigation were represented by a bootstrap consensus tree constructed from 1000 replicates [35]. Branches that had less than 50% support from bootstrap collapsed. Next to each branch is the proportion of replicates in which the taxa grouped together in the bootstrap test (1000 repetitions) [35]. Evolutionary distances are reported as the number of base substitutions per site and were computed using the Maximum Composite Likelihood approach [36]. With regard to codon locations 1st+2nd+3rd+Noncoding, 19 nucleotide sequences were considered in this research. Pairwise deletion was used to eliminate ambiguous places. The final set of places in the dataset was 29904. With MEGA11, all evolutionary analyses were carried out [30].

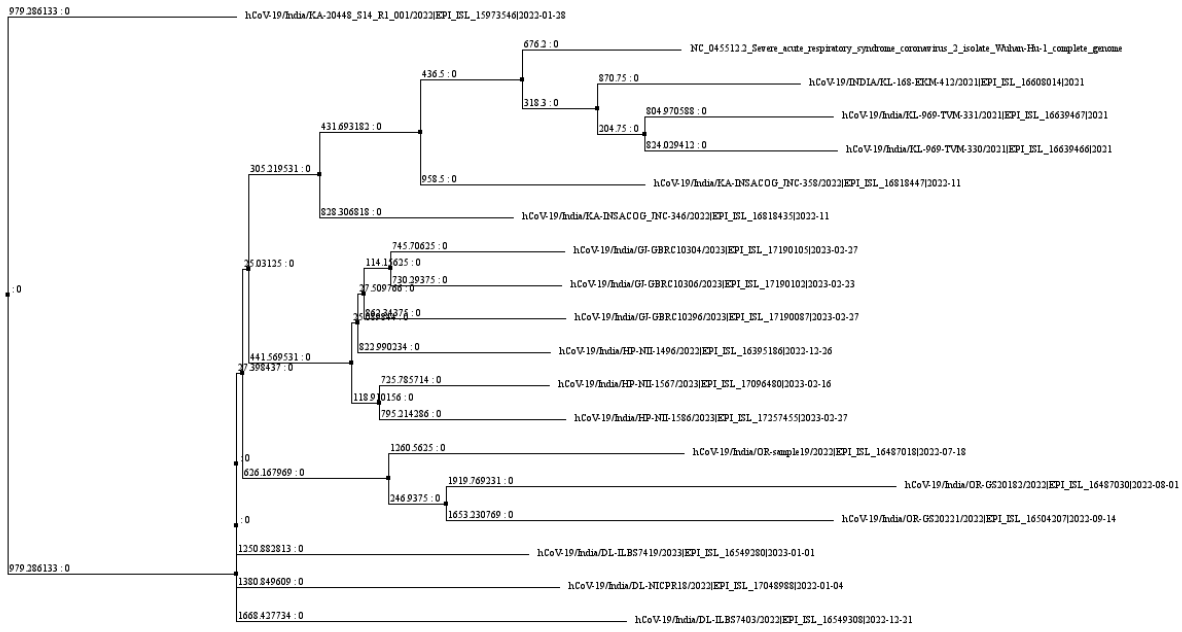


Figure 4.5.1.1: Neighbor Joining Tree.

4.5.2) The Tamura-Nei model and the Maximum Likelihood approach were used to infer evolutionary history [29]. Presented is the tree with the greatest log-likelihood (-43094.63). Alongside branches is an indication of the percentage of trees where related species gathered together [35]. Using Neighbor-Join and BioNJ algorithms applied to pairwise distance matrices calculated using the Tamura-Nei model, the initial trees for heuristic search were automatically created, choosing the topology with the best log likelihood value. 19 nucleotide sequences were used in this study, with codon positions 1st+2nd+3rd+Noncoding taken into account. The final set of places in the dataset was 29904. With MEGA11, all evolutionary analyses were carried out[30].

Figure 4.5.2.1: Maximum-Likelihood tree

4.5.3) Clustering- analysis-of the maximum-likelihood -phylogenetic tree gave 4 results which are said to be similar with 99% accuracy.

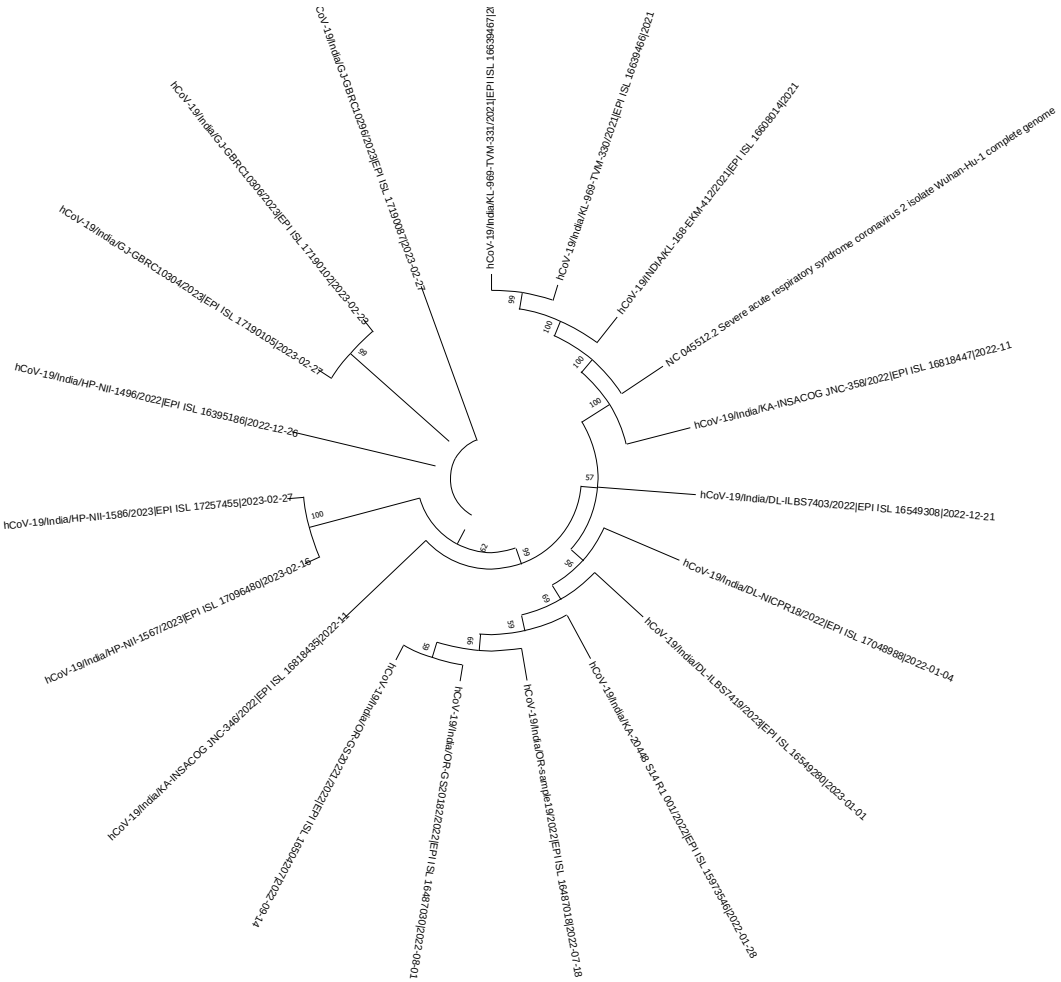


Figure 4.5.3.1: Bootstrap Consensus Tree for Maximum Likelihood

CHAPTER 5

CONCLUSION & FUTURE SCOPE

5.1 CONCLUSION

The current research investigation looked at the phylogenetic reconstruction and characterization of the whole genome sequence of SARS-CoV-2 isolates from the Indian population, specifically from the states of Gujarat, Himachal Pradesh, Delhi, Odisha, Kerala, and Karnataka.

Only 18 of the hundreds of complete genome sequences 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 utilize 18 sequences of sufficient accuracy instead of several hundred sequences of low or doubtful integrity.

The maximum sequence set taken for the research showed similarity in the genome sequence although research sets were from different locations throughout India. The likelihood of SNPs analysis were around 100+ snps which were found from 29900 nts of the reference sequence of COVID-19 along with different nts lengths for other test sets.

Our research set took data from 2023 which was after the extreme pandemic had started to settle down and subside. Despite the pandemic receding, the detection of SNPs remained crucial. SNPs, variations in a single nucleotide at specific genomic positions, play pivotal roles in understanding genetic predispositions to diseases, including infectious ones like COVID-19.

With the results released throughout this study give us the idea that there is a resemblance of each set to the reference genome sequence, it will also help to consistently research on a larger scale worldwide.

After considering all of the research, we have concluded that the latest strain of virus is relatively new, having originated between October and December of

2019. There were 215 SNPs found in all, 103 of which were monoallelic, 105 of which were diallelic, and 7 of which were triallelic.

Along with this, all the tree formations have revealed that 4 major clades were recognized by the phylogenetic tree.

5.2 FUTURE SCOPE

Even as the pandemic wanes, examining SNPs could provide valuable insights into how individuals react to the virus, potentially guiding future public health approaches and treatment methods. It is crucial to highlight that the diverse applications of SNPs would not be possible without technological advancements enabling the discovery, prediction, and validation of SNPs. Undoubtedly, advancements in bioinformatics are indispensable for effectively studying SNPs.

My research with the Indian population illuminates the identification of single nucleotide polymorphisms (SNPs) across various species. The frequency of occurrences, cost-effectiveness of test development, and adaptability of such assays across different research facilities supported the utilization of SNPs for investigating genetic variations within specific populations, including humans, plants, or microorganisms & holds significant promise across multiple domains.

Further, this study helps in advancing personalized medicine and public health measures to uncover population genetics and bolster forensic practices, this research stands to profoundly impact healthcare, genetic exploration, and societal progress in India and beyond. By persisting in the exploration and application of findings from this study, we can foster a deeper comprehension of human genetic diversity, ensure fair access to healthcare, and lay the groundwork for targeted, efficient interventions tailored to India's genetic makeup.

REFERENCES

1) Peiris, J. (2012). Coronaviruses. *Medical Microbiology*, 587–593. https://doi.org/10.1016/b978-0-7020-4089-4.00072-x

2) C. Drosten *et al.*, "Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1967–1976, May 2003, doi: 10.1056/nejmoa030747.

3) S. M. Peiris and L. L. M. Poon, "Severe Acute Respiratory Syndrome (SARS)," *Encyclopedia of Virology*, pp. 552–560, 2008, doi: 10.1016/b978-012374410-4.00780-9.

4) R. Fauver *et al.*, "Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States," *Cell*, vol. 181, no. 5, pp. 990-996.e5, May 2020, doi: 10.1016/j.cell.2020.04.021.

5) M. F. Osuchowski *et al.*, "The COVID-19 puzzle: deciphering pathophysiology and phenotypes of a new disease entity," *The Lancet Respiratory Medicine*, vol. 9, no. 6, pp. 622–642, Jun. 2021, doi: 10.1016/s2213-2600(21)00218-6.

6) L. Singh *et al.*, "Modulation of Host Immune Response Is an Alternative Strategy to Combat SARS-CoV-2 Pathogenesis," *Frontiers in Immunology*, vol. 12, Jul. 2021, doi: 10.3389/fimmu.2021.660632.

7) Hadfield *et al.*, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, May 2018, doi: 10.1093/bioinformatics/bty407.

8) "Chris Gunter, Ph.D.," *Genome.gov*. [Online]. Available: https://www.genome.gov/staff/Chris-Gunter-PhD

9) Zhao *et al.*, "Relationship Between the ABO Blood Group and the Coronavirus Disease 2019 (COVID-19) Susceptibility," *Clinical Infectious Diseases*, vol. 73, no. 2, pp. 328–331, Aug. 2020, doi: 10.1093/cid/ciaa1150.

10) D. Schoeman, B. Gordon, and B. C. Fielding, "Coronaviruses," *Encyclopedia of Infection and Immunity*, pp. 241–258, 2022, doi: 10.1016/b978-0-12-818731-9.00052-5.

11) T. Estola, "Coronaviruses, a New Group of Animal RNA Viruses," *Avian Diseases*, vol. 14, no. 2, p. 330, May 1970, doi: 10.2307/1588476.

12) A. Barthorpe and J. P. Rogers, "Coronavirus infections from 2002 to 2021: neuropsychiatric manifestations," *Sleep Medicine*, vol. 91, pp. 282–288, Mar. 2022, doi: 10.1016/j.sleep.2021.11.013.

13) WHO (2003a) "WHO issues a global alert about cases of atypical pneumonia," Mar. 12, 2003. [Online]. Available: https://www.who.int/news/item/12-03-2003-who-issues-a-global-alert-about-cases-of-atypical-pneumonia

14) WHO (2003b) WHO Issues Global Alert About Cases of Atypical Pneumonia: Cases of Severe Respiratory Illness May Spread to Hospital Staff. Accessed on 15 October 2007 at http://www.who.int/csr/sars/archive/

15) World Health Organization (2003c). Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. Retrieved from https://www.who.int/csr/sars/country/table2004_04_21/en/

16) N. Lee *et al.*, "A Major Outbreak of Severe Acute Respiratory Syndrome in Hong Kong," *New England Journal of Medicine*, vol. 348, no. 20, pp. 1986–1994, May 2003, doi: 10.1056/nejmoa030685.

17) D. Schoeman, B. Gordon, and B. C. Fielding, "Coronaviruses," *Encyclopedia of Infection and Immunity*, pp. 241–258, 2022, doi: 10.1016/b978-0-12-818731-9.00052-5.

18) L. Wang, S. Su, Y. Bi, G. Wong, and G. F. Gao, "Bat-Origin Coronaviruses Expand Their Host Range to Pigs," *Trends in Microbiology*, vol. 26, no. 6, pp. 466–470, Jun. 2018, doi: 10.1016/j.tim.2018.03.001.

19) K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of SARS-CoV-2," *Nature Medicine*, vol. 26, no. 4, pp. 450–452, Mar. 2020, doi: 10.1038/s41591-020-0820-9.

20) C. J. Burrell, C. R. Howard, and F. A. Murphy, Chapter 31- Coronaviruses, Book *Fenner and White's Medical Virology*. Academic Press, 2016, 437-446 DOI: 10.1016/B978-0-12-375156-0.00031-X

21) C. Singh *et al.*, "Effectiveness of COVID-19 vaccine in preventing infection and disease severity: a case-control study from an Eastern State of India," *Epidemiology and Infection*, vol. 149, 2021, doi: 10.1017/s0950268821002247.

22) M. Hoffmann *et al.*, "SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor," *Cell*, vol. 181, no. 2, pp. 271-280.e8, Apr. 2020, doi: 10.1016/j.cell.2020.02.052.

23) S. Matsuyama, N. Nagata, K. Shirato, M. Kawase, M. Takeda, and F. Taguchi, "Efficient Activation of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein by the Transmembrane Protease TMPRSS2," *Journal of Virology*, vol. 84, no. 24, pp. 12658–12664, Dec. 2010, doi: 10.1128/jvi.01542-10.

24) E. Hartenian, D. Nandakumar, A. Lari, M. Ly, J. M. Tucker, and B. A. Glaunsinger, "The molecular virology of coronaviruses," *Journal of Biological Chemistry*, vol. 295, no. 37, pp. 12910–12934, Sep. 2020, doi: 10.1074/jbc.rev120.013930.

25) Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, vol. 22, no. 13, Mar. 2017, doi: 10.2807/1560-7917.es.2017.22.13.30494.

26) "SARS-CoV-2 Genomes from Nigeria Reveal Community Transmission, Multiple Virus Lineages and Spike Protein Mutation Associated with Higher Transmission and Pathogenicity," *Virological*, May 29, 2020.

27) BioRender. "An In-depth Look into the Structure of the SARS-CoV2 Spike Glycoprotein." [Online]. Available: https://app.biorender.com/biorender-templates/figures/5e99f5395fd61e0028682c01/t-5f1754e62baea000aee86904-an-in-depth-look-into-the-structure-of-the-sars-cov2-spike-g (accessed on 1 August 2020)

28) A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, “Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein,” *Cell*, vol. 183, no. 6, p. 1735, Dec. 2020, doi: 10.1016/j.cell.2020.11.032.

29) “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.,” *Molecular Biology and Evolution*, May 1993, doi: 10.1093/oxfordjournals.molbev.a040023.

30) K. Tamura, G. Stecher, and S. Kumar, “MEGA11: Molecular Evolutionary Genetics Analysis Version 11,” *Molecular biology and evolution*, Apr. 23, 2021.

31) J. Felsenstein, “CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP,” *Evolution*, vol. 39, no. 4, pp. 783–791, Jul. 1985, doi: 10.1111/j.1558-5646.1985.tb00420.x.

32) L. M. Van Blerkom, “Role of viruses in human evolution,” *American Journal of Physical Anthropology*, vol. 122, no. S37, pp. 14–46, 2003, doi: 10.1002/ajpa.10384.

33) C. Wang *et al.*, “The establishment of reference sequence for SARS-CoV-2 and variation analysis,” *Journal of Medical Virology*, vol. 92, no. 6, pp. 667–674, Mar. 2020, doi: 10.1002/jmv.25762.

34) J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, “Haploview: analysis and visualization of LD and haplotype maps,” *Bioinformatics*, Aug. 05, 2004. [PubMed ID: 15297300]

35) Saitou N. and Nei M. “The neighbor-joining method: a new method for reconstructing phylogenetic trees.,” *Molecular Biology and Evolution*, Jul. 1987, doi: 10.1093/oxfordjournals.molbev.a040454.

36) K. Tamura, M. Nei, and S. Kumar, “Prospects for inferring very large phylogenies by using the neighbor-joining method,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 30, pp. 11030–11035, Jul. 2004, doi: 10.1073/pnas.0404206101.

Thesis

ORIGINALITY REPORT

13%	10%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Christopher J. Burrell, Colin R. Howard, Frederick A. Murphy. "Coronaviruses", Fenner and White's Medical Virology Internet Source	2%
2	I.A. Taiwo, N. Adeleye, F.O. Anwoju, A. Adeyinka, I.C. Uzoma, T.T. Bankole. "Sequence analysis for SNP detection and phylogenetic reconstruction of SARS-cov-2 isolated from Nigerian COVID-19 cases", New Microbes and New Infections, 2022 Publication	2%
3	www.mdpi.com Internet Source	1%
4	www2.mdpi.com Internet Source	1%
5	wicri-demo.istex.fr Internet Source	<1%
6	www.ncbi.nlm.nih.gov Internet Source	<1%
7	www.researcherslinks.com	

	Internet Source	<1%
8	Submitted to Jaypee University of Information Technology Student Paper	<1%
9	Submitted to Ain Shams University Student Paper	<1%
10	Asmita Ray, Avantika Tiwari, D. Chandra Mouli. "Chapter 11 Early Screening of COVID-19 from Chest CT Using Deep Learning Technique", Springer Science and Business Media LLC, 2021 Publication	<1%
11	ebin.pub Internet Source	<1%
12	www.coursehero.com Internet Source	<1%
13	Submitted to Democritus University Student Paper	<1%
14	Submitted to CTI Education Group Student Paper	<1%
15	en.wikipedia.org Internet Source	<1%
16	repository.helmholtz-hzi.de Internet Source	<1%

17	LP Awasthi, HN Verma. "COVID-19 current scenario", Journal of Human Virology & Retrovirology, 2021 Publication	<1 %
18	sciendo.com Internet Source	<1 %
19	(4-10-03) http://134.214.92.116/gratuits/imprim.php?page=nav&action=list_item&num=610&&imprim=yes& Internet Source	<1 %
20	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
21	Submitted to Malta College of Arts,Science and Technology Student Paper	<1 %
22	www.expresshealthcaremgmt.com Internet Source	<1 %
23	www.worldometers.info Internet Source	<1 %
24	Submitted to South University Student Paper	<1 %
25	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1 %
	Submitted to University of Nottingham	

26	Student Paper	<1%
27	m.moam.info Internet Source	<1%
28	E. C. Freundt, L. Yu, C. S. Goldsmith, S. Welsh et al. "The Open Reading Frame 3a Protein of Severe Acute Respiratory Syndrome-Associated Coronavirus Promotes Membrane Rearrangement and Cell Death", Journal of Virology, 2009 Publication	<1%
29	M Dunowska. "Cross-species transmission of coronaviruses with a focus on severe acute respiratory syndrome coronavirus 2 infection in animals: a review for the veterinary practitioner", New Zealand Veterinary Journal, 2023 Publication	<1%
30	bmcgenomics.biomedcentral.com Internet Source	<1%
31	www.ghtcoalition.org Internet Source	<1%
32	Jessica A. Plante, Brooke M. Mitchell, Kenneth S. Plante, Kari Debbink, Scott C. Weaver, Vineet D. Menachery. "The Variant Gambit: COVID's Next Move", Cell Host & Microbe, 2021	<1%

	Publication	
33	tudr.thapar.edu:8080 Internet Source	<1 %
34	Submitted to La Trobe University Student Paper	<1 %
35	Qi Lu, Yuan Shi. "Coronavirus disease (COVID-19) and neonate: What neonatologist need to know", Journal of Medical Virology, 2020 Publication	<1 %
36	fdocuments.in Internet Source	<1 %
37	quarxiv.authorea.com Internet Source	<1 %
38	www.frontiersin.org Internet Source	<1 %
39	drrajivdesaimd.com Internet Source	<1 %
40	www.limsforum.com Internet Source	<1 %
41	www.researchgate.net Internet Source	<1 %
42	www.uspharmacist.com Internet Source	<1 %
43	Feng He, Yu Deng, Weina Li. "Coronavirus Disease 2019 (COVID-19): What we know?",	<1 %

Journal of Medical Virology, 2020

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		