GENOMIC SURVEILLANCE OF SARS-CoV-2 TO RECONSTRUCT INFECTION DYNAMICS AND PHYLODYNAMICS USING PHYLOGENETIC INFERENCE OF NEPAL

A Report for Master's Thesis
(2022)

Submitted To:
Central Department of Biotechnology
Tribhuvan University
Kirtipur, Kathmandu, Nepal

For partial fulfillment of the requirement of the M Sc. degree in Biotechnology

Submitted by:
Binod Khadka
Roll no.: BT 607/075
T.U. Regd. No.: 5-2-37-0247-2013

GENOMIC SURVEILLANCE OF SARS-CoV-2 TO RECONSTRUCT INFECTION DYNAMICS AND PHYLODYNAMICS USING PHYLOGENETIC INFERENCE OF NEPAL

A Report for Master's Thesis
(2022)

Submitted To:
Central Department of Biotechnology
Tribhuvan University
Kirtipur, Kathmandu, Nepal

For partial fulfillment of the requirement of the M Sc. degree in Biotechnology

Submitted by:
Binod Khadka
Roll No.: BT 607/075
T.U. Regd. No.: 5-2-37-0247-2013

Prof. Krishna Das Manandhar, PhD
Supervisor

Mr. Rajindra Napit
Co- Supervisor

Date: Dec 5, 2022

Recommendation

This is to certify that the research work entitled "**GENOMIC SURVEILLANCE OF SARS-CoV-2 TO RECONSTRUCT INFECTION DYNAMICS AND PHYLODYNAMICS USING PHYLOGENETIC INFERENCE OF NEPAL**" has been carried out by **Mr. Binod Khadka** under my supervision.

This thesis work was performed for the partial fulfillment of the Master of Science in Biotechnology under the course code BT 621. The result presented here is his original findings. I/we, hereby, recommend this thesis for final evaluation

.....
Prof. Krishna Das Manandhar, PhD
Supervisor
Head of Department
Central Department of Biotechnology
Tribhuvan University, Kirtipur, Nepal

.....
Mr. Rajindra Napit
Co- Supervisor
Lab manager
Center for Molecular Dynamics Nepal
Thapathali, Kathmandu, Nepal

Date: .....

Certificate of Evaluation

This is to certify that the thesis entitled "**GENOMIC SURVEILLANCE OF SARS-CoV-2 TO RECONSTRUCT INFECTION DYNAMICS AND PHYLODYNAMICS USING PHYLOGENETIC INFERENCE OF NEPAL"** presented to evaluation committee by Mr. Binod Khadka is found satisfactory for the partial fulfillment of Master of Science in Biotechnology.

.....

Prof. Krishna Das Manandhar, Ph.D
Head of Department
Central Department of Biotechnology
Tribhuvan University
Kirtipur, Kathmandu, Nepal

.....

Nishan Katuwal
(External Examiner)
Molecular Biologist and Incharge,
Molecular and Genome sequencing Research Lab,
Dhulikhel Hospital, Kathmandu University Hospital.

.....

Asst. Prof. AlinaShree Sapkota
(Internal Examiner)
Central Department of Biotechnology
Tribhuvan University
Kirtipur, Kathmandu, Nepal

.....

Prof. Krishna Das Manandhar, Ph.D
(Supervisor)
Central Department of Biotechnology
Tribhuvan University
Kirtipur, Kathmandu, Nepal

.....

Rajindra Napit
(Co-Supervisor)
Lab manager
Center for Molecular Dynamics Nepal
Thapathali, Kathmandu, Nepal

Acknowledgement

A dissertation journey that seemed never-ending has finally come to a close, and on this occasion of accomplishment, I would like to convey my heartfelt gratitude to everyone who made this possible.

First and foremost, I would like to express my heartfelt gratitude to my supervisors for their supervision of my thesis work. **Prof. Dr. Krishna Das Manandhar**, Head of Department of Central Department of Biotechnology, Tribhuvan University, Kirtipur, whose guidance, support, and motivation in every step of this thesis work has been a huge strength in completing all the associated works with ease and on time has been a huge strength. Despite his hectic schedule, he was always concerned about everyone's work in the lab and encouraged us to find new ways to better our research.

My co-supervisor **Mr. Rajindra Napit,** Assistant Professor, Central Department of Biotechnology, Tribhuvan University. It was a privilege to have the opportunity to work with him. I would not have been able to complete any of my study work without his help in understanding the various computational tools that I used during my research. He was constantly concerned and helpful during my work. I consider myself extremely fortunate to be his disciple.

I am grateful towards **Kirtipur Municipality-TU Biotech Laboratory** for providing me with the samples required for my wet-lab activities.

I would want to convey my heartfelt gratitude **Central Department of Biotechnology, Tribhuvan University**, for providing me the platform for this research and all respected **faculty members** of Central Department of Biotechnology. I am extremely grateful for all those guidance that they provided me during my stay in this department.

The Masters' Thesis Support grant from **University Grants Commission** has been a very motivating factor for the research and I am very much grateful towards **University Grants Commission** for providing me with this research grant.

I must also thank my seniors, **Dr. Eanstara Tuladhar, Ms. Sabita Prajapati, Ms. Roji Raut, Mr. Ramanuj Rauniyar, Mr. Suresh Joshi, Mr. Sailesh Adikari** who were always there to

help me with my lab work or any other important tasks. Their presence offered me assurance and confidence in my work. I've enjoyed learning from and working with them.

My gratitude would be inadequate without the support of my friends: **Sujeeta Maharjan, Salin Maharjan, Smita Shrestha, Suruchi Karna, Sushma Acharya**, and all other classmates. They turned every moment of the research into wonderful memories to enjoy for the rest of their lives.

A special thanks to all the lab assistants, staff, seniors and juniors of Tribhuvan University's Central Department of Biotechnology, who have been invaluable in their teachings, guidance, and help throughout this journey in this department, and all the work I've done so far is the result of their contributions as well. So, I am deeply grateful and honored to be a member of this institution that has introduced me to such lovely people.

Finally, and most significantly, none of this would have been possible without my family's love and patience. I'd want to dedicate this work to my family and show my heartfelt gratitude for their ongoing encouragement, prayers, and unwavering support.

Glossary Acronyms

COVID-19	Coronavirus disease 2019
DNA	Deoxyribonucleic Acid
E gene	Envelope protein Gene
EDCD	Epidemiology and Diseases Control Division
GISAID	Global Initiative on Sharing All Influenza Data
HMM	Hidden Markov Models
HPD	Highest Posterior Density
M gene	Membrane Glycoprotein
MERS	Middle East Respiratory Syndrome
MoHP	Ministry of Health and Population
MAFFT	Multiple Sequence Alignment Based on Fast Fourier Transform
N gene	Nucleocapsid Glycoprotein Gene
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NHRC	Nepal Health Research Council
NPHL	National Public Health Laboratory
nsp	Non-structural proteins
ORF	Open Reading Frame
RT-qPCR	Reverse Transcription Quantitative Polymerase Chain Reaction

S gene	Spike Glycoprotein gene
SARS	Severe Acute Respiratory Syndrome
SARS-CoV-2	Severe Acute respiratory Syndrome Coronavirus 2
SNPs	Single Nucleotide Polymorphisms
TMRCA	Time to Most Recent Common Ancestor
WGS	Whole Genome Sequence

Table of Contents

List of Figures

ABSTRACT

GENOMIC SURVEILLANCE OF SARS-CoV-2 TO RECONSTRUCT INFECTION DYNAMICS AND PHYLODYNAMICS USING PHYLOGENETIC INFERENCE OF NEPAL

Severe acute respiratory syndrome (SARS-CoV-2) have caused an unprecedented impact global public health and the economy. Both current and prospective interventions require a firm understanding of evolutionary and epidemiological parameters of novel SARS-CoV-2. Phylogenetic and phylodynamic approaches have provided critical insights into the spread of SARS-CoV-2 in international level, aided in tracking virus genetic changes, allowed the investigations of outbreaks and transmission chains, and informed for public health strategy. Nepal lack any such phylodynamic study on the SARS-CoV-2 genome to infer the epidemiological evolutionary state of virus using phylogenetic networks and growth trends. This study aims to investigate and reconstruct the evolutionary and epidemiological dynamics of SARS-CoV-2 virus in Nepal using the 278 genomes of SARS-CoV-2 sequenced in Biotechnology laboratory, TU. In this study, we used the Bayesian Phylodynamic pipeline and TransPhylo to analyze and evaluate the evolutionary, epidemiological and infection dynamics of SARS-CoV-2. Phylogenetic tree was inferred to study the evolution of SARS-CoV-2 variants, Reproduction number of virus and transmission chain of COVID-19 infection were estimated and the possible unsampled cases of SARS-CoV-2 was predicted. Phylogenetic analysis showed the presence of Omicron, Delta and Alpha variants from our dataset. Depending upon the Bayesian time-scaled phylogenetic analysis using the best fitting model showed us that the estimated evolutionary rate of SARS-CoV-2 was 1.226×10^{-3} substitutions per site per year. Estimated reproduction number showed two growing phases, one during early 2021 and other during early 2022. The mean estimated R value ranged from 0.495 to 4.8472. Similarly, inference of viral transmission chain using TransPhylo showed that inferred unsampled sources greatly outnumber the actual sequenced samples in the transmission network. Prediction of unsampled cases using the available genome sequences also suggested that very high cases are not being sequenced and are acting as unsampled source of infection. Our findings highlight the critical importance of establishing genomic surveillance programs to guarantee the current state of the epidemic and to ensure impactful decision making for the allotment of intervention initiatives against the most relevant variants.

Also, the study suggests the usefulness of various phylogenetic and phylodynamic approaches in supporting the surveillance of COVID-19 and other emerging disease outbreaks.

Keywords: SARS-CoV-2, Genomic surveillance, Phylogenetic, Phylodynamic, TransPhylo, Bayesian Inference, Reproduction number, Transmission chain

Chapter I

INTRODUCTION

1.1 Background of COVID infection

A novel coronavirus (previously designated as 2019-CoV), was identified as the etiological agent of a cluster of pneumonia cases in Wuhan City, Hubei Province, China with the first outbreak being discovered on 12 December 2019 (Dhama et al., 2020). The World Health Organization (WHO) on February 11, 2020, declared the CoV-associated disease as COVID-19 (Coronavirus Disease 2019), the illness caused by infection of this novel pathogen, SARS-CoV-2, which spread rapidly and with 11800 cases recorded from 114 countries, on 11 march 2020, WHO declared a pandemic (Poon & Peiris, 2020; Wu et al., 2020; Chiara et al., 2021; WHO Director, 2020). As of 1st April 2022, COVID-19 has infected over 200 nations, with over 486 million documented individual infections and a death toll of more than 6 million, constituting among the largest global health and socioeconomic dangers (Chiara et al., 2021; WHO, 2022).

Between humans, SARS-CoV-2 is predominantly transferred by respiratory droplets and personal contact, while airborne transmission also appears to be possible, with incubation period typically ranged between 2 to 14 days, but longer intervals have also been documented (Riou & Althaus, 2020; van Doremalen et al., 2020; Lauer et al., 2020). The most frequent symptoms include fever, dry cough, and general weariness or fatigue. Muscle discomfort, nasal congestion, runny nose, sore throat, and diarrhea are less-common symptoms. A small percentage of individuals developed pneumonia, severe acute respiratory syndrome and/or renal failure (Deng et al., 2020; Cevik et al., 2020: Tay et al., 2020).

Through metatranscriptomics methods, aided by PCR and sanger sequencing, the first complete genomic sequences of the new betacoronavirus were extracted in late December 2019 (Chiara et al., 2021). The first sequence of 2019-nCoV was presented online on behalf of Dr. Yong-Zhen Zhang and scientists at Fudan University in Shangai. Following that, five additional 2019-nCoV sequences from institutes across China (Chinese

CDC, Wuhan Institute of Virology, Chinese Academy of Medical Sciences, and Peking Union Medical College) were deposited on the GSAID database on January 11th, allowing researchers throughout the globe to begin analyzing the new virus (Gralinski & Menachery, 2020).

1.2 Brief History

Coronaviruses have been linked with major disease outbreaks in East Asia and the Middle East, the severe acute respiratory syndrome (SARS) first appeared in 2002 and Middle East respiratory syndrome (MERS) emerged in 2012 (Rodriguez-Morales et al., 2020). A novel betacoronavirus, SARS-CoV, that emerged in Guangdong, southern China, in November 2002 resulted in over 8000 human infections and 774 deaths in 37 countries during the 2002-03 outbreak (Chan-Yeung & Xu, 2003; Peiris et al., 2004). Similarly, being first detected in Saudi Arabia in 2012, the Middle East respiratory syndrome coronavirus (MERS-CoV), has been responsible for 2494 confirmed cases of infection and 858 deaths (Lee et al., 2016). Recent novel coronavirus, SARS-CoV-2 emerged in late 2019 causing a coronavirus disease 2019 (COVID-19), which has caused a continuing pandemic in numerous countries and territories, posing a worldwide danger in health. The International Committee on Taxonomy of Viruses (ICTV) classified SARS-CoV-2 into Severe Acute Respiratory Syndrome-related Coronavirus (SARSr-CoV) (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020). After its first instances in late December 2019 in Wuhan, Hubei Province, China, SARS-CoV-2 quickly spread throughout the globe, prompting the World Health Organization (WHO) to declare it as a worldwide pandemic on march 11, 2020. COVID-19 has destroyed numerous nations and overloaded many healthcare systems and long-term shutdowns caused by the epidemic have led to the loss of livelihoods which has had a rippling impact on the worldwide economy.

SARS-CoV-2 shows several hallmarks of these previous zoonotic outbreaks. It has striking resemblance to SARS-CoV, which infected people in Foshan, Guangdong province, China, in November 2002 and Guangzhou, Guangdong province, in 2003. Both SARS-CoV emergence occurrences were linked to live animal markets and featured species such as civets and raccoon dogs which were also sold live in Wuhan markets in 2019 and are

known to be vulnerable to SARS-CoV-2 infection (Guan et al., 2003; Xu et al., 2004; Xiao et al., 2021). SARS-CoV-2 also exhibit resemblance with the four endemic human coronaviruses: human coronavirus-OC43 (HCoV-OC43), human coronavirus-HKU1 (HCoV-HKU1), human coronavirus-229E (HCoV-229E), and human coronavirus NL63 (HCoV-NL63) (Holmes et al., 2021).

Epidemiological statistics show the Huanan market in Wuhan as an early and major epicenter of SARS-CoV-2 infection. Two of the three earliest confirmed COVID-19 cases, as well as 28 percent of all cases reported in December 2019, were directly related to this market selling wild animals (WHO, 2021). Thousands of live wild animals, including high-risk species like civets and raccoon dogs, were traded at Wuhan marketplaces in 2019, including the Huanan market. Following its shutdown, SARS-CoV-2 was found in environmental samples taken from the Huanan market, mainly in the western sector that sold wildlife and domestic animal items, as well as in associated drainage areas (Holmes et al., 2021; WHO, 2021).

1.3 History of Covid-19 in Nepal

Nepal reported its first case of covid-19 from a 32-year-old Nepalese student at Wuhan University of Technology in Wuhan, China, with no history of comorbidities, who returned to Nepal with COVID-19 infection. He arrived with a cough to the outpatient department of Sukraraj Tropical and Infectious Disease Hospital in Kathmandu. He became unwell on January 3, only six days before flying to Nepal. He stated that he had no exposure to Wuhan's so-called wet market. Throat swabs from the patient were found to be positive for 2019-nCoV in real-time RT-PCR tests at the WHO laboratory in Hong Kong (Bastola et al., 2020). After 8 weeks of first case, the second COVID-19 case was recorded, followed by the third on March 26, 2020. The fourth case was proclaimed on March 27, 2020, while the fifth case was declared on March 28, 2020. Both incidents were reported from outside the Kathmandu Valley. During the first week of April, four additional cases were discovered (Piryani et al., 2020). By April 10, 2020, a total of 3525 samples from suspected COVID-19 cases were screened using RT-PCR, of which 9 were found positive (0.2%) (Pun et al., 2020).

Beginning on March 19, 2020, a lockdown was implemented throughout the nation to stop the spread of COVID-19. In addition, the government announced the suspension of all physical educational classes, the postponement of examinations, the suspension of all government services, restrictions on travel, and a ban on gatherings of more than 25 people. Sharing of an open border between Nepal and India became one of the main causes of COVID-19 spreading to Nepal. The first verified case in South Asia is also the first instance in Nepal. There were 277461 confirmed cases of COVID-19 in Nepal at the end of the first wave, of which 272851 were recovered and 3031 passed away. Infection with COVID-19 had been dropping in Nepal for a few months when it abruptly began to increase at the beginning of the second wave. Since March 2021, there was an unanticipated exponential rise in the number of daily COVID-19 cases in India. Numerous studies indicate that this is a result of the extremely contagious delta-variant strain of SARS-CoV-2 that was discovered in India. The COVID-19 second wave began in Nepal as a result of the open border issue. A sampling study revealed that the delta-variant was responsible for 97% of the new instances. The health of those infected became much worse than it was during the first wave of COVID-19 due to the more deadly and highly transmittable characteristics of this delta-variant, which caused a large increase in infection and fatality rates (Pandey et al., 2022; Kandikattu et al., 2021; Joshi et al., 2021). After June, cases steadily reduced, and by the end of November 2021, the infection rate among confirmed patients was 2% per day. Just two weeks after it was discovered for the first time in South Africa on November 24, 2021, the Omicron (B.1.1.529) variant was discovered for the first time in Nepal on December 6, 2021. When Omicron first appeared, there was an increase in daily new cases that peaked on January 20, 2022, when there were a record-breaking 10,000 instances. Two months later, the third wave was over (by February 2022). The third wave had much fewer fatalities each day (n = 32) than the second wave (n = 246) (Pandey et al., 2022). **Error!**

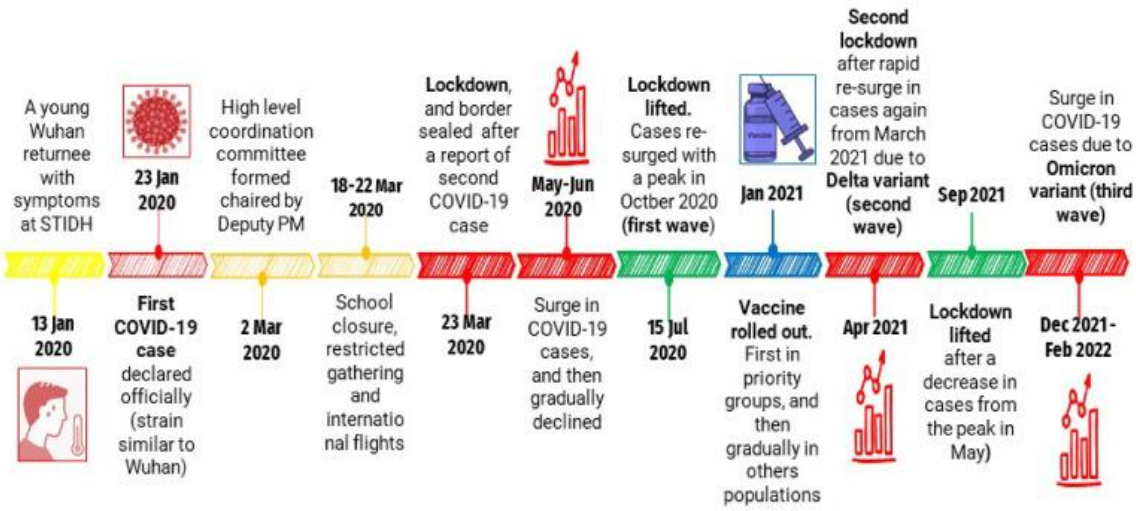


Figure 1 Chronology of key events of COVID-19 in Nepal. (Pandey et al., 2022)

1.4 Brief about corona virus

The diameters of coronaviruses (CoVs) range from around 80 to 120 nm, and they are generally spherical in shape with an envelope. Coronaviruses are distinguished by the club-shaped projections (spikes) that emerge from the surface of the virion. Its name comes from the fact that the spikes mimic the solar corona. Because of its significant sequence similarity to SARS-CoV, SARS-CoV-2 is assumed to have the same structure as SARS-CoV-2 (Kumaret al., 2020). The coronavirus's membrane, envelope, and surface viral protein spike are all embedded in a lipid bilayer produced from the host membrane that encases the virus's helical nucleocapsid which contains viral RNA (Kumar et al., 2020).

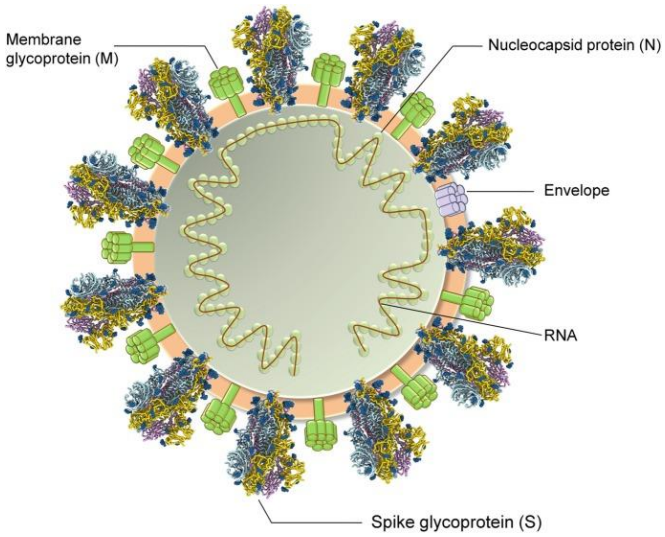


Figure 2 Structure of SARS-CoV-2 (Kumar et al., 2020)

Coronavirus comprises of positive sense single stranded RNA which is non segmented. The genomes of coronaviruses ranges between 26 and 32 kilobases (kb) in size, with 6-11 ORFs that code for 9680 amino acid polyprotein (G. Li et al., 2020). Sixteen different non-structural proteins (nsps) are encoded by the first open reading frame (ORF), which accounts for roughly 67% of the genome. The other ORFs code for various accessory and structural proteins. Genome of SARS-CoV-2 are without the hemagglutinin-esterase gene. It does, however, have two untranslated regions (UTRs) on either end: one at the 5' end (of 265 nucleotides) and one at the 3' end (of 358 nucleotides). No significant differences in ORFs and nsps were found between SARS-CoV-2 and SARS-CoV after analyzing sequence variation. Other than nsps, spike surface glycoprotein (S), envelope (E), membrane, and nucleocapsid protein (N), are the four primary structural proteins encoded by ORFs along with some accessory proteins (Chan et al., 2020).

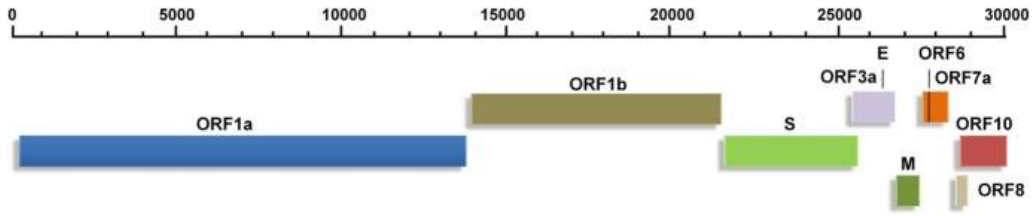


Figure 3 Genomic organization of SARS-CoV-2 (Kumar et al., 2020)

1.5 Genomic surveillance of SARS-CoV-2

Going beyond merely testing for SARS-CoV-2, genome sequencing enables researchers to categorize a virus as a specific variant and establish its lineage. Throughout the COVID-19 epidemic, genomic surveillance has been a crucial aspect of public health initiatives. Genomic sequencing offers vital information to study the development and spread of SARS-CoV-2, to improve molecular diagnostics, therapeutics, and vaccines, and to direct public health responses. Using next-generation sequencing (NGS) for genomic surveillance allows for the monitoring of infectious disease transmission and the detection of novel coronavirus strains, among other emerging pathogens. We can better combat the spread of infectious diseases by keeping an eye out for new cases with the use of nearly entire genome sequence data of pathogens (Brito et al., 2021).

The identification of SARS-CoV-2 as the causal agent of COVID-19 and the investigation of its global spread have both been proven to be greatly aided by virus genome sequencing.

The dynamics of an outbreak, such as the size of the epidemic over time, its spatial and temporal distribution, and the pathways through which it was propagated can be studied by analyzing virus genome sequences. Genomic sequences can also be used to aid in the development of diagnostic tests, pharmaceuticals, and vaccinations, and to track any potential change in effectiveness over time that could be traced back to mutations in the virus genome. As they multiply and propagate through a population, viruses of all kinds undergo modifications (or mutations). SARS-CoV-2 (the virus that causes COVID-19) and influenza, both of which employ RNA as their genetic material, change at an exponentially higher rate than DNA-based viruses. Each cycle of replication affords SARS-CoV-2 the chance to evolve. Since many mutations do not alter the key proteins involved in infection and transmission, the virus is able to continue to spread and cause disease despite them. There may be a competitive advantage over other lineages of SARS-CoV-2 when one of these alterations does affect the virus's capacity to propagate or cause disease. These advantageous lineages spread through a population and eventually become the norm. The Centers for Disease Control and Prevention (CDC) may label a genetic variation as "variant of interest" or "variant of concern" if it has properties that have an effect on public health (CDC, 2022).

Variants of concern (VOCs) including Alpha/B.1.1.7, Beta/B.1.351, Gamma/P.1, and Delta/B.1.617.2 have been identified and assessed early thanks to genomic surveillance. Because of their higher transmissibility and probable immunological escape from neutralizing antibodies elicited by natural infections and/or vaccinations, these lineages represent heightened worldwide public health hazards (The Lancet, 2021). The transmission, severity, and antigenicity of Variants of interest (VOIs) must also be tracked along with VOCs on a regular basis. It is crucial to monitor the variety of SARS-CoV-2 lineages that are moving around the world in near real-time to assist direct public health responses to emerging variants (Brito et al., 2021). A record number of SARS-CoV-2 viral genomes have been made available in public databases, with more than 11 million consensus genome sequences shared via EpiCoV database in the GISAID data science initiative as of 15th June, 2022.

1.6 Research Hypothesis

Null hypothesis: Genomic surveillance cannot be used to study the present context and to resolve the public health impact from SARS-CoV-2 like pandemics.

Alternative hypothesis: Genomic surveillance can be used to understand the present context and to resolve the public health impact from SARS-CoV-2 like pandemic.

1.7 Objectives

1.7.1 General Objective

1) To reconstruct the infection dynamics and phylodynamics of SARS-CoV-2 using genomic surveillance and phylogenetic inference.

1.7.2 Specific Objective

1. Whole Genome Sequencing of SARS-CoV-2 from COVID-19 cases.
2. Exploration of distribution of different variants of SARS-CoV-2 based on GISAID metadata.
3. Analysis of different viral variants circulating in Nepal through phylogenetic tools.
4. Estimation of Reproduction number of SARS-CoV-2
5. Estimation of Unsampled cases of SARS-CoV-2
6. Inference of Viral Transmission chain

1.8 Rationale of study

The virus strain is brand-new, and knowledge of the genomic epidemiology of SARS-CoV-2 allows for the recognition of transmission clusters, an understanding of its biological evolution rate, and valuable insights into the mechanisms underlying viral drug resistance, immune escape, and virulence/pathogenesis. New SARS-CoV-2 genomes are being acquired in several nations throughout the world, which will allow for monitoring of the pandemic's many characteristics. Genetic variety, relationships with clinical and epidemiological patterns and profiles, the value of diagnostic techniques, and the logical design of medicines and vaccine candidates are a few of these.

Despite a significant COVID-19 load in Nepal, only a small amount of data from complete, high-quality sequences is now available. To determine whether new viral variations were disseminating throughout the nations, it is necessary to increase our understanding of the genetic makeup of the new SARS-CoV-2 virus. As a result, in order to better understand the molecular epidemiology of COVID-19 in Nepal, we sought to characterize the SARS-CoV-2 whole genome sequence isolated from patients diagnosed as COVID-19 positive and to examine genetic variations. Furthermore, there is lack of study of epidemiological dynamics of viral epidemics in Nepal. Phylogenetic and phylodynamic approaches tend to play crucial role in studying viral spread, identify the outbreaks, analyze the transmission chains, predicts the growth rate and reproduction rate, evaluate and keep track of variants of concern and mutations related to them. We aim to estimate various phylodynamic and phylogenetic measures from obtained genome sequences that are significant to study the ongoing trend of infections.

Chapter II

LITERATURE REVIEW

2.1 The VIRUS (SARS-CoV-2):

SARS-CoV-2 belongs to the order Nidovirales, family Coronaviridae, and subfamily Orthocoronavirinae, which is further classified into four genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. Alphacoronavirus and Betacoronavirus are bat-derived viruses, whereas Gammacoronavirus and Deltacoronavirus emerged from avian and swine gene pools (Dhama et al., 2020).

2.1.1 Genomic organization

Coronaviruses have a nearly 30 kb non-segmented positive-sense RNA genome. The genome has a 5’ cap structure and a 3’ poly (A) tail, which allows it to function as an mRNA for replicase polyprotein translation. The replicase gene, which encodes non-structural proteins (nsps), takes up two-thirds of the genome, or approximately 20 kb, whereas structural and accessory proteins take up only around 10 kb. Present at the 5’ end of the genome are a leader sequence and an untranslated region (UTR) containing numerous stem loop structures essential for RNA replication and. Furthermore, at the start of each structural or accessory gene are transcriptional regulatory sequences (TRSs) that are essential for the expression of each of these gene. RNA structures necessary for viral RNA replication and synthesis are also present in 3’ UTR. The coronavirus genome is organized as 5’-leader-UTR-replicase-S(Spike)-E(Envelope)-M(Membrane)-N(Nucleocapsid)-3’ UTR-poly (A) tail with accessory genes interspersed within the structural genes at the 3’ end of the genome (Fehr & Perlman, 2015).

SARS-CoV-2 has a genomic length of 29,891 bp and a G+C content of 38%. The genome of coronaviruses is not atypical. At the 5’ end of the genome, the replicase gene, which is made up of two lengthy, overlapping open reading frames called ORF1a and ORF1b (Sawicki et al., 2007; Sola et al., 2015), takes up two-thirds of the genome. Polyprotein 1a (pp1a) is produced by ORF1a, whereas polyprotein 1b (pp1ab) is produced by 1 ribosomal frameshifting. Following that, the polyproteins are converted into 16 nonstructural proteins (nsps), which are essential for viral genome replication and transcription (Chiara

et al., 2021). The virus-encoded proteases (Main protease [Mpro], chymotrypsin-like protease [3CLpro], and papain-like proteases [PLPs]) cleave polyproteins into individual nsps. SARS-CoV-2 differs from other CoVs in that a unique short putative protein was discovered inside the ORF3 band, a secreted protein having an alpha helix and beta-sheet with six strands encoded by ORF8 (Dhama et al., 2020; Fuk-Woo Chan et al., 2020).

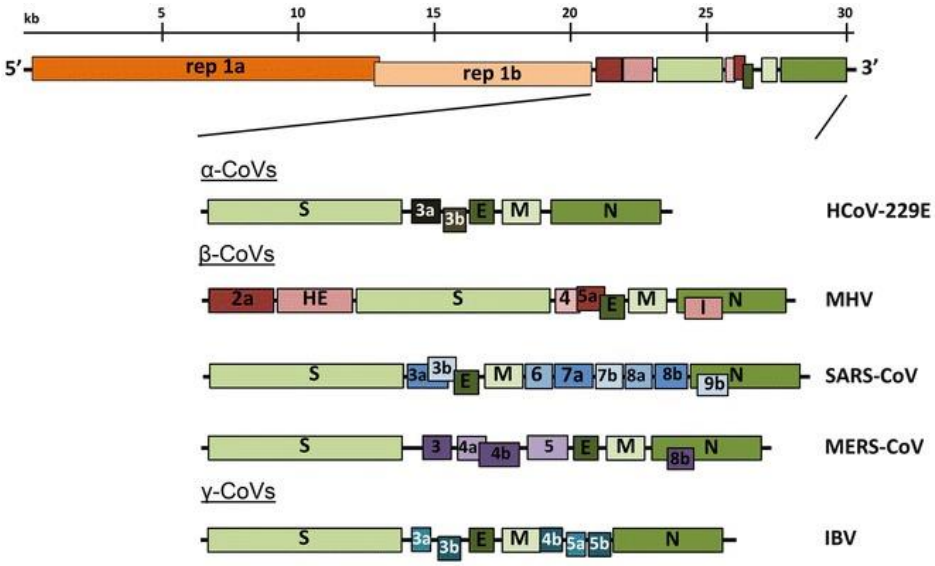


Figure 4 Representation of CoV's genomic organization. At the top is an illustration of the MHV genome. The structural and accessory proteins in the 3' sections of the HCoV-229E, MHV, SARS-CoV, MERS-CoV, and IBV are shown in the expanded portions below. The size of the genome and of individual genes are approximated but not drawn to scale using the legend at the top of the diagram. HCoV-299E: Human Coronavirus 229E, MHV: Mouse Hepatitis Virus, SARS-CoV: Severe Acute Respiratory Syndrome Coronavirus, MERS-CoV: Middle East Respiratory Syndrome Coronavirus, IBV: Infectious Bronchitis Virus (Fehr & Perlman, 2015).

2.1.2 Viral Structure

Coronaviruses are spherical in shape and have a diameter of around 125 nanometers (Fehr & Perlman, 2015). Coronaviruses are known for their club-shaped spike projections that protrude from the virion's surface. These spikes are a distinguishing characteristic of the virion, giving it the appearance of a crown, therefore the name coronavirus. The nucleocapsid is found within the virion's envelope. Coronavirus nucleocapsids are helically symmetrical, which is unusual among positive-sense RNA viruses but significantly more prevalent among negative-sense RNA viruses (Fehr & Perlman, 2015; Dhama et al., 2020).

Coronaviruses encode four primary structural proteins: spike (S), membrane (M), envelope (E), and nucleocapsid (N), which are all encoded inside the viral genome's 3' end.

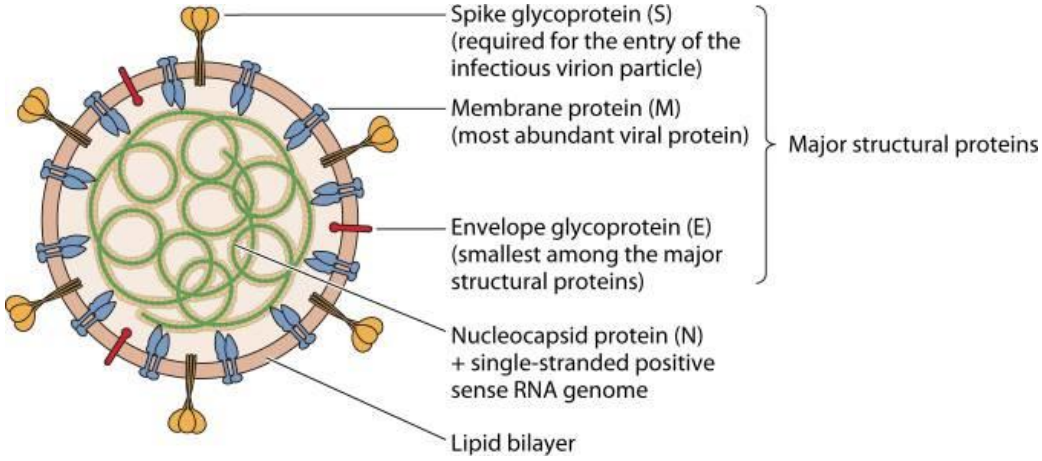


Figure 5 SARS-CoV-2 Virus structure (Dhama et al., 2020)

These structural proteins are essential for producing the structurally complete viral particle. Entry of coronavirus into host cells is guided by spike glycoprotein.

The unique spike structure on the virus's surface is made up of homotrimers of the virus's encoded S protein. The trimeric S glycoprotein is a class I fusion protein that facilitates host receptor attachment (Khan et al., 2020; Fehr & Perlman, 2015). In the majority of coronaviruses, a host cell furin-like protease cleaves the S protein into 2 distinct polypeptides S1 and S2. S1 subunit of S protein is responsible for the receptor-binding portion, while S2 subunit is responsible for the spike's stalk (Degroot RJ et al., 1987). The first (S1) is subdivided into N-terminal Domain (NTD) and C-terminal Domain (CTD). Both of these domains have a key role as receptor-binding domains, facilitating interactions with a wide range of host receptors. A receptor-binding motif (RBM) is found in the S1 CTD. Each coronavirus spike protein consists of a trimeric S1 subunit located atop a trimeric S2 stalk (F. Li, 2016).

The M protein being the most abundant protein in the virion particle, gives the viral envelope a definite structure. By attaching with the nucleocapsid it serves as the coronavirus assembly's key organizer (Dhama et al., 2020). M protein is a type III membrane protein that has three domains: a short N-terminal ectodomain, a triple-spanning transmembrane domain, and a C-terminal endodomain. The M protein is crucial during the budding, envelope development, assembly, and pathogenic stages of the virus lifecycle (Neuman et al., 2011).

E protein is smallest among the key structural proteins of the coronavirus and is the least understood one with the size of 8.4 to 12KDa. During virus's pathogenesis, assembly and release, this protein serves several functions and the inactivation or lack of this causes change in viral pathogenicity due to alterations in shape and tropism of viruses (Schoeman & Fielding, 2019; Dhama et al., 2020).

Coronavirus N protein has several functions. It performs various pivotal tasks, including complex formation with the viral genome, facilitating M protein interactions during virion assembly, and increasing the virus's transcription efficiency (Sheikh et al., 2020). N protein has three distinctive and highly conserved domains, including a CTD, an RNA-binding domain or a linker region (LKR), and an NTD. The NTD is significantly divergent in length and sequence and associates with the 3' end of the viral genome, possibly through electrostatic interactions (Dhama et al., 2020).

A subset of beta-coronaviruses also has a fifth structural protein called hemagglutinin-esterase (HE). Protein has acetyl-esterase activity, binds sialic acids on surface glycoproteins, and functions as a hemagglutinin. Some researchers believe that these actions promote S protein-mediated cell entrance and spread of virus through mucosa (Klausegger et al., 1999).

Other than these important structural proteins, genome of SARS-CoV-2 encodes 15 non-structural proteins (nsps), including nsps 1 through 10 and 12 through 16, as well as 8 accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and ORF14). Each of these proteins serves a distinct function in the process of viral replication (Wu et al., 2020).

2.2 Next Generation Sequencing and sequencing of SARS-CoV-2

Sanger sequencing technique was heavily used in traditional molecular diagnostic methods. Though efficient for tiny DNA segments, it is time-consuming and inefficient for big sequence pieces. Recent improvements in genome sequencing have resulted in the development of next generation sequencing (NGS) technologies. NGS refers to a group of technologies that use massively parallel sequencing methodologies to generate millions of short read sequences in significantly less time, at a lower cost, and with higher throughput than Sanger sequencing (Kang et al., 2020).

The next generation sequencing, with its rapid advancement, has provided unparallel capacity to tackle issues in multiple fields of molecular biology, resulting in numerous discoveries and fresh insights. Thanks to the development of novel library preparation methods, computational pipelines for processing massive amounts of sequencing data, and improved analysis methodologies, NGS is beings used in a variety of fields.

Next generation sequencing (NGS) employs massively parallel technologies in which millions of sequencing reactions take place and are recorded at the same time. The stages for completing NGS may be generalized, and they are as follows (Clark et al., 2019).

1. isolation of genomic DNA or RNA (which is reverse transcribed into cDNA);
2. creation of an NGS library;
    a. fragmentation of DNA into tiny double-stranded pieces of about identical size;
    b. modification of the fragments to make them compatible with the sequencing platform.
3. Separate the library pieces and build clusters of identical copies on a solid surface;
4. Sequence each cluster; and
5. Analyze the results.

NGS sequencing methods have quickly become the technique of choice for a variety of applications in virology, including the discovery of novel viruses from metagenomic data, the reconstruction of whole or nearly complete viral genome sequences, and the study of viral evolution and quasispecies (Smits et al., 2014; Domingo et al., 2012). Among the greatest benefits of NGS-based techniques is the ability to rebuild full-length viral genomes even for unknown or poorly defined viruses, either from culture-enriched viral preparations or straight from clinical samples. In the instance of SARS-CoV-2, both second and third generation NGS technologies have been effectively employed, and multiple specific library preparation techniques have been separately created by different manufacturers (Chiara et al., 2020).

With the advancement in sequencing techniques, genomic sequences from the virus that causes COVID-19, SARS-CoV-2, are now being created and published at an unprecedented rate. SARS-CoV-2 genomes may now be sequenced within hours or days of a case being discovered, thanks to recent technical developments. Virus genome sequencing has already been helpful in identifying SARS-CoV-2 as the causal agent of COVID-19 and tracking its global dissemination. Furthermore, viral genome sequences may be utilized to study outbreak dynamics, such as variations in epidemic size over time, spatiotemporal spread, and transmission mechanisms. Furthermore, genomic sequences can aid in the development of diagnostic tests, medicines, and vaccinations, as well as in determining if variations in effectiveness over time are due to changes in the viral genome. Analysis of SARS-CoV-2 viral genomes can thereby supplement, augment, and assist COVID-19 control measures (WHO, 2021).

In the first two decades of the 21st century, there has been a revolutionary change in the application of virus genomes to disease outbreaks, with the ability to explore genomic epidemiology in near-real time, as opposed to the long procedures and retrospective analyses of the past. Rapid declines in per-base cost and sample-to-result turnaround time, increases in the volume of generated data and the computational capacity needed to process it, and also the advancement of easily deployable, price-effective bench-top sequencing equipment have all contributed to the widespread application of sequencing (Roy et al., 2016). Thus, sequencing has developed into an essential instrument in clinical microbiology for detection and characterization of viral pathogens in clinical samples, aiding in infection control, shedding light on epidemiological inquiries, and illuminating the evolutionary viral responses to vaccines and treatments (Houldcroft et al., 2017; WHO, 2021a).

Genomic reactions to the severe acute respiratory syndrome (SARS) outbreak in 2002-2003 and the present COVID-19 pandemic serve as contrasting examples of the growing usefulness of viral genomic sequencing for both clinical and epidemiological research (WHO, 2021a). Within the first month after a coronavirus was identified as the causal pathogen during the SARS epidemic, only three virus genomes were made publicly available, and within the following three months, only 31 were made available. While genomics was utilized to create molecular assays that may link the illness to the new

coronavirus at issue (Peiris et al., 2003; WHO, 2021a), the field was not yet advanced enough to study virus epidemiology in real time on a wide scale. However, metagenomic sequencing was used to determine the cause of unexplained pneumonia during the COVID-19 pandemic only a week after the first cases were reported (Zhu et al., 2020). In early January of 2020, the virus was officially identified as a novel coronavirus (SARS-CoV-2, also known as 2019- nCoV). Before the middle of January, six genomes were made public, paving the way for the swift creation of diagnostic assays and methodologies for deep virus genomic sequencing. More than 60,000 nearly full viral genomes were sequenced in the first six months after the detection of SARS-CoV-2, and this number is still rising as the virus spreads around the world. Genomes have often been generated within days of case identification and utilized to understand virus propagation during a pandemic (WHO, 2021a). SARS-CoV-2 has hence originated as scientific setting where genome sequences can be created quickly and more easily, and used to answer a wide range of public health problems.

2.3 Illumina MiSeq

Major developments in next-generation sequencing (NGS) technology over the past ten years have produced several significant advances in the study of human disease, the detection of cancer mutations, metagenomics, agricultural, and evolutionary biology. First introduced in 2011, The MiSeq system works as a small benchtop sequencer that utilized Illumina's sequencing-by-synthesis (SBS) technology (San Diego, CA). Benchtop technologies for sequencing libraries, like MiSeq, can swiftly sequence libraries, generate hundreds of gigabytes of data, and carry out data analysis in a single integrated sequencing run. The MiSeq system comes with onboard cluster generation, data analysis, and access to BaseSpace®, the Illumina genomic analysis platform, which offers real-time data uploading on-site or via the internet (cloud), data analysis tools, and run monitoring (Ravi et al., 2018). The MiSeq System is the ideal platform for quick and affordable genetic analysis thanks to the NGS power it packs into a small footprint.

Figure 6 MiSeq System, a sequencing platform (Source of picture: official website of illumina)

2.3.1 Workflow of Illumina MiSeq NGS:

The Illumina next-generation sequencing (NGS) method is based on sequencing-by-synthesis (SBS) and reversible dye-terminators that make it possible to identify single bases as they are added to DNA strands. The basic workflow of Illumina sequencing involves following steps (https://www.cd-genomics.com/blog/principle-and-workflow-of-illumina-next-generation-sequencing/) (Ravi et al., 2018)

Step 1. Library preparation

Through fragmentation, genomic DNA is broken into pieces that are 200 to 500 bp long. The 5' and 3' adapters are then added to the ends of these small segments. This process, called "tagmentation," combines the fragmentation and ligation reactions into a single step, making the library preparation process much more efficient. The adapter-linked fragments are then amplified with PCR. The library of sequences is put together.

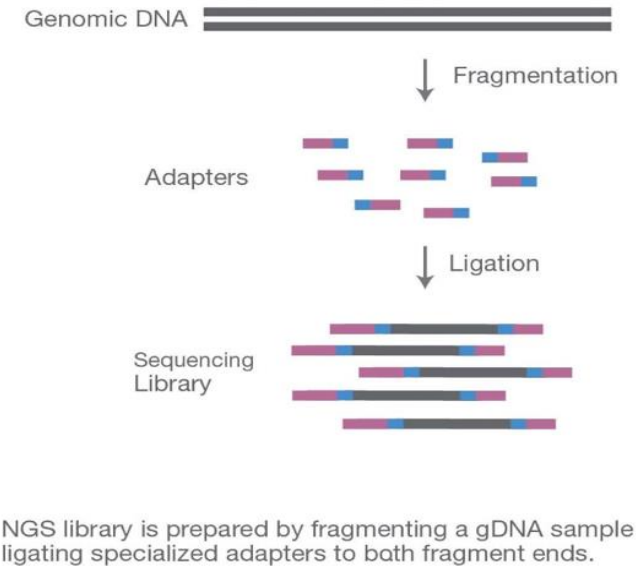


Figure 7 Library preparation for NGS (Source of Picture: from website of cd-genomics)

Step 2. Cluster generation

The sequencing reaction is carried out entirely within the flow cell, which is both an adsorption route for mobile DNA fragments and the central reactor vessel for sequencing. When the sequencing library is passed through a flow cell, random DNA fragments will adhere to the lanes. Each flow cell has eight lanes, and on the surface of each lane are a set of adapters that can pair with those added to the ends of the DNA fragment during the library building process. This allows the flow cell to adsorb the DNA after the fragment has been assembled, and it also allows the DNA to undergo amplification via bridge PCR. Bridge PCR was done using the adapters on the surface of the flow cell as templates. After many cycles of amplification, each DNA fragment will ultimately be clustered together in bundles at its respective location. Each bundle will have many copies of the same DNA template. The base signal is amplified in order to fulfill the signal requirements of the sequencing process. After the clusters are generated, the templates can be used for sequencing (Mardis, 2008).

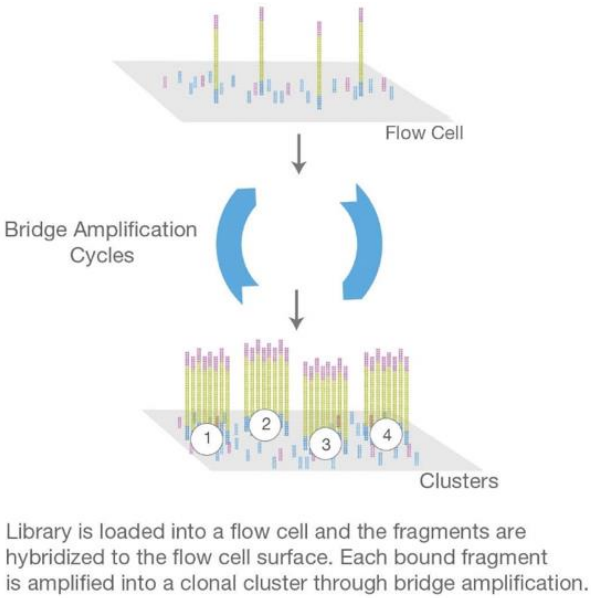


Figure 8 Cluster generation through bridge amplification on the surface of flow cell (Source of Picture: from website of cd-genomics)

Step 3. Sequencing

Sequencing-by-synthesis (SBS) is the basis for the sequencing procedure. The reaction system was supplemented with DNA polymerase, connector primers, and 4 dNTPs labeled with fluorescent markers specific to their respective bases. In order to ensure that only one base is inserted at a time during sequencing, the 3'-OH of these dNTPs is chemically protected. After the synthesis reaction is complete, any leftover free dNTPs and DNA polymerase are eluted. A fluorescence excitation buffer solution is then added, a laser is used to excite the fluorescence signal, and the resulting fluorescence is recorded using optical equipment. Computer processing of the optical signal is then used to convert those optical signals into sequencing base pairs. A chemical reagent is applied after the fluorescence signal is recorded to remove the dNTP 3'-OH protecting group and to quench the fluorescence signal, allowing the next round of sequencing reaction to proceed (Bentley et al., 2008).

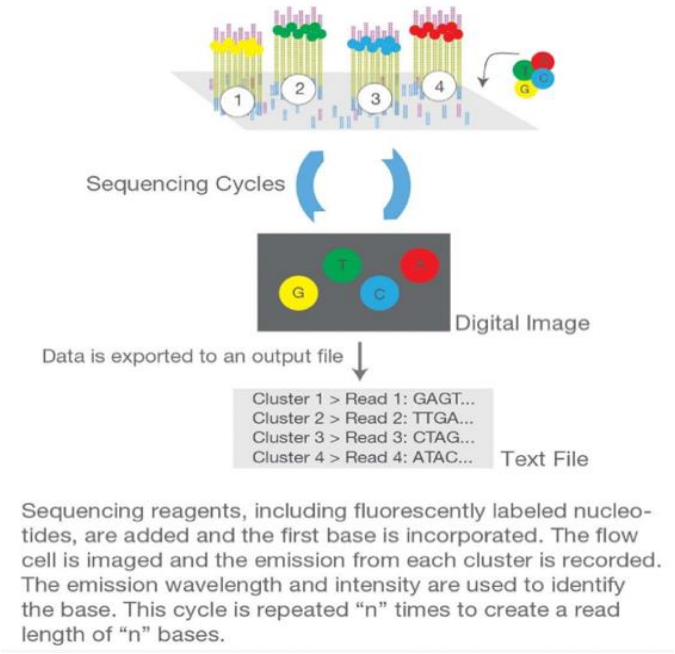


Figure 9 Sequencing through basis of Sequencing-by-synthesis(Source of Picture: from website of cd-genomics)

Step 4. Alignment & Data analysis

After that the sequence reads from a new sample are aligned to a reference genome, a wide range of bioinformatics analyses can be performed, including SNP/InDel/SV/CNV calling, annotation and statistics, population genetics, pathway enrichment analysis, and many more.

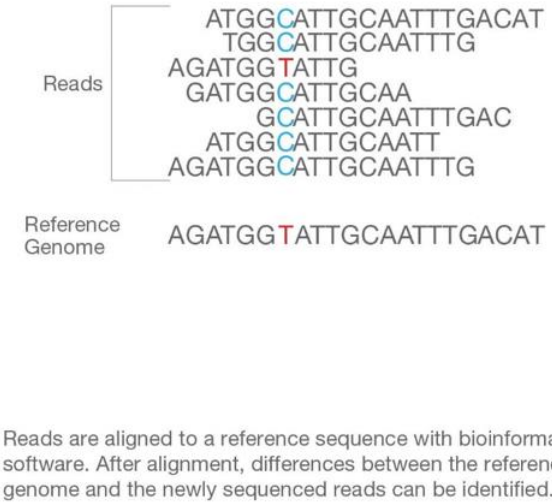


Figure 10 Sequence alignment and Data analysis (Source of Picture: from website of cd-genomics)

2.4 Phylodynamic and phylogenetic study and its importance

As a result of the 2019 coronavirus disease (COVID-19) pandemic, a massive international effort to sequence the genomes of pathogens was launched, and millions of genomes were created and made available to the public at an unprecedented rate. Throughout the pandemic, the epidemiological and virological situation on a worldwide scale was continually changing, making genome sequence analysis crucial for keeping up with changes. Phylogenetic and phylodynamic techniques are frequently used in concert with other data sources to decode information in sampled genomes. Quantifying the global spread of a virus, locating an outbreak and its transmission chain, estimating its growth and reproductive rates, adjusting for gaps and lags in surveillance, locating and tracking mutations of interest, discovering and analyzing variants of concern, and studying the intra-host viral evolution are all examples of applications of such phylogenetic and phylodynamic analyses. Phylogenetic and phylodynamic analyses have served as the foundation for several useful applications of virus genomes in informing public health actions (Ingle et al., 2021).

2.4.1 Phylogenetics

Phylogeny is the term used to describe the relationship between biological lineages that share a common ancestor (Gorbalenya, 2008). This term also describes the process used to reconstruct these relationships. Using character-state data, phylogenetics offers a way for creating hypotheses regarding ancestor-descendant relationships. The resulting phylogeny makes an effort to explain the character states that have been observed in the sequences we sampled as having been evolved from a single common ancestor in the past, via a series of typically unobserved (unsampled or extinct) postulated intermediate ancestors indicated by internal nodes or branches on a bifurcating tree. Phylogenetic approaches often look for the answer that involves the fewest possible evolutionary stages (parsimony) or one that maximizes the likelihood of the data given the tree. A third option is a Bayesian strategy, which utilizes Bayes theorem in order to estimate a probability distribution for important population parameters. The approach has an advantage above maximum likelihood estimate since it can incorporate prior data (priors)

for the occurrences, such as a prior distribution for the onset timing of the outbreak (Attwood et al., 2022).

Phylogenies have also been used to create a way to identify, define, and keep an eye on outbreak clusters and variants of concern (VOCs). While the WHO has adopted nomenclatures that provide names to specific constellations of substitutions that frequently occur together (for instance, VOC delta), the majority of other existing nomenclatures are based on lineage (for example, Pango and Nextstrain). According to the Pango nomenclature, lineages either roughly or precisely correlate to clades inferred on a reference tree. A clade is a monophyletic subtree on a phylogeny that only contains the offspring of the most recent common ancestor shown by the node connecting them to the entire tree. However, Pango lineages can include any relatively cohesive and exclusive (or nearly so) clustering of sequences on the global SARS-CoV-2 phylogeny. This is especially true when the cluster is linked to an outbreak, an epidemiologically significant phenotype (such as greater transmissibility), or any other notable trait, whether established or still under investigation (Attwood et al., 2022; O'Toole et al., 2022; WHO, 2021b).

2.4.2 Phylodynamics

Grenfell and colleagues used the word "phylodynamics" in 2004 to characterize the fusion of pathogen population dynamics with evolutionary dynamics (Grenfell et al., 2004). Phylodynamics aims to estimate the population dynamic parameters such as growth rate of an epidemic, reproduction number and generation time, using genetic sequence data and molecular phylogenies. Mathematical modeling techniques have been developed to incorporate data from phylogenetic trees into population dynamic models in parallel with the quickly developing techniques for sequencing pathogen genomes. Epidemiological data is complemented by genetic data, and the two together may help to better understand how infections spread and the efficacy of population-level interventions. Studying the genetic makeup of the pathogen population in a modeling framework may provide insight into the fraction of undiagnosed cases, particularly in circumstances when not all infected patients are discovered and identified (Kretzschmar, 2016). Phylodynamics has been used to add epidemiological data to phylogenetic analysis of the

pandemic. Using such models, demographic or epidemic parameters can be estimated over time. These parameters frequently include change in selection coefficients and changes in relative population size (including reproductive number and growth rate). By calculating the time to the most recent common ancestor (TMRCA) of a clade, phylodynamics can be used to date the first cases in a location and give public health experts an estimate of the delay between importation and first-case detection. Phylogeography, which employs phylogenetic techniques to comprehend the spatial distribution of lineages, has been employed throughout the pandemic to predict rates of virus (lineage) transit between locations (Attwood et al., 2022).

When it comes to pathogens that evolve quickly, the sampling times of the genome sequences could be used to establish the molecular clock, and the data can be regarded as if it came from a population that is measurably evolving. Under these conditions, the times at which branches form in a phylogenetic tree can show transmission dynamics (Ingle et al., 2021). Most models of transmission dynamics have been based on coalescence and birth-death processes and assume that branching events are closely related to the effective population size, N_e, from which the data were taken. From these models, epidemiological parameters can be figured out. For example, if you assume that the size of the population of infected hosts follows a mathematical function, you can infer the Re and sampling fraction (Frost & Volz, 2010; Stadler et al., 2012).

For phylodynamic inference, there are different ways to use statistics. Bayesian approaches are becoming more and more popular because they can be used to describe very complicated models, they can include independent information through the prior, and they can usually estimate parameters of interest in a single framework. Maximum likelihood techniques, on the other hand, usually require that analyses be done in separate steps. Because genome data sets are getting bigger and more complicated, fully Bayesian approaches are sometimes computationally impossible. This has led to the development of hybrid techniques, such as when the phylogenetic tree is figured out using maximum likelihood and epidemiological parameters are figured out using a Bayesian analysis (Ingle et al., 2021).

2.4.3 Reproduction number

Since the COVID-19 started in China, epidemiological models have been used to make predictions and projections about the epidemic. These predictions and projections have been at the center of planning and implementing strategies to stop the epidemic. Researchers have implemented different models to figure out the optimal extent a number of cases reaches and the time for that peak to occur (Marimuthu et al., 2021a). There are various ways to study the disease progression. One of the ways to measure viral spread is the **basic reproduction number (R_0).** R_0 is the number of secondary infections caused by an infected person in a group of a completely susceptible population during the infection period of that disease (Dietz, 1993). In general, a disease spreads if the basic reproduction number is more than unity, and it dies out if the value is smaller than unity.

Effective reproduction number (R_e or R_t)

When figuring out the basic reproduction number, R_0, one of the major assumptions is that all people who are not infected with the virus have the same susceptibility towards the virus. Even though this is true for the most part at the start of an epidemic, the populations most likely to get infected change over time as some people recover or become immune against the disease or because of policies that protect them. This means that the actual number of secondary infections per infected person changes over time. The number of secondary cases per infected person in a population with both susceptible as well as non-susceptible people is the effective reproduction number and is denoted by R_e or R_t (You et al., 2020).

The reproduction number is based on three variables: how long an infected person can spread the disease, how likely it is for an infected person to spread the disease to a susceptible person when they come into contact, and how often people come into contact. The pathogen's ability to spread and how long it can spread are biological constants, but the amount of contact between people will vary, so R will change depending on this parameter. This shows why social distance was so important during the COVID-19 pandemic (Achaiah et al., 2020). At an effective reproduction rate of 2.5, it is predicted that 90% of the ongoing pandemic can be contained if 80% of the contacts can be successfully identified, quarantined, or isolated. If the effective reproduction number

could be reduced to less than 1.5, greater levels of control could be attained with less intensive contact tracing, and if it increases to 3.5 or more, the epidemic's course would be swift (Hellewell et al., 2020).

2.5 Computational Tools

2.5.1 MAFFT (Multiple Sequence Alignment Based on Fast Fourier Transform):

Multiple sequence alignment (MSA) is a fundamental method in many molecular biological research, from identifying important functional residues to inferring evolutionary history. MAFFT is an MSA program, first released in 2002. The program MAFFT can be used with several alignment strategies, including progressive alignment done on its own (using the Fast Fourier Transform) or progressive alignment done first, then iterative refining. MAFFT can include up to three steps. First, a progressive alignment is built using shared 6-tuples to build an estimated distance between each pair of sequences. Additionally, UPGMA generates a guide tree with changed linkage, after which sequences are aligned in accordance with the tree's branching structure (this step alone is called strategy FFT-NS-1). Based on the data gathered in the previous phase, the second step recalculates a distance matrix, and the progressive alignment is then completed again using a tree produced using the new matrix as a starting point (till this, the strategy is called as FFT-NS-2 and is the default step used by the software). The next stage is the iterative refinement, which employs a group-to-group alignment along with the tree-dependent restriction partition technique to optimize the Gotoh's weighted sum of pairs (WSP) score. The process is known as FFT-NS-i, meaning it uses an FFT method to quickly detect homologous regions existing in the sequences and is proceeded by an iterative phase of refining. FFT transforms each amino acid in a sequence into a vector that represents volume and polarity, two crucial aspects of substitution events, enabling the program to accurately forecast their occurrence (Katoh et al., 2002; Katoh & Standley, 2014; Nuin et al., 2006).

MAFFT version 7 has options for different alignment strategies, such as progressive methods (PartTree, FFT-NS-1, and L-INS-1), iterative refinement methods (FFT-NS-i, L-INS-i, E-INS-i, and G-INS-i), and structural alignment methods for RNAs (Q-INS-i and X-INS-i).

MAFFT can quickly find some region where homology is more obvious. After finding these areas, slower methods of dynamic programming are used to put these parts together into aa whole arrangement. So, speed was the best thing about the early versions of MAFFT. It is also one of the programs that is more accurate and can be used on its own as standalone app or through the web which gives many different kinds of output, including interactive phylogenetic tree (Mohamed et al., 2018).

2.5.2 AliView

AliView is an alignment viewer and editor tailored to the needs of phylogenetic datasets generated in the era of next-generation sequencing. AliView supports the most popular alignment formats, including FASTA, Phylip, Nexus, Clustal, and MSF, and can process large alignments of any size. Inspecting, sorting, erasing, merging, and realigning sequences during the manual filtering of massive datasets is a breeze because to the straightforward graphical interface. Furthermore, AliView may be used as a simple alignment editor for both small and large datasets (Arvestad, 2018; Larsson, 2014).

More complex and faster alignment editors are in demand as DNA and protein datasets continue to grow in size. What was absent in previously available programs was a streamlined and user-friendly interface for aligning, rearranging, deleting, and merging sequences, as well as a method for identifying degenerate primers in chosen semiconserved areas and ability to visually highlight different conserved regions. Som e of these functions may already be available individually in existing alignment editors, but not in combination. In addition to the basic functions that meet these specific needs, AliView is made with a full set of easy-to-use general functions that meet the commonest needs for preparation of multiple sequence alignment. AliView is presented as an alignment viewer plus editor with a powerful collection of capabilities that makes it possible to handle enormous datasets with ease. The user-friendly interface allows for endless alignment sizes and provides a clear visual overview for navigation (Larsson, 2014).

2.5.3 IQ-TREE

As successor to IQPNNI and TREE- PUZZLE, the IQ-TREE program was developed (hence the name IQ-TREE). The increasing growth of phylogenomic data necessitated IQ-TREE to

meet the growing demand for efficient phylogenomic software capable of processing massive amounts of data and delivering more sophisticated models of sequence evolution. In order to speed up the analysis, IQ-TREE can take advantage of multicore processors and distributed parallel computing. Standard sequence alignment formats like as PHYLIP, FASTA, Nexus, Clustal, and MSF are all supported as input in IQ-TREE. IQ-output TREE's consists of two files: a NEWICK tree file (.treefile) that can be viewed in tree viewer software like FigTree, Dendroscope, or iTOL, and a self-readable reporting file (name suffix.iqtree). Maximum likelihood (ML) based phylogenetic inference has been increasingly popular in recent years, and IQ-TREE is a popular and freely available software tool for doing this analysis. IQ-exceptional TREE's performance is the result of the clever integration of modern phylogenetic approaches that enhance the three essential processes in phylogenetic analysis: rapid model selection using ModelFinder, a powerful tree search algorithm, and an unique ultrafast bootstrap approximation. In comparison to other well-known ML phylogenetics software like RAxML and PhyML, IQ-tree TREE's search method shows promising performance in terms of computing times and likelihood maximization. When it comes to research in the medical field, IQ-TREE is also an essential part of the software ecosystem. Among the many widely used open-source programs that rely on it are Galaxy, Nextstrain, OrthoFinder, and QIIME 2 (Minh et al., 2020; Nguyen et al., 2015).

2.5.4 FigTree

FigTree is a software that can e used to graphically visualize the phylogenetic trees. Its primary goal is to display the summarized and annotated files that come from BEAST and other tools. The program's graphical user interface allows users to change different components of tree including the tree's rooted position, node labels, label for the tip of the tree, and scale axes. Print-ready PDFs of tree diagrams can be exported for use in publications or for use as templates in other graphics programs (Rambaut, 2018). FigTree can display maximum clade credibility trees generated by Bayesian phylogenetic analysis as well as other tree formats. The source tree file can be used to annotate both external and internal nodes, and the inclusion of a time axis is made possible by the inclusion of temporal information within the tree (Theys et al., 2019).

2.5.5 BEAST 2

BEAST 2 is cross-platform tool used for Bayesian phylogenetic analysis of the molecular sequences. It computes strict or a relaxed molecular clock models to estimate rooted, time-measured phylogenies. It can be used to reconstruct phylogenies and also provides a framework for evaluating evolutionary theories without being dependent on a particular tree topology. BEAST 2 use Markov chain Monte Carlo (MCMC) to compute an average throughout the tree space, giving each tree a weight based on the posterior probability of that tree. A graphical user interface for configuring common analyses is included in BEAST 2, as well as a number of tools for analyzing the outcomes (Bouckaert et al., 2014)(Bouckaert et al., 2019).

The BEAST tool is used to perform MCMC-based Bayesian phylogenetic inference. Its fundamental components are rooted time trees (also time networks in more recent advancements), which can be deduced from a variety of data sources. BEAST accepts sequencing data including nucleotides, amino acids, discrete and continuous morphological traits, language, codon models, microsatellites, SNPs, and user-defined discrete and biogeographical data (Bouckaert et al., 2019). Thanks to Bayesian inference, it is possible to incorporate data from a wide variety of sources—including, for example, DNA sequences from both living and extinct species. In addition to allowing for the inference of rooted time trees, BEAST also enables the answer to a wide range of micro- as well as macroevolutionary questions, including figuring out the age and the location of the origins of various species, mutation rate and migration rate, and the rate at which epidemics spread (Hinchliff et al., 2015).

BEAST 2 fills the same niche as other well-known Bayesian evolutionary analysis platforms like BEAST, MrBayes, and RevBayes and so uses many of the same models (Heled & Drummond, 2010; Hohna et al., 2016; Huelsenbeck & Ronquist, 2001). Despite being a complete redesign of BEAST 1, BEAST 2 keeps many of the same basic model elements, such as relaxed molecular clock models, Bayesian skyline models for nonparametric coalescent studies, multispecies coalescent inference with BEAST, and phylogeographical models. A BEAST 2 analysis is configured using input XML files, same as BEAST 1. These

files can be simply prepared or the majority of common analysis utilizing (BEAUti 2), a graphical user interface (Heled & Drummond, 2010).

Design requires active modeling decisions from the BEAST 2 user; there is no longer possibility to simply run an analysis by "default". This active involvement makes it possible to customize studies to individual data sets and research issues, considerably enhancing the strength of the package. However, it also significantly adds to the complexity and makes it simpler to erroneously utilize the wrong models or introduce errors resulting beginners to opt for other simpler options(Barido-Sottani et al., 2018). BEAST 2 is fundamentally a time-tree model that predicts rooted phylogenies () using sequencing data ($\mathcal{D}$) with branch lengths expressed in calendar time. It simultaneously estimates population dynamics parameters (η) (including speciation/extinction or transmission/recovery rates) and evolutionary parameters (θ) (namely the substitution rate). In order to draw conclusions, BEAST 2 samples from the posterior distribution, using the Markov chain Monte Carlo (MCMC) algorithm (Barido-Sottani et al., 2018),

$$\Pr[\mathcal{T}, \eta, \theta|\mathcal{D}] = \frac{\Pr[\mathcal{D}|\mathcal{T}, \theta]\Pr[\mathcal{T}|\eta]\Pr[\eta]\Pr[\theta]}{\Pr[\mathcal{D}]}$$

The analysis's output is a log-file providing samples of these states ($\mathcal{T}$, η, θ) that the MCMC algorithm encountered. The output of BEAST 2 (after removing the burn-in phase) is a set of samples from the posterior distribution. This occurs after a so-called burn-in phase in which each value ($\mathcal{T}$, η, θ) is reached by the chain at such a frequency proportionate to its posterior probability. To analyze the data and respond to the relevant research question, the user must carefully follow a number of stages in order to accomplish a successful and accurate study. A multileveled (i.e., hierarchical) model with several interrelated parts must be specified by the researcher. These parts include: (i) a suitable model explaining the evolution of the sequence information on a time-tree, along with the substitution as well as molecular-clock models (Pr[$\mathcal{D}$| $\mathcal{T}$, θ]); (ii) a phylodynamic model explaining the growth of the tree over time (Pr[$\mathcal{T}$, η]); and (iii) reasonable prior distributions for every parameters of the evolutionary models (Pr[θ] and Pr[η]).

The researcher must specify and perfect MCMC operators that suggest novel states for the model parameters in adding to the model components ($\mathcal{T}$, η, θ). An MCMC analysis is

rather likely to efficiently sample the posterior distribution by selecting suitable proposal algorithms. The researcher must then determine if the MCMC chain has converged and recovered a significant signal from the data once it has sampled a reasonably enough number of states (Barido-Sottani et al., 2018).

2.5.6 Tracer v1.7

Discovering the evolutionary relationships between taxa, such as genes, genomes, people, or species, is made easier with the use of Bayesian inference of phylogeny utilizing Markov chain Monte Carlo (MCMC). Given DNA sequence data and a variety of evolutionary models, MCMC techniques produce samples of the model parameter values, together with the phylogenetic tree that is estimated from their posterior distribution. To approximate the posterior values of interest that one publishes as the result of a Bayesian phylogenetic study, one must visualize, tabulate, and marginalize these samples. To make this effort easier, the Tracer (version 1.7) software package was created, which processes MCMC trace files with parameter sample data and allows users to interactively explore the posterior distribution in high dimensions. In addition to BEAST, BEAST2, LAMARC, Migrate, MrBayes, and RevBayes, Tracer can also process sample output from MCMC programs in other domains (Barido-Sottani et al., 2018).

With respect to continuous, integer, and categorical parameters, Tracer analyzes the posterior samples from all of the available parameters from a trace and displays their statistical summaries and visualizations. Tracer can also aggregate samples from several files and analyze it in addition to analyzing a single trace. The effective sample size (ESS), one such statistic that enables users to gauge the number of effective independent samples from the posterior distribution represented by the trace. In order to reconstruct epidemic dynamics, Tracer offers demographic reconstruction that produces a graphical plot. Constant size, exponential and logistic growth, as well as the non-parametric Bayesian skyline, skyride, and skygrid models, are among the models that are available (Drummond et al., 2005; Gill et al., 2013; Minin et al., 2008).

2.5.7 TransPhylo

TransPhylo is a R tool that uses genomic data to reconstruct infectious disease transmission. The input for TransPhylo is a dated phylogeny, with leaves representing

pathogens isolated from known affected hosts. The major result is a transmission tree that reveals who infected whom, as well as the possibility of unsampled persons acting as missing/unsampled transmission links. TransPhylo operates by coloring the phylogenetic branches with a different color for each host, be it sampled or not. Each part of the tree with a unique color reflects pathogen evolution inside a specific host. Color changes on branches correspond to the transmission events happening from one host to another (Didelot et al., 2021). Previous methodologies for phylogenetic building and inference of transmission tree have limitations in that they presume that all cases of outbreak have been sampled and sequenced and the outbreak has ended. These assumptions make transmission tree inference much easier, but they do not match epidemiological reality. An outbreak is rarely thoroughly sampled, all the cases may not be recorded to public health or may lack nucleic acid for sequencing, but genomic epidemiology studies are frequently unfolding in real-time, implying that an outbreak is being investigated before it is over. The few approaches that can handle unsampled instances do so by assuming no within-host diversity. TransPhylo new Bayesian technique for inferring transmission events using a temporal phylogeny that may be applied to an outbreak that is partially sampled, ongoing, or both. This method allows to infer when these transmission episodes happened, which, when combined with the person-to-person inference, resulting in a thorough and epidemiologically applicable outbreak reconstruction (Didelot et al., 2017).

Chapter III

MATERIALS AND METHODS

3.1 Research Design

The research design of the entire study is presented as below:

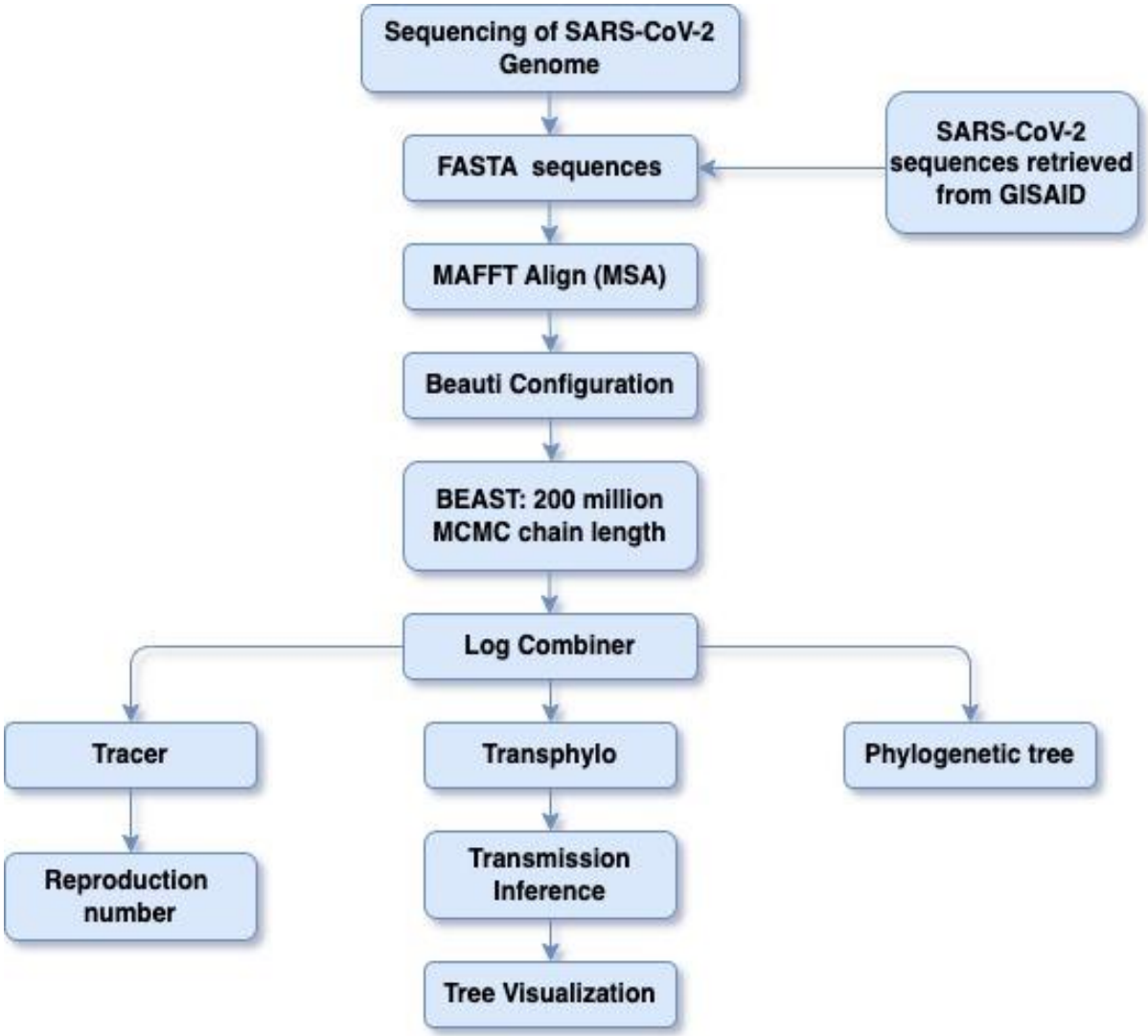


figure 11. General Outline of the research

3.2 Ethical Approval

The ethical approval for this research was obtained from Nepal Health Research Council (NHRC) (Registration Number: 274/2020). The privacy of the samples was maintained as

samples were kept anonymous by masking the patient's identity with unique sample code, only other details such as age, sex, address and date of sample collection were used.

3.3 Samples selection

Samples for whole genome sequencing of SARS-CoV-2 were collected from Kirtipur Municipality-TU Biotech Corona Laboratory based on the inclusion criteria of sufficient viral load (C_t value of <30).

3.4 Sample processing

Extracted RNA from SARS-CoV-2 samples were provided by Kirtipur Municipality- TU Biotech Corona Laboratory. According to Kirtipur Municipality- TU Biotech Corona Laboratory, RNA was extracted using automated extraction platform Liferiver EX3600 and SARS-CoV-2 was screened with the RT-qPCR in Kirtipur Municipality- TU Biotech Corona Laboratory. After confirming that the samples were positive for SARS-CoV-2, RNA were provided for further processing of library preparation for sequencing.

3.5 Library preparation and sequencing

All qPCR qualified samples were selected for the whole genome sequencing (WGS). The WGS of SARS-CoV-2 was carried out using commercially available COVIDSeq Test (RUO Version) (Illumina-USA) kit following manufacturer instructions. The prepared library was quality controlled for fragment size using Agilent Bioanalyzer DNA 1000 kit (50671504), while quantified using dsDNA High Sensitivity kit (Q32851) on qubit 3.0 (Invitrogen-USA). The final library was diluted to 4nm concentration and was denatured as per standard Illumina protocol to obtain final loading concentration of 10 pM library to be run on Illumina MiSeq platform using MiSeq V2 reagent kit (MS-102-2002). The detail protocol for library preparation using the extracted RNA is available in appendix 1.a and the steps for denaturation and sequencing is available in appendix 1.b.

3.6 Data Analysis

The obtained sequences were then analyzed with the usage of nextflow based pipeline nfcore-Viralrecon (r 2.4.1) (Ewels et al., 2020). It is an automated pipeline with integrated sequence quality control using FastQC, FastP followed by iVar for variant analysis and

bcftools to obtain the consensus sequences. The output from same tool was utilized to obtain variant calling file (VFC) which was verified from GISAID and Nextclade lineage check web app. The detail steps and codes are present in https://github.com/nf-core/viralrecon.

3.7 Phylogenetic reconstruction

SARS-CoV-2 full genome sequences were first aligned using Multiple Alignment using Fast Fourier Transform (MAFFT) v7.0 with default setting, subsequently Aliview was used to visualize and create phylip format. To evaluate the phylogenetic signals from our dataset we performed maximum likelihood analysis. Maximum likelihood tree was prepared from all the sequences (n=278 samples) using iqtree v2.2.0 (Nguyen et al., 2015) on auto mode for substitution model detection, 1000 bootstrap was run to infer final consensus tree. The tree was then refined with Augur v 17.7.0 from Nextstrain with tip dates. All the phylogenetic tree was visualized and edited in Figtree V1.4.4.

3.8 Bayesian Inference

For the estimation of the Bayesian molecular clock phylogenies of SARS-CoV-2, Bayesian inference analysis were performed for our dataset using a Markov Chain Monte Carlo (MCMC) framework implemented in BEAST v2.6.7. Foremost, all the sequences were filtered to obtain high coverage sequences containing <5% N to perform further Bayesian inference using BEAST V 2.6.7. Multiple sequence alignment (MSA) was performed using the MAFFT v7.0 for efficiency of the algorithm for long reads. The phylogenetic parameter selection was completed with the help of bModelTest on BEAST v 2.6.7. Parameter setup and xml file containing details for Bayesian inference was prepared using BEAUTI and was subsequently run-on BEAST. The most suitable substitution model was found to be Tamura Nei 93 distance. Further, Birth and Death skyline serial prior was selected with tip dates activated and relaxed log normal clock model.

3.8.1 Phylogenetic tree generation

Two hundred million long chain of Markov chain Monte Carlo (MCMC) iterations were carried out. The log of MCMC run was analyzed using Tracer 1.7.1 ensuring Effective sampling Size (ESS) of >200 in all the parameters. The trees were summarized using

Treeannotator 2.6.7 to obtain consensus tree having posterior probability of more than 0.6, maximum clade credibility and common ancestor node heights. The tree was then visualized and edited in Figtree v1.4.4.

3.8.2 Phylogenetic Estimation of Reproductive number and Become uninfectious interval

The Bayesian birth-death skyline (BDSKY) model was used to estimate rates of epidemics at different time period by measuring the change in basic reproduction number (R_e) and the rate of becoming uninfectious (denoted by δ) was also inferred. For this the log file of MCMC output was loaded into the bdskytools package in R to plot the BDSKY results. The required codes are described in https://taming-the-beast.org/tutorials/Skyline-plots/

3.9 Transmission tree Inference

The Bayesian based program TransPhylo which was created specifically to reconstruct transmission networks using dated phylogenetic data, was used to do transmission tree inference. TransPhylo has been employed and found well suited for various COVID-19 dataset analyses. TransPhylo makes it possible to completely infer transmission trees for an ongoing pandemic by predicting unsampled sources of infection and date of infection. Such estimation of unsampled nodes prediction is very important in the pandemic and epidemic condition particularly in understanding the path of viral transmission and how certain clusters were associated. The program was executed with viral generation time of 1 to 14 days, median of 5.5 days and sampling collection window of 2 to 14 days of infection.

3.9.1 Visualization of the viral transmission networks

The Transmission data, obtained from TransPhylo inference was visualized in Gephi 0.9.2. The built-in clustering algorithms Force Atlas 2 and Yifan Hu were used to re-order tree clustering with clustered transmission network for identification, as well as to visualize the data in comprehensive manner. The Pipeline used in this study was obtained from a previous similar study by (Perera et al., 2021).

Chapter IV

RESULTS

4.1 Study site:

Infectious and Viral Disease Research Laboratory, Central Department of Biotechnology-Tribhuvan University [IVDRL-CDBT-TU].

4.2 Sample descriptive analysis

A total of 278 samples were utilized for this research. However, only 2 samples, namely BB598 and BB466 were subjected to whole genome sequencing in our laboratory. The genome sequences of remaining 276 samples, sequenced in Biotech Laboratory, and their corresponding metadata were retrieved from GISAID for phylogenetic and phylodynamic inferences. Further analysis includes results using all 278 genomes which are presented accordingly.

4.2.1 Gender based distribution

Out of 278 cases study consisted of 51% male (n=141) and 31% female (n=87), while 18% (n=50) of the cases were with no gender information.

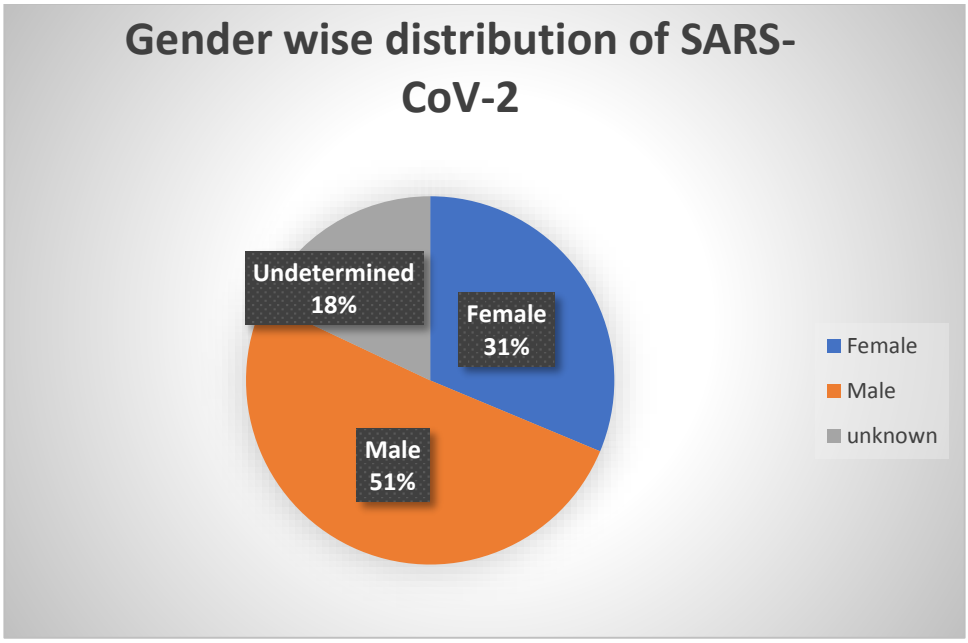


Figure 12 Gender wise distribution of samples. Male show dominance over female by 51% to 31% and the color code is as per indicated by the legend in the figure.

4.2.2 Age group

The age group ranged from as young as 8 months old to 90 years old. Among the 278 samples 50 of the samples had no information about their ages. Highest number of samples were observed in the age group of 31-40(n=52) followed by 21-30 (n=44) as shown in figure 13.

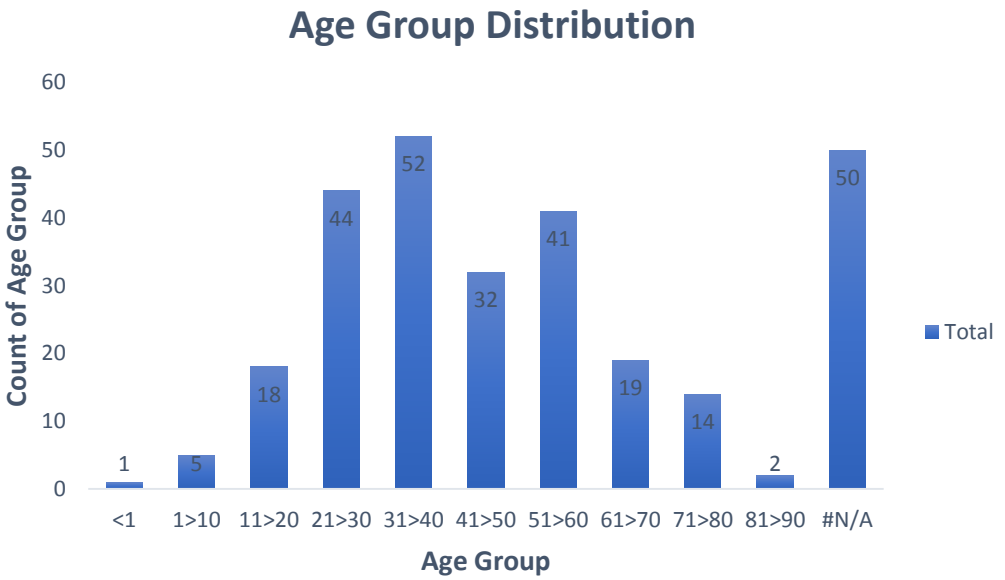


Figure 13 Age wise distribution of SARS-CoV-2 positive samples for sequencing. All the samples (y-axis: number of individuals) were distributed among different age groups (x-axis: different age groups). 50 samples at the end of the x-axis are with no information about which age group they belonged to

4.2.3 Location wise distribution

The distribution of samples across the district is indicated in following order: Kathmandu (141), Lalitpur (34) and Kaski (13). A bar graph showing respective number of genome samples belonging to different location is shown in figure 14 below.

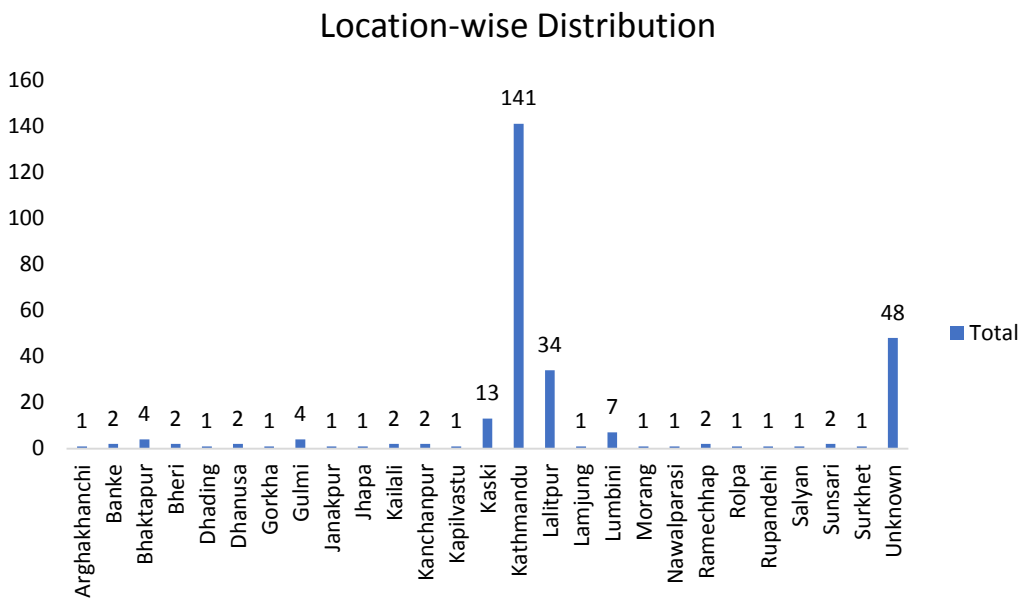


Figure 14 Location-wise Distribution of samples. All the 278 samples (y-axis: number of individuals) were distributed against different locations (y-axis: different locations)

4.2.4 Month-wise Distribution of cases

The samples were collected from April 2021 to June 2022. The sample size was inconsistent throughout the period. During the year of 2021, a small rise in sample collection was seen in the month of July (n=22), which gradually decreased in the succeeding months. The sample size peaked from the month of December (n=44) and reached the highest (n=65) on June of 2022. The month wise distribution of sample size is shown in Figure 15.

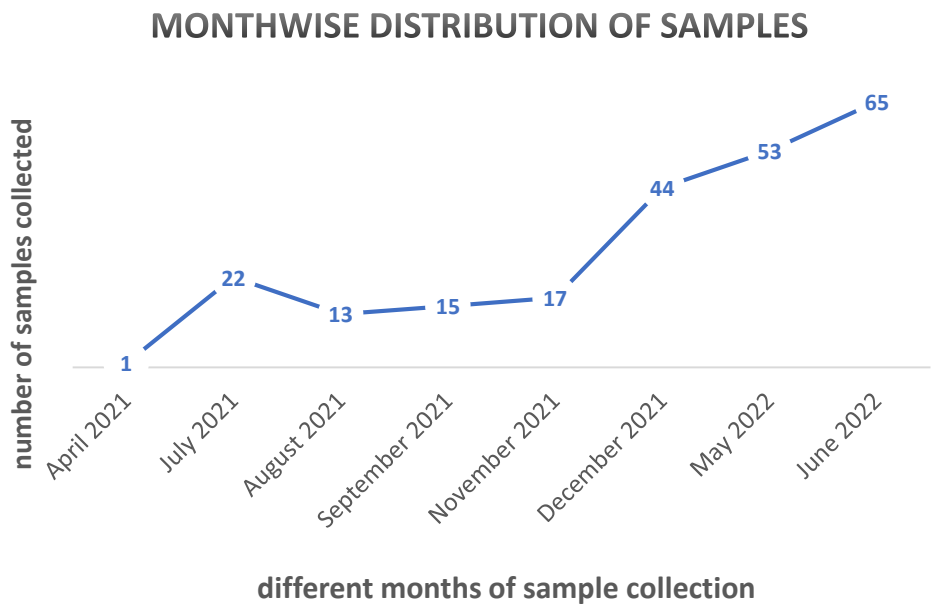


Figure 15 Month-wise distribution of samples showing sample frequency at different time interval. y-axis shows the number of samples and x-axis shows the progression of time

4.2.5 Variant based study

According to our data, Omicron, Delta, and Alpha variants were present, with Omicron having a genomic prevalence of 56% (n=156), Delta having a prevalence of 41%(n=114), and Alpha having a genomic prevalence of only one. Among the Omicron cases BA.2 was the most prevalence lineage (n=73) followed by BA.2.38 lineage (n=45). Similarly, among the Delta variants, Delta AY.39 lineage was the most prevalent (n=40) followed by the B.1.617.2 lineage (n=24).

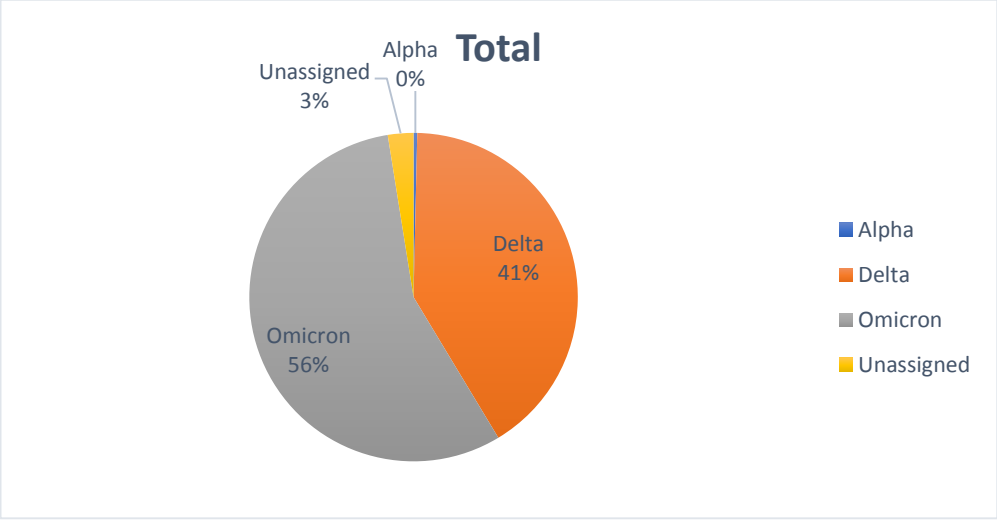


Figure 16 Frequency of different variants observed. Different variants are represented by different color codes as indicated by the legends.

4.2.6 GISAID Clades based distribution

Analysis of SARS-CoV-2 genomes using GISAID clade showed the occurrence of 5 different clades with clade GRA(n=130) being the dominant clade followed by GK (n=109). Clade GH was the least common clade with a single sample reported.

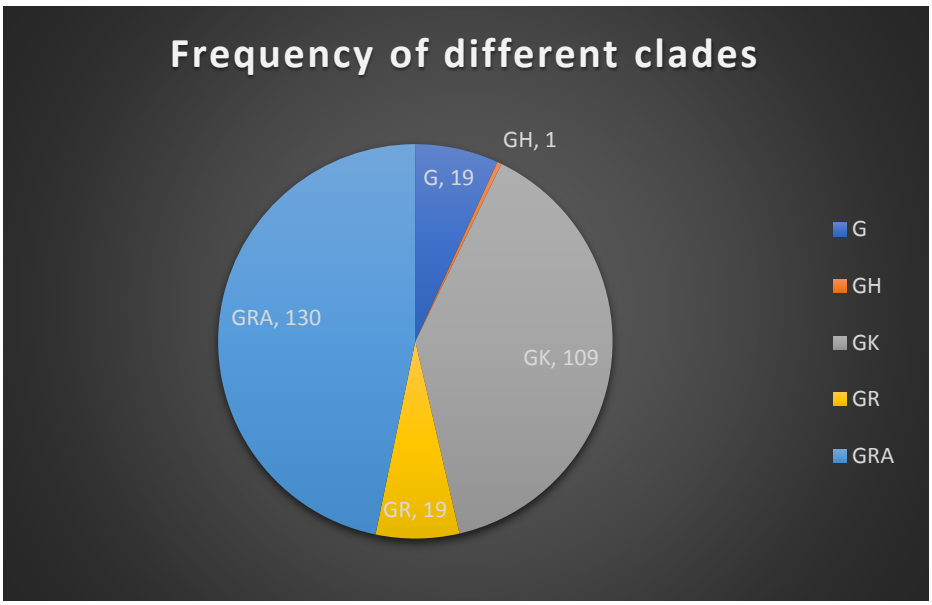


Figure 17 GISAID Clade-wise distribution. Different clades are represented by different colors as indicated by legend

A timeline-based analysis figure 18, shows us that at first alpha variants were observed and were present till July 2021. The second variant to be observed was delta variant. It was prevalent till December 2021, yet few cases were seen till may 2022 as well. Figure 18 shows that Omicron variant was seen during early December 2021 and soon became the most prevalent variant.

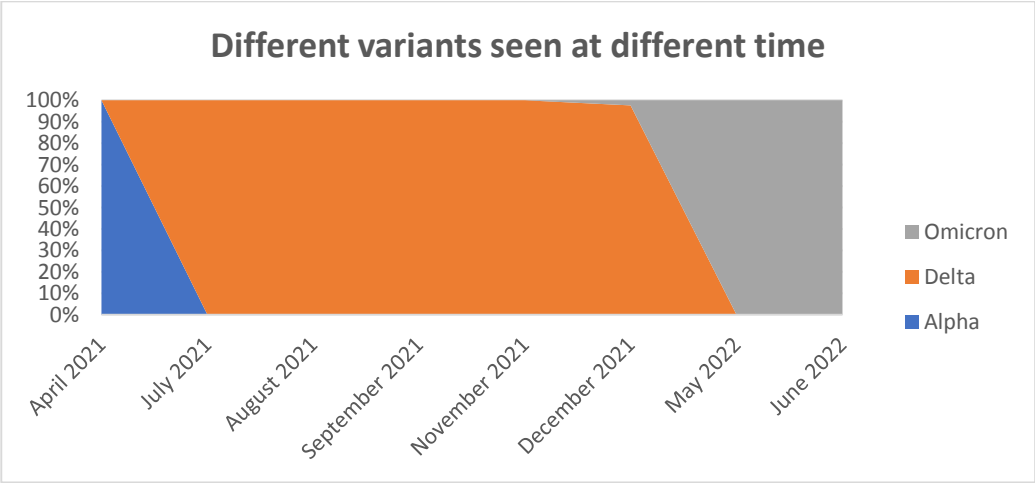


Figure 18 Timeline based study of occurrence of different variants

4.3 Phylogenetic analyses

The best fitting model was found to be Tamura Nei 93 distance substitution model. The estimated mean evolutionary rate from our study was found to be 1.226×10^{-3} substitutions per site per year. The phylogenetic constructed showed 3 distinct clusters of population in terms of evolution (Figure 19). One cluster primarily containing isolates of Delta variants, another comprised of isolates from Omicron variants. While one single isolate belonging to Alpha variant was seen as a separate cluster.

The two sequences from the samples obtained from Kirtipur Municipality- TU Biotech Corona Laboratory were found within two distinct clusters. Isolate BB595 was present in the cluster representing the Delta variants and was very close to the strain hCoV-19/Nepal/CDBT-TU-SQ1778/2021 isolated from NPHL. Meanwhile, BB466 was found in the cluster representing the Omicron variants (Figure 20). Phylogenetic tree suggests that BB466 was the early isolates present among the omicron variants.

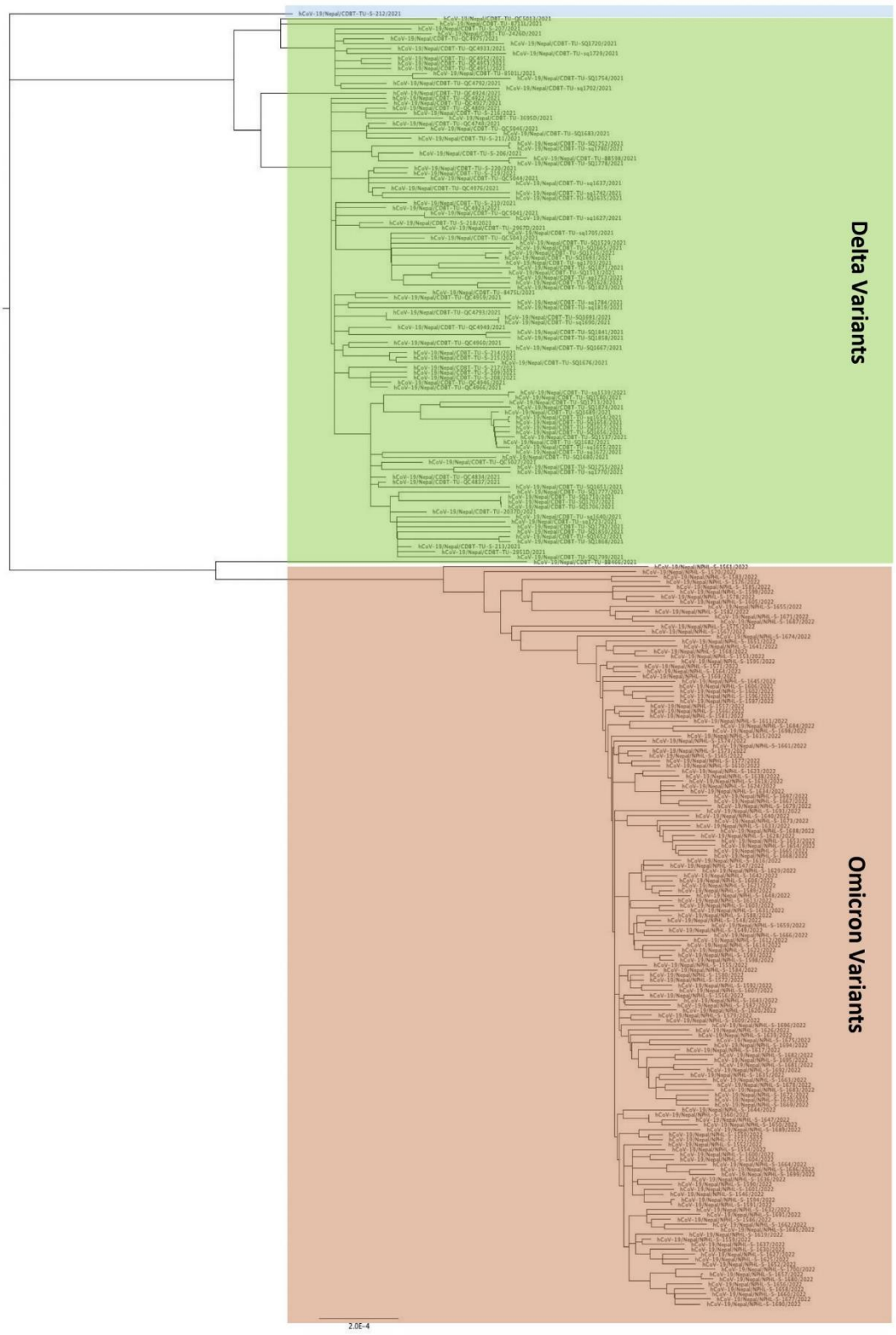


Figure 19 Phylogenetic tree constructed from our dataset. Bayesian Phylogenetic tree was constructed using BEAST v2.6.7 with Tamura Nei 93 model. Bayesian maximum clade credibility tree of SARS-CoV-2 virus sequences, having a posterior probability >0.60 and common ancestor node heights. Tree was visualized in FigTree v1.4.4. Three different variants are represented by 3 different color shades: Alpha variant by Blue, Delta variant by Green and Omicron variants by Red.

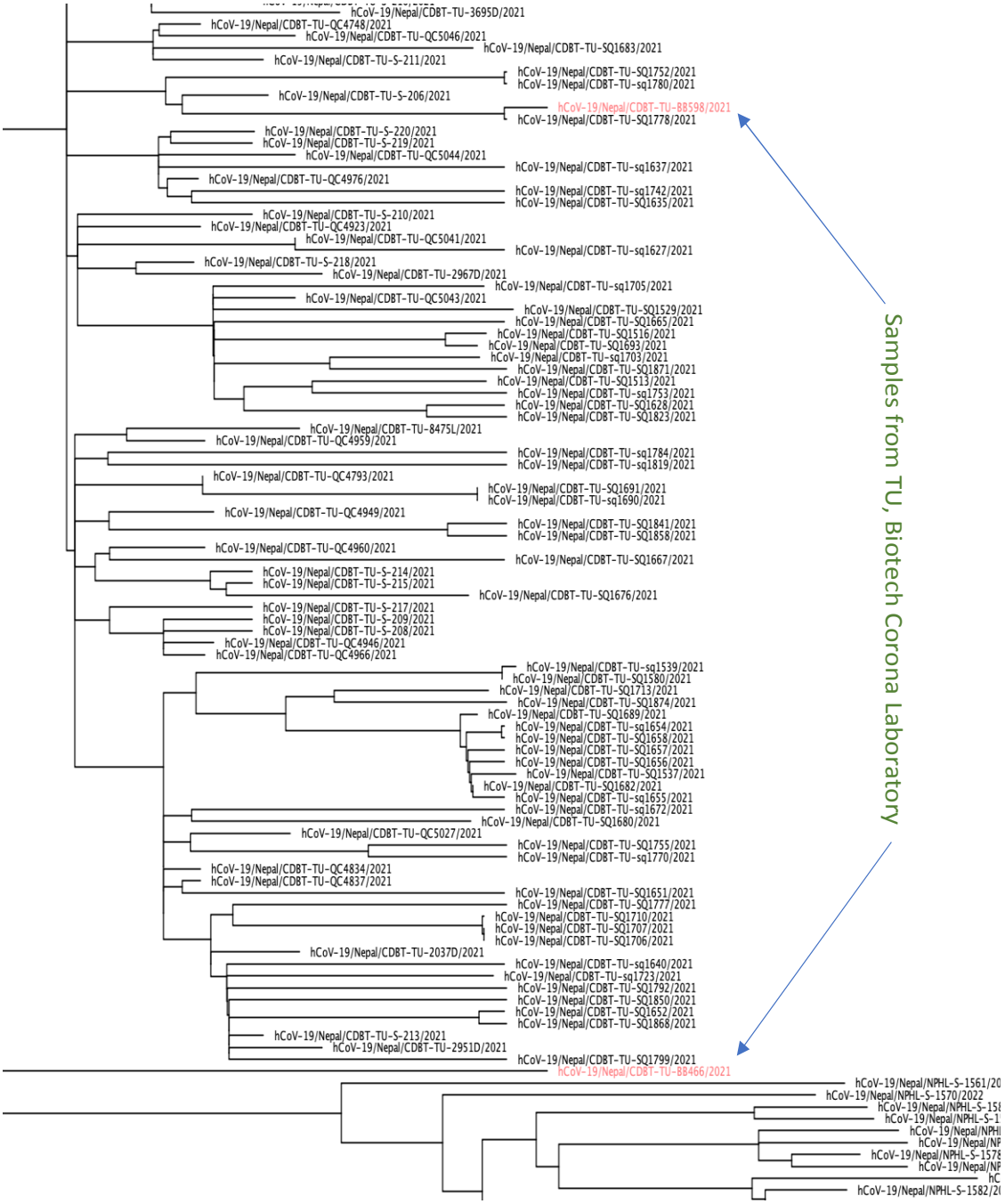


Figure 20 Phylogenetic tree showing our 2 sequenced samples BB466 and BB598. Phylogenetic tree constructed using BEAST v2.6.7 with Tamura Nei 93 model. Zoomed view of the tree showing two isolates sequenced from Kirtipur Municipality- TU Biotech Corona Laboratory where these two isolates are highlighted in red color.

4.4 Phylodynamic of SARS-CoV-2

4.4.1 Estimated Reproduction Number

Analysis through BDSKY-plot shows that the R_e estimates from our SAR-CoV-2 genomic datasets shows a complex phylodynamic pattern characterized by at least two growing phases (figure 21). The first rise in Basic reproduction number was observed in early January 2021 where median R_e reached peak value of 4.8472. This R_e value remained almost same and then declined sharply during June 2021. The second rise in R_e value was seen during January 2022 and reached maximum of 3.195. The first growth phase shows more uncertainty than the 2nd phase because it had wider 95% confidence interval of R_e estimates. Furthermore, analysis shows that the estimated time to become uninfectious interval for our dataset was around 3.6526 (95% HPD:2.1146-4.6761).

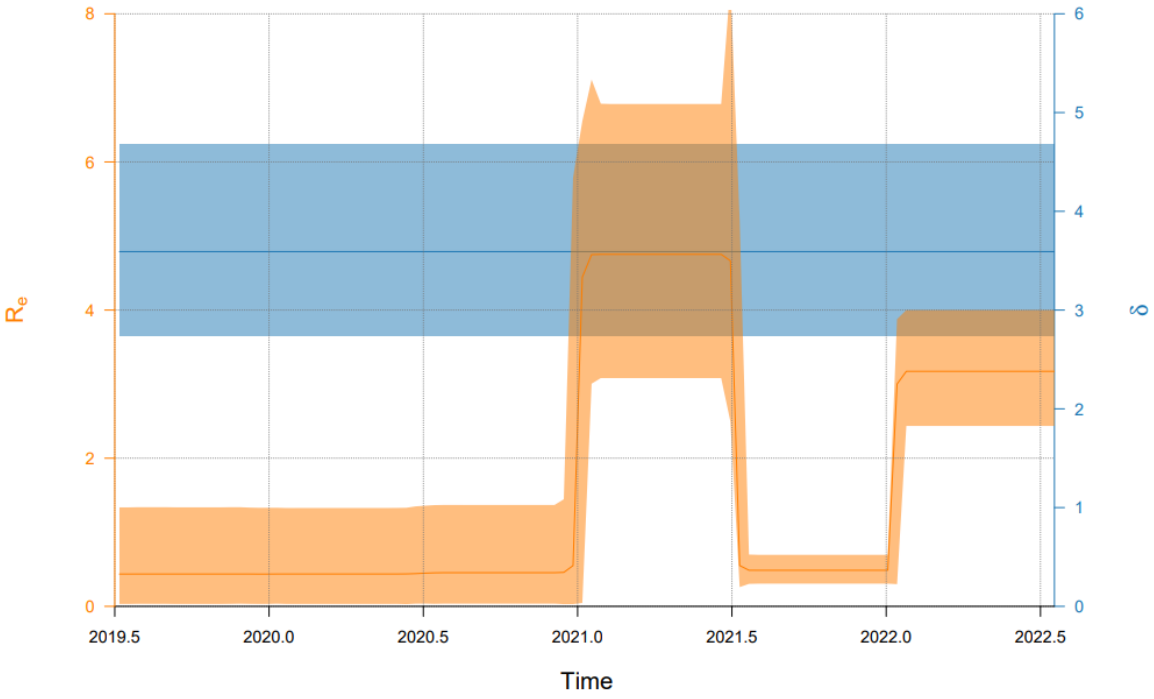


Figure 21 Estimates of the inferred R_e (orange) over time and the estimate of the becoming uninfectious rate(blue). R_e estimates obtained using Bayesian birth-death skyline model. Horizontal dotted line represents epidemiological threshold (R_e=2). Shaded area represents 95% BCI (Bayesian Confidence interval). Y-axis indicates log population and x-axis indicates the progression through time.

4.4.2 Lineage through time estimation of SARS-CoV-2

The y-axis shows the number of lineages; the x-axis shows the estimated time before present. The solid line represents the absolute value of the log of the number of ancestral lineages (different branches) present at each time interval. The shading is the 95% confidence interval for the number of lineages. The rate of accumulation of new lineages is initially high till mid 2021 and then levels off which again rises slightly during first quarter of 2022 and again levels off (figure 22). Lineage diversity estimation/analysis of SARS-CoV-2 s lineage diversity has decreased in omicron wave compared to delta wave at the start of pandemic The slope of the lineage-accumulation curve represents the net rate of diversification, which is the rate of formation of new lineages minus the rate of loss of lineages by extinction. Dan Rabosky and Irby Lovette show that this pattern can be explained only by a high initial speciation rate followed by a deceleration in the rate of diversification. This pattern is consistent with density-dependent speciation, or the early diversification of the lineage as it filled the available ecological space before being constrained by resource limitation.

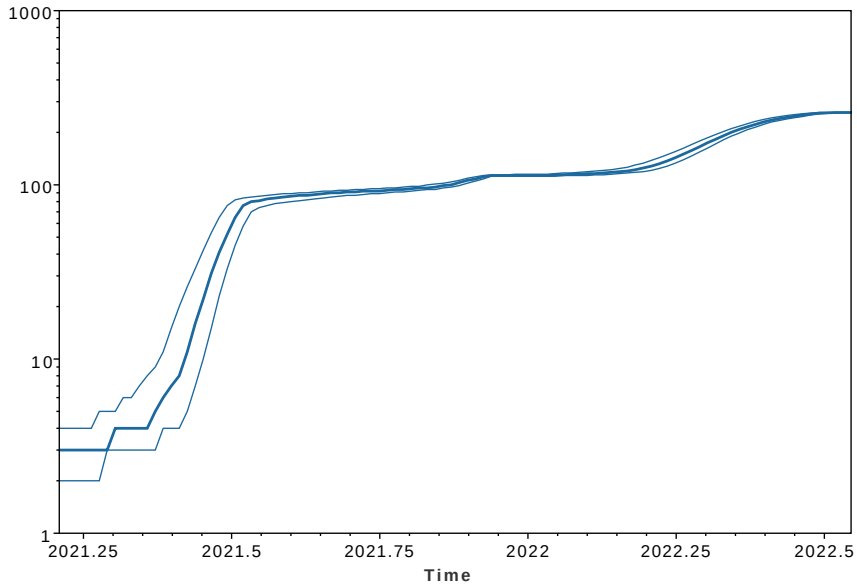


Figure 22 Lineage diversity through time estimation of SARS-CoV-2. Y-axis shows the number of lineages, the x-axis shows the estimated time before present. The solid line represents the absolute value of the log of the number of ancestral lineage present at each time interval

4.4.3 Estimation of unsampled cases of SARS-CoV-2 and study of viral Transmission chain

Using TransPhylo the total possible unreported cases present within the study community at specific study period was predicted. TransPhylo illustrates transmission within many regions through inferring various unsampled sources. Of total 1939 nodes present in obtained transmission network, 260 (13.4%) of them were sampled nodes while remaining 1679 nodes that account for (86.59%%) of the network were inferred as unsampled sources. The inferred unsampled sources indicated by brown color greatly outnumber the actual sampled sources (figure 23). This data is further supported by the graph of Figure 24 that shows very high number of inferred unsampled cases. In our study the source of infection is predicated to be an unsampled source that has been shown by black arrow in figure 23.

The transmission chain shows Kathmandu to be main cluster for infection cases of SARS-CoV-2 though it is not conclusive as unsampled nodes are overwhelming. The sampled cases are present only at the tip of the nodes with large unsampled cases which points out lack of enough testing and genomic surveillance. Additionally, we can see few events occurring outside of the main transmission network indicating that some infections occurred separately with different source of infection independent to the cases present within the main transmission chain.

Figure 23 Estimation of viral transmission chain over the study period using TransPhylo, with node colored as per the legends

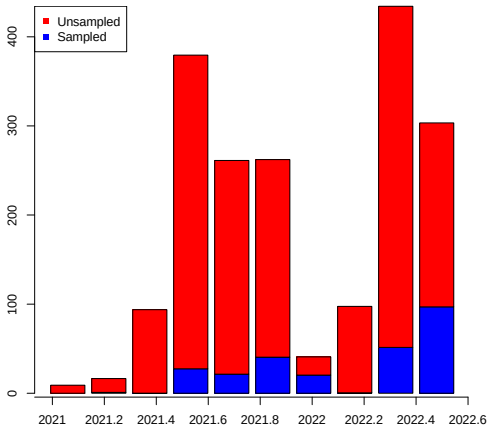


Figure 24 Outbreak plot showing the numbers of sampled and unsampled cases through time

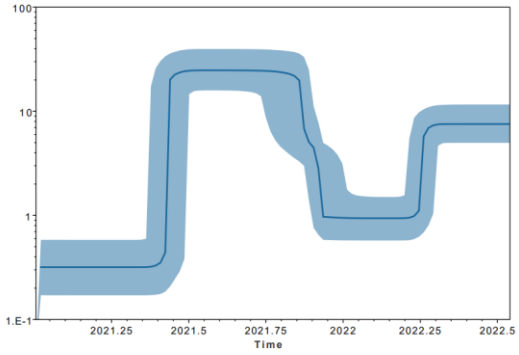


Figure 25 effective population size estimation from Bayesian Phylogenetic analysis. Solid dark line and shaded blue region respectively represent the median and 95% high posterior density interval estimates of effective population size(y-axis) over time (x-axis)

In figure 24 the number of cases is plotted over time next to their respective infection dates (x axis), and are colored according to if they were sampled (blue) or not (red). The graph shows the estimated numbers of unsampled cases with respect to the reported cases through time. Result allows us to infer that very high number of unsampled cases are contributing in the transmission. Figure 24 depicts the two major waves of an outbreak, an early peak around mid 2021 and the second rise from early 2022 each with the very high proportion of inferred unsampled cases. This figure is further supported by another figure 25 which shows the estimated effective population estimated by BEAST. In accordance to figure 24, this figure also shows that there are two major waves of infection. First wave is seen during mid 2021 and the second wave is obtained during early 2022.

Chapter V

DISCUSSION

The coronavirus disease 2019 (COVID-19) emerged as a global pandemic affecting the world health and economy. Nepal is not an exception and encountered its own sets of hurdles to prevent the spread of covid 19 infection. Nepal reported its first confirmed case of COVID-19 on Jan 23, 2020 in a student returning home from Wuhan, China. This was also the first import case reported in South-East Asia. In subsequent weeks, with the hope to combat the disease, the country established its first COVID-19 testing facility, identified dedicated hospitals to treat infected patients, setup screening at the airport, enhanced surveillance in land borders and increased public awareness campaign (Piryani et al., 2020).

Two months later, on March 23, the second import case was confirmed, by then, only 610 RT PCR test was performed. The following day, Government announced the nationwide lockdown (Sharma et al., 2021). With the beginning of April, the horrors of pandemic became apparent with the reports of first local transmission in Kailali district (Province 7), nearing Indian border. The immediate unprecedented lockdown, resulted in unmanaged outflow of large mass of people from the capital. In spite of precautionary measures embarked, within few weeks, the number of cases rise exponentially throughout the nation — from 1000 cases in May to 10,000 at the end of July 2020. The rising number of cases coincided with opening of borders for the returning migrant workers. In addition, limited quarantine centers, shortage of medical staffs, testing facilities, overcrowd, delayed test results, and improper healthcare infrastructures acted as a catalyst. The start of July seems to be a tipping point of the outbreak, as the average new case load began to decline. Although, the flattening of curve might reflect the success of lockdown, along with implementation of other nonpharmaceutical interventions (NPIs), such as use of face masks, social distancing measures; the definitive cause remains unclear. In response to fledgling economy and growing public pressure, on July 21 government lifted the lockdown with partial restrictions. The ease of restriction coincided with the festive month and thus caused a rapid surge in transmission causing the inevitable first wave of

COVID-19 that began around early August 2020 with the peak in October 2020 (Pandey et al., 2022).

Despite the rising number of COVID-19 cases and PCR tests, Nepal lacked the infrastructure for the genome sequencing of the SARS-CoV-2 virus, prevalent in its population. Till December of 2020 only one SARS-CoV-2 genome was available in GISAID from Nepal, which was the sequence being published on 14 February 2020, and was sequenced in The University of Hong Kong from a sample collected in Nepal Public Health Laboratory (NPHL) on 13 January of 2020 (GISAID, 2022). On 31 December 2021, 15 more sequences were uploaded in GISAID. Due to a shortage of appropriate sequencing equipment in the nation, samples from the initial wave were overlooked for genetic surveillance. This effect was also evident in our work, as we could only begin to sequence the samples that had been collected after April 2021.

Male patients made up 51% of the study's total participants, female patients made up 31%, and 18% of the samples had no gender information available. The age range in the samples with the highest case rate was between 31 and 40 years (18.7%), followed by 21 to 30 years (15.82%) and 51 to 60 years (14.74%). In a similar study presented by (Dawadi et al., 2022) revealed that in a gender-wise distribution of COVID 19 infection in Nepal, males (58.7%) were more impacted as compared to female (41.23%). This study also demonstrated that the age group between 31-40 years were infected the most. The cause of this might be these group belongs to most active working population and the group to socialize more in comparison to other age groups. A gender specific assessment of covid 19 impact reasons that males being more outgoing, a higher rate of smoking and role of testosterone which is immunosuppressive could result in higher incidence of infection in male (Basnet et al., 2021).

Similar to our research, a study conducted in 2022 by Dawadi et al. showed that a significantly higher cases were seen in Kathmandu valley and then the Terai region in Nepal. This may be related to the influx of people from surrounding COVID-19-affected areas as Nepal's open border with India and megacities like Kathmandu valley had higher cases due to high density of population and higher availability of resources for testing of COVID-19 infection (Piryani et al., 2020).

SARS-CoV-2 has a proofreading enzyme, yet it adapts and mutates quickly, producing new variants with improved fitness. The World Health Organization had identified a total of five VOCs by March 29th, 2022, including VOCs that were once in circulation (Alpha, Beta, and Gamma) and VOCs that are now in circulation (Delta and Omicron) (WHO, 2021b). SARS-CoV-2 genomic surveillance is crucial for preventing and tracking the emergence and spread of novel variations as well as for providing data for public health interventions (Massi et al., 2022). This study reports SARS-CoV-2 genomic surveillance using the whole genome sequences of 278 samples sequenced in CDBT, TU laboratory. Phylogenetic analysis showed that the most isolates belonged to BA.2 lineage belonging to omicron variants. In case of our study the most common variant among the sampled cases was Omicron variant (56%) followed by Delta variant (41%) and a single case of alpha variant. But it should be noted that, for our study higher number of samples were collected during later stage of infection when the omicron variant was dominant in country. Scarcity of infrastructures for sequencing led to paucity of genomic sequencing data in the country. Facilities for sequencing were available only during later stages of pandemic as a result more genomic sequences were available from samples isolated during late stages than during early stage of pandemic. Very few samples were collected for sequencing till the end of second wave in country that created difficulty in estimating the strains dominant during early stage of infections. To gain a clear picture of what was going on in nation, regular sequencing would have been quite beneficial. Such study might have kept track of which variations were spreading around the nation over time and such regular time-lapsed data could have helped in better planning of public policies.

A timeline-based study from our available data (figure 17) showed alpha variant as the early variant identified. Delta variant was introduced on late April 2021 which soon replaced alpha variant within July of 2021. The omicron variant then appeared at the December of 2021 and became dominant variant replacing older delta. Similar result was seen in case of a temporal analysis of SARS-CoV-2 variants in Nepal done by (Paudel et al., 2021). According to their data, alpha, kappa and delta strains were present in early part of 2021. Despite all three strains were present, the other two variants were outcompeted by delta, which was the only variant to appear in July of 2021. The omicron variant was discovered for the first time in Nepal on December 6, 2021, just 2 weeks after it was first

discovered in South Africa. With appearance of omicron there was increase in daily new cases with record-breaking 10,000 cases on 20 January 2022, giving rise to third wave of country which lasted till February 2022 (Pandey et al., 2022).

In an epidemiological study, one of the most crucial measures while monitoring an epidemic is R_o (i.e., the number of secondary cases caused by a single infected individual in a susceptible healthy population). In course of epidemic its value keeps on changing and hence called effective reproduction number (R_e). R_0 is typically calculated based on the rate of increase in number of cases. Thanks to recently established evolutionary models, it is now possible to predict such epidemiological parameters based on phylogenesis. The basic reproduction number (R_0), is calculated under the premise that all people who are not infected are equally vulnerable to the virus. Even while this is typically true at the start of an epidemic, susceptibility populations vary over time as certain members of the population may recover from the disease, develop an immunity to it, or become protected as a result of intervention programs. Other factors such as super-spreading events, social interventions, disease controlling strategies etc. may cause change in the reproduction number over time (Marimuthu et al., 2021b).

The analysis from our dataset showed an increase in effective reproduction number (R_e) during early 2021 from less than 1 and reached its peak to 4.8471(95% HPD: 3.08-6.78). The R_e then declined during mid 2021 to 0.495(95% HPD: 0.3-0.69) and again raised during early January 2022 to 3.1957 (95% HPD: 2.43-4). This result is supported by the data of actual SARS-CoV-2 positive count in Nepal (as shown below in figure 26) which shows that there was actual rise in daily covid-19 positive cases as predicted by R_e from our study. Data shows that there was upsurge of covid cases during early 2021 which soon start to decline after the end of mid-2021 and again there was rise in daily covid cases during early 2022 just as predicted by our study.

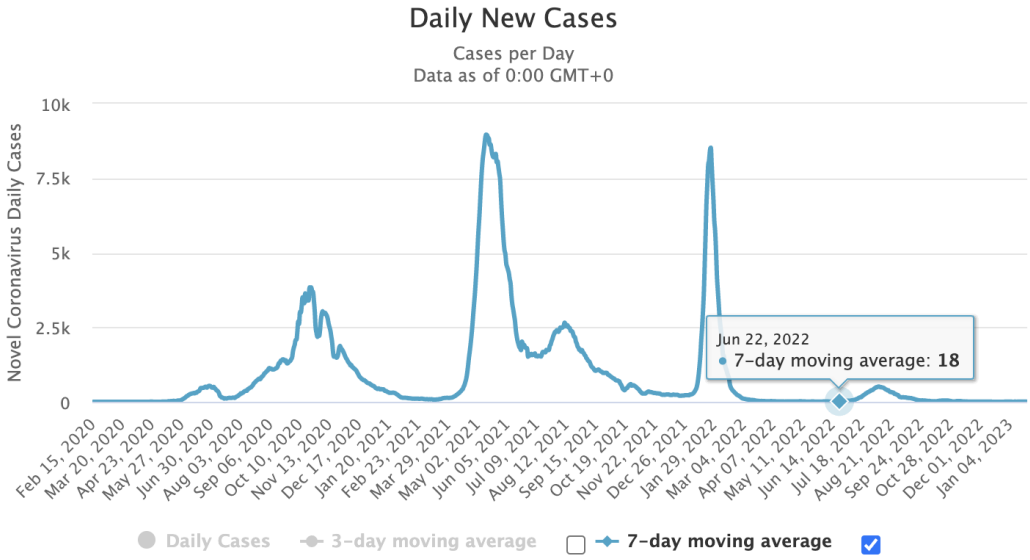


figure 26 Daily new cases count of SARS-CoV-2 in Nepal. The plot shows the graphical representation of total number of daily covid positive cases (y-axis) against the time (x-axis)

In a similar study the R_e value predicted during second wave in Nepal was 4.2 which then decreased during June 2021 due to implementation of the lockdown and again raised following the partial relaxation of lockdown (Adhikari et al., 2022). Typically, the value of Re declines during an epidemic due to reduction in the population of susceptible individuals. Meanwhile, an increase in R_e can be due to increased transmissibility of the virus, or may be due to increase in contact rates between the individual within the population (Lai et al., 2020). In an estimation of effective reproduction number of Nepal by another study showed that the effective reproduction number was consistently below unity till June 2020. Despite an early surge in COVID-19 cases, a decreasing trend of infection (R<1) during the initial phase was observed. This could be reasoned with the idea that the initial increase in cases was caused by imported cases rather than local transmission (Pantha et al., 2021). The reproduction number is based on three variables: how long an infected person can spread the disease, how likely it is for an infected person to spread the disease to a susceptible person when they come into contact, and how often people come into contact. The pathogen's ability to spread and how long it can spread are biological constants, but the amount of contact between people will vary, so R will change depending on this parameter. This shows why social distance was so important during the COVID-19 pandemic (Achaiah et al., 2020).

There are several intriguing possibilities that can arise from the capacity to infer the existence of unsampled sources of infection. Using the optimized TransPhylo for COVID-19 we were able to infer the transmission of COVID-19 in Nepal and get a fairly clearer picture of the pandemic's progression. In our study, the transmission tree estimated from our dataset using TransPhylo showed very high number of non-sampled sources of infection. Proportion of inferred unsampled sources (86.59%) of infection was overwhelmingly higher than actual sampled cases. In a comparable study (Perera et al., 2021), same approach was used to infer the transmission network in Russia and New York where TransPhylo inferred 61.06% as non-sampled source in Russia and 65.73% of events as unsampled source of infection in New York state. Furthermore, the transmission tree showed that most of the sampled events were at the tip of the transmission nodes and in some cases, some events reside outside of the main transmission network. These indicate the absence of clustering of nodes from the same region or location, suggesting that there were possibly numerous introductions through travel, lack of any super-spreader and numerous clusters of infections (Perera et al., 2021). Similarly, using TransPhylo the total number of possible cases were inferred by estimating the number of unsampled cases along with their respective reported cases through time (Figure 23). The result showed overwhelmingly high number of unsampled cases and allocate that the key transmission events may be due to unsampled cases, suggesting that the genomic screening and investigation during an outbreak were not comprehensive (Didelot et al., 2017). The plot showed two major waves of outbreak. The first peak coincides with the second covid-19 wave of Nepal which was peaked during mid-May to June of 2021 where rapid surge in infection was seen with the number of cases averaged as high as 9000 cases per day (Kharel, 2021), and the second peak resembles the third wave of Nepal that began on early 2022 (IOM, 2020).

From our study, we acknowledge that unsampled cases in our analysis may not have received a SARS-CoV-2 diagnosis or may have undergone testing and tested positive for SARS-CoV-2 but not had the virus sequenced. In these instances, cases might not be sequenced due to a lack of a clinical specimen or other sequence-able materials. In addition, cases might proceed unsampled for many reasons, like asymptomatic infections where a person might refuse to seek medical attention. Additionally, this is a typical

problem in places lacking sophisticated sequencing procedures or infrastructure, meaning that it will be impossible to sequence the full symptomatic infected population there. Our study is at its most useful in these kinds of settings. Such inferences about the size of the unsampled population, together with the known number of SARS-CoV-2-positive tests in the public, may shed light on the proportion of the general population that is asymptomatic or showing early signs of infection but has not yet been identified. Therefore, the regulating body will have a better understanding of the current epidemic.

Since we only had access to a small number of sequence data and lacked the comprehensive sequencing data, unsampled people might be underestimated as a result. However, even with that caveat, the inferences drawn were in concordance with real events. Increases in the availability of sequencing data for diverse locations will allow for more precise estimates of the number of people who went undiagnosed and showed no symptoms. Without extensive contact tracking required, such study will contribute to better epidemiological modeling and the interpretation of the effects of public health initiatives on the containment and spread of infectious diseases.

Chapter VI

CONCLUSION

This research is among the first in Nepal to apply a Bayesian Phylodynamic pipeline to analyze the epidemiology and evolutionary characteristics of SARS-CoV-2 viruses that are prevalent in the region. We made an effort to characterize the demographic condition, evolutionary dynamics, infection dynamics and epidemic transmission patterns, as well as other phylodynamic inferences, based on the whole genome sequences of the isolates sequenced in our lab. We were able to estimate the reproduction number of SARS-CoV-2 infection and was found to be similar to the outcomes obtained from other conventional epidemiological methods. We were also able to reconstruct the inferred transmission pattern and picture the infection dynamics of population despite the incomplete sampling for sequencing. Our study highlights the importance of genomic surveillance programs to ensure impactful decision-making programs for intervention aimed at most relevant variants. Study also emphasizes the importance of implementing the phylogenetic and phylodynamic approaches for monitoring and surveillance of emerging infection and to shed light on the roles of multiple initiatives implemented from limiting the spread of virus infection. Precise understanding of the epidemic dynamics of such viral outbreaks in real time is absolutely crucial for guiding effective prevention efforts.

LIMITATIONS OF THE STUDY

1. Samples were collected only after April 2021 where first wave of infection was subsided already.
2. Sampling is not unbiased as most of the samples were from Kathmandu valley as they were easily accessible ones.
3. Only limited number of samples were included in this study and no comprehensively distributed samples were available.

RECOMMENDATIONS / FUTURE PERSPECTIVES

1. It is recommended that clinical and epidemiological data be obtained for a more comprehensive and prolific analysis of the virus's importation and circulation in the country.
2. It is suggested that more diverse samples be obtained in terms of geographical location, demography, and clinical characteristics.
3. Our findings highlight the importance of sequencing SARS-CoV-2 genomes in conjunction with clinical history in terms of recovery, hospitalization, and co-morbidity in order to identify actionable variants that are also relevant for prognosis and epidemiological understanding of the virus and disease.

REFERENCES

Achaiah, N. C., Subbarajasetty, S. B., & Shetty, R. M. (2020). R0 and re of covid-19: Can we predict when the pandemic outbreak will be contained? *Indian Journal of Critical Care Medicine*, *24*(11), 1125–1127. https://doi.org/10.5005/jp-journals-10071-23649

Adhikari, K., Gautam, R., Pokharel, A., Dhimal, M., Uprety, K. N., & Vaidya, N. K. (2022). Insight into Delta variant dominated second wave of COVID-19 in Nepal. *Epidemics*, *41*(December 2021), 100642. https://doi.org/10.1016/j.epidem.2022.100642

Arvestad, L. (2018). Alv: a Console-Based Viewer for Molecular Sequence Alignments. *Journal of Open Source Software*, *3*(31), 955. https://doi.org/10.21105/joss.00955

Attwood, S. W., Hill, S. C., Aanensen, D. M., Connor, T. R., & Pybus, O. G. (2022). Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews*, *23*. https://doi.org/10.1038/s41576-022-00483-8

Barido-Sottani, J., Bošková, V., Plessis, L. Du, Kühnert, D., Magnus, C., Mitov, V., Müller, N. F., Pečerska, J., Rasmussen, D. A., Zhang, C., Drummond, A. J., Heath, T. A., Pybus, O. G., Vaughan, T. G., & Stadler, T. (2018). Taming the BEAST - A Community Teaching Material Resource for BEAST 2. *Systematic Biology*, *67*(1), 170–174. https://doi.org/10.1093/sysbio/syx060

Basnet, P. S., Sharma, D., Singh Karki, S., Karki, S., Bhandari, H. L., & Shrestha, D. (2021). Clinico-Demographic Profile of Covid-19 Patients Admitted in COVID- HDU and its Association with Conjunctivitis. *Nepalese Medical Journal*, *4*(2), 482–484. https://doi.org/10.3126/nmj.v4i2.41492

Bastola, A., Sah, R., Rodriguez-Morales, A. J., Lal, B. K., Jha, R., Ojha, H. C., Shrestha, B., Chu, D. K., Poon, L. L., Costello, A., Morita, K., & Pandey, B. D. (2020). The first 2019 novel coronavirus case in Nepal. *The Lancet Infectious Diseases*, 20(3),279–280. https://doi.org/10.1016/s1473-3099(20)30067-0.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, *10*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4), 1–28. https://doi.org/10.1371/journal.pcbi.1006650

Brito, A. F., Semenova, E., Dudas, G., Hassler, G. W., Kalinich, C. C., Kraemer, M. U. G., Hill, S. C., Danish Covid-19 Genome Consortium, Sabino, E. C., Pybus, O. G., Dye, C., Bhatt, S., Flaxamn, S., Suchard, M. A., Grubaugh, N. D., Baele, G., & Faria, N. R. (2021). Global disparities in SARS-CoV-2 genomic surveillance. *MedRxiv : The Preprint Server for Health Sciences*, 1–24. https://doi.org/10.1101/2021.08.21.21262393

CDC. (2022). What is Genomic Surveillance? *Cdc.Gov*, 22–24. https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html

Cevik, M., Bamford, C., & Ho, A. (2020). COVID-19 pandemic-a focused review for clinicians. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 26(7), 842–847. https://doi.org/10.1016/j.cmi.2020.04.023.

Chan-Yeung, M., & Xu, R. H. (2003). SARS: epidemiology. Respirology (Carlton, Vic.), 8 *Suppl(Suppl 1)*, S9–S14. https://doi.org/10.1046/j.1440-1843.2003.00518.x.

Chiara, M., D'Erchia, A. M., Gissi, C., Manzari, C., Parisi, A., Resta, N., Zambelli, F., Picardi, E., Pavesi, G., Horner, D. S., & Pesole, G. (2021). Next generation sequencing of SARS-CoV-2 genomes: Challenges, applications and opportunities. *Briefings in Bioinformatics*, *22*(2), 616–630. https://doi.org/10.1093/bib/bbaa297

Clark, D. P., Pazdernik, N. J., & McGehee, M. R. (2019). DNA Sequencing. *Molecular Biology*, 240–269. https://doi.org/10.1016/B978-0-12-813288-3.00008-2

Dawadi, P., Syangtan, G., Lama, B., Kanel, S. R., Raj Joshi, D., Pokhrel, L. R., Adhikari, R., Joshi, H. R., & Pavel, I. (2022). Understanding COVID-19 Situation in Nepal and Implications for SARS-CoV-2 Transmission and Management. *Environmental Health Insights*, *16*. https://doi.org/10.1177/11786302221104348

DEGROOT RJ, LUYTJES W, HORZINEK MC, VANDERZEIJST BAM, SPAAN WJM, & LENSTRA JA. (1987). Evidence for a Coiled-Coil Structure in the Spike Proteins of Coronaviruses. *Journal of Molecular Biology*, *196*(4), 963–966. http://apps.isiknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=6&SID=2CCd8cPAGIPDodJLBpD&page=1&doc=1

Deng, Y., Liu, W., Liu, K., Fang, Y. Y., Shang, J., Zhou, L., Wang, K., Leng, F., Wei, S., Chen, L., & Liu, H. G. (2020). Clinical characteristics of fatal and recovered cases of coronavirus disease 2019 in Wuhan, China: a retrospective study. *Chinese medical journal,* 133(11), 1261–1267. https://doi.org/10.1097/CM9.0000000000000824.

Dhama, K., Khan, S., Tiwari, R., Sircar, S., Bhat, S., Malik, Y. S., Singh, K. P., Chaicumpa, W., Bonilla-Aldana, D. K., & Rodriguez-Morales, A. J. (2020). Coronavirus Disease 2019-COVID-19. *Clinical microbiology reviews,* 33(4), e00028-20. https://doi.org/10.1128/CMR.00028-20.

Didelot, X., Fraser, C., Gardy, J., Colijn, C., & Malik, H. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, *34*(4), 997–1007. https://doi.org/10.1093/molbev/msw275

Didelot, X., Kendall, M., Xu, Y., White, P. J., & McCarthy, N. (2021). Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. *Current Protocols*, *1*(2), 1–23. https://doi.org/10.1002/cpz1.60

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, *2*(1), 23–41. https://doi.org/10.1177/096228029300200103

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral quasispecies evolution. *Microbiology and molecular biology reviews: MMBR,* 76(2), 159–216. https://doi.org/10.1128/MMBR.05023-11

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Fehr, A. R., & Perlman, S. (2015). Coronaviruses: An overview of their replication and pathogenesis. *Coronaviruses: Methods and Protocols*, 1–23. https://doi.org/10.1007/978-1-4939-2438-7_1

Frost, S. D. W., & Volz, E. M. (2010). Viral phylodynamics and the search for an effective number of infections. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1548), 1879–1890. https://doi.org/10.1098/rstb.2010.0060

Fuk-Woo Chan, J., Kok, K.-H., Zhu, Z., Chu, H., Kai-Wang To, K., Yuan, S., & Yuen, K.-Y. (2020). *Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan*. https://doi.org/10.1080/22221751.2020.1737364

GISAID, 2022. https://www.epicov.org/epi3/frontend#1557d3

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., & Suchard, M. A. (2013).

Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*, *30*(3), 713–724. https://doi.org/10.1093/molbev/mss265

Gorbalenya, A. E. (2008). Phylogeny of Viruses. *Encyclopedia of Virology*, *2005*, 125–129. https://doi.org/10.1016/B978-012374410-4.00712-3

Gralinski, L. E., & Menachery, V. D. (2020). Return of the Coronavirus: 2019-nCoV. *Viruses*, 12(2), 135. https://doi.org/10.3390/v12020135.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., & Holmes, E. C. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, *303*(5656), 327–332. https://doi.org/10.1126/science.1090727

Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., Butt, K. M., Wong, K. L., Chan, K. W., Lim, W., Shortridge, K. F., Yuen, K. Y., Peiris, J. S., & Poon, L. L. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science (New York, N.Y.)*, 302(5643), 276–278. https://doi.org/10.1126/science.1087139.

Heled, J., & Drummond, A. J. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, *27*(3), 570–580. https://doi.org/10.1093/molbev/msp274

Hellewell, J., Abbott, S., Gimma, A., Bosse, N. I., Jarvis, C. I., Russell, T. W., Munday, J. D., Kucharski, A. J., Edmunds, W. J., Sun, F., Flasche, S., Quilty, B. J., Davies, N., Liu, Y., Clifford, S., Klepac, P., Jit, M., Diamond, C., Gibbs, H., … Eggo, R. M. (2020). Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *The Lancet Global Health*, *8*(4), e488–e496. https://doi.org/10.1016/S2214-109X(20)30074-7

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Dail Laughinghouse, H., McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., … Cranston, K. A. (2015). Synthesis of phylogeny and taxonomy into a

comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(41), 12764–12769. https://doi.org/10.1073/pnas.1423041112

Hohna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, *65*(4), 726–736. https://doi.org/10.1093/sysbio/syw021

Holmes, E. C., Goldstein, S. A., Rasmussen, A. L., Robertson, D. L., Crits-Christoph, A., Wertheim, J. O., Anthony, S. J., Barclay, W. S., Boni, M. F., Doherty, P. C., Farrar, J., Geoghegan, J. L., Jiang, X., Leibowitz, J. L., Neil, S., Skern, T., Weiss, S. R., Worobey, M., Andersen, K. G., Garry, R. F., … Rambaut, A. (2021). The origins of SARS-CoV-2: *A critical review. Cell,* 184(19), 4848–4856. https://doi.org/10.1016/j.cell.2021.08.017.

Houldcroft, C. J., Beale, M. A., & Breuer, J. (2017). Clinical and biological insights from viral genome sequencing. *Nature Reviews Microbiology*, *15*(3), 183–192. https://doi.org/10.1038/nrmicro.2016.182

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Ingle, D. J., Howden, B. P., & Duchene, S. (2021). Development of Phylodynamic Methods for Bacterial Pathogens. *Trends in Microbiology*, *29*(9), 788–797. https://doi.org/10.1016/j.tim.2021.02.008

IOM. (2020). *Rapid Assessment on Impacts of COVID-19 on Returnee Migrants and Responses of Local Governments of Nepal*.

Jasper Fuk-Woo Chan, Kin-Hang Kok, Zheng Zhu, Hin Chu, Kelvin Kai-Wang To, Shuofeng Yuan, & Kwok-Yung Yuen. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerging Microbes and Infections*, *9*(1), 540.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid

multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2014). MAFFT: Iterative refinement and additional methods. *Methods in Molecular Biology*, *1079*, 131–146. https://doi.org/10.1007/978-1-62703-646-7_8

Khan MI, Khan ZA, Baig MH, Ahmad I, Farouk AE, et al. (2020). Comparative genome analysis of novel coronavirus (SARS-CoV-2) from different geographical locations and the effect of mutations on major target proteins: An in silico insight. *PLOS ONE* 15(9): e0238344. https://doi.org/10.1371/journal.pone.0238344.

Kharel, P. (2021). *Nepal ' s fight against the second wave of COVID-19 pandemic*. *2*, 1–8.

Klausegger, A., Strobl, B., Regl, G., Kaser, A., Luytjes, W., & Vlasak, R. (1999). Identification of a Coronavirus Hemagglutinin-Esterase with a Substrate Specificity Different from Those of Influenza C Virus and Bovine Coronavirus. *Journal of Virology*, *73*(5), 3737–3743. https://doi.org/10.1128/jvi.73.5.3737-3743.1999

Kretzschmar, M. (2016). Measurement and Modeling: Infectious Disease Modeling. In *International Encyclopedia of Public Health* (Second Edi, Vol. 4). Elsevier. https://doi.org/10.1016/B978-0-12-803678-5.00229-0

Kumar, S., Maurya, V. K., Prasad, A. K., Bhatt, M. L. B., & Saxena, S. K. (2020). Structural, glycosylation and antigenic variation between 2019 novel coronavirus (2019-nCoV) and SARS coronavirus (SARS-CoV). *VirusDisease*, *31*(1), 13–21. https://doi.org/10.1007/s13337-020-00571-5

Kumar, S., Nyodu, R., Maurya, V. K., & Saxena, S. K. (2020). *Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)*. *2*, 23–31. https://doi.org/10.1007/978-981-15-4814-7_3

Lai, A., Bergna, A., Acciarri, C., Galli, M., & Zehender, G. (2020). Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *Journal of Medical Virology*, *92*(6), 675–679. https://doi.org/10.1002/jmv.25723

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of internal medicine,* 172(9), 577–582. https://doi.org/10.7326/M20-0504.

Lee, J., Chowell, G., & Jung, E. (2016). A dynamic compartmental model for the Middle East respiratory syndrome outbreak in the Republic of Korea: A retrospective analysis on control interventions and superspreading events. J*ournal of theoretical biology,* 408, 118–126. https://doi.org/10.1016/j.jtbi.2016.08.009.

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annual Review of Virology*, *3*, 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Li, G., Fan, Y., Lai, Y., Han, T., Li, Z., Zhou, P., Pan, P., Wang, W., Hu, D., Liu, X., Zhang, Q., & Wu, J. (2020). Coronavirus infections and immune responses. *Journal of Medical Virology*, *92*(4), 424–432. https://doi.org/10.1002/jmv.25685

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, *9*, 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

Marimuthu, S., Joy, M., Malavika, B., Nadaraj, A., Asirvatham, E. S., & Jeyaseelan, L. (2021a). Modelling of reproduction number for COVID-19 in India and high incidence states. *Clinical Epidemiology and Global Health*, *9*(May 2020), 57–61. https://doi.org/10.1016/j.cegh.2020.06.012

Marimuthu, S., Joy, M., Malavika, B., Nadaraj, A., Asirvatham, E. S., & Jeyaseelan, L. (2021b). Modelling of reproduction number for COVID-19 in India and high incidence states. *Clinical Epidemiology and Global Health*, *9*(June 2020), 57–61. https://doi.org/10.1016/j.cegh.2020.06.012

Massi, M. N., Abidin, R. S., Farouk, A. E. A., Halik, H., Soraya, G. V., Hidayah, N., Sjahril, R., Handayani, I., Hakim, M. S., Gazali, F. M., Setiawaty, V., & Wibawa, T. (2022). Full-genome sequencing and mutation analysis of SARS-CoV-2 isolated from Makassar, South Sulawesi, Indonesia. *PeerJ*, *10*, 1–18. https://doi.org/10.7717/peerj.13522

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., Lanfear, R., & Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Minin, V. N., Bloomquist, E. W., & Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, *25*(7), 1459–1471. https://doi.org/10.1093/molbev/msn090

Mohamed, E. M., Mousa, H. M., & keshk, A. E. (2018). Comparative Analysis of Multiple Sequence Alignment Tools. *International Journal of Information Technology and Computer Science*, *10*(8), 24–30. https://doi.org/10.5815/ijitcs.2018.08.04

Neuman, B. W., Kiss, G., Kunding, A. H., Bhella, D., Baksh, M. F., Connelly, S., Droese, B., Klaus, J. P., Makino, S., Sawicki, S. G., Siddell, S. G., Stamou, D. G., Wilson, I. A., Kuhn, P., & Buchmeier, M. J. (2011). A structural analysis of M protein in coronavirus assembly and morphology. *Journal of Structural Biology*, *174*(1), 11–22. https://doi.org/10.1016/j.jsb.2010.11.021

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nuin, P. A. S., Wang, Z., & Tillier, E. R. M. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, *7*, 1–18. https://doi.org/10.1186/1471-2105-7-471

O'Toole, Á., Pybus, O. G., Abram, M. E., Kelly, E. J., & Rambaut, A. (2022). Pango lineage

designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics*, *23*(1), 1–13. https://doi.org/10.1186/s12864-022-08358-2

Pandey, B. D., Ngwe Tun, M. M., Pandey, K., Dumre, S. P., Nwe, K. M., Shah, Y., Culleton, R., Takamatsu, Y., Costello, A., & Morita, K. (2022). How an Outbreak of COVID-19 Circulated Widely in Nepal: A Chronological Analysis of the National Response to an Unprecedented Pandemic. *Life (Basel, Switzerland),* 12(7), 1087. https://doi.org/10.3390/life12071087

Pantha, B., Acharya, S., Joshi, H. R., & Vaidya, N. K. (2021). Inter-provincial disparity of COVID-19 transmission and control in Nepal. *Scientific Reports*, *11*(1), 1–16. https://doi.org/10.1038/s41598-021-92253-5

Paudel, S., Dahal, A., & Bhattarai, H. K. (2021). Temporal Analysis of SARS-CoV-2 Variants during the COVID-19 Pandemic in Nepal. *Covid*, *1*(2), 423–434. https://doi.org/10.3390/covid1020036

Peiris, J. S., Guan, Y., & Yuen, K. Y. (2004). Severe acute respiratory syndrome. *Nature medicine*, 10(12 Suppl), S88–S97. https://doi.org/10.1038/nm1143.

Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Guan, Y., Yam, L. Y. C., Lim, W., Nicholls, J., Yee, W. K. S., Yan, W. W., Cheung, M. T., Cheng, V. C. C., Chan, K. H., Tsang, D. N. C., Yung, R. W. H., Ng, T. K., & Yuen, K. Y. (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet*, *361*(9366), 1319–1325. https://doi.org/10.1016/S0140-6736(03)13077-2

Perera, D., Perks, B., Potemkin, M., Liu, A., Gordon, P. M. K., Gill, M. J., Long, Q., & van Marle, G. (2021). Reconstructing SARS-CoV-2 infection dynamics through the phylogenetic inference of unsampled sources of infection. *PLoS ONE*, *16*(12 December), 1–16. https://doi.org/10.1371/journal.pone.0261422

Piryani, R. M., Piryani, R. M., Piryani, S., & Shah, J. N. (2020). *Nepal ' s response to contain COVID-19 Infection Nepal ' s Response to Contain COVID-19 Infection*. *18*, 128–134.

Poon, L., & Peiris, M. (2020). Emergence of a novel human coronavirus threatening human health. *Nature medicine,* 26(3), 317–319. https://doi.org/10.1038/s41591-020-0796-5.

Pun, S. B., Mandal, S., Bhandari, L., Jha, S., Rajbhandari, S., Mishra, A. K., Sharma Chalise, B., & Shah, R. (2020). Understanding covid-19 in Nepal. Journal of Nepal Health Research Council, 18(1), 126–127. https://doi.org/10.33314/jnhrc.v18i1.2629. Ravi, R. K., Walton, K., & Khosroheidari, M. (2018). Miseq: A next generation sequencing platform for genomic analysis. *Methods in Molecular Biology*, *1706*, 223–232. https://doi.org/10.1007/978-1-4939-7471-9_12

Rambaut,A. 2018. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.

Riou, J., & Althaus, C. L. (2020). Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin,* 25(4), 2000058. https://doi.org/10.2807/1560-7917.ES.2020.25.4.2000058.

Rodriguez-Morales, A. J., Bonilla-Aldana, D. K., Balbin-Ramon, G. J., Rabaan, A. A., Sah, R., Paniz-Mondolfi, A., Pagliano, P., & Esposito, S. (2020). History is repeating itself: Probable zoonotic spillover as the cause of the 2019 novel Coronavirus Epidemic. *Le infezioni in medicina,* 28(1), 3–5.

Roy, S., Laframboise, W. A., Nikiforov, Y. E., Nikiforova, M. N., Routbort, M. J., Pfeifer, J., Nagarajan, R., Carter, A. B., & Pantanowitz, L. (2016). *Next-Generation Sequencing Informatics Challenges and Strategies for Implementation in a Clinical Environment*. https://doi.org/10.5858/arpa.2015-0507-RA

Sawicki, S. G., Sawicki, D. L., & Siddell, S. G. (2007). A Contemporary View of Coronavirus Transcription. *Journal of Virology*, *81*(1), 20–29. https://doi.org/10.1128/jvi.01358-06

Schoeman, D., & Fielding, B. C. (2019). Coronavirus envelope protein: current knowledge. *Virology journal,* 16(1), 69. https://doi.org/10.1186/s12985-019-

1182-0

Sharma, K., Banstola, A., & Parajuli, R. R. (2021). Assessment of COVID-19 Pandemic in Nepal: A Lockdown Scenario Analysis. *Frontiers in Public Health*, *9*(April), 1–12. https://doi.org/10.3389/fpubh.2021.599280

Sheikh, A., Al-Taher, A., Al-Nazawi, M., Al-Mubarak, A. I., & Kandeel, M. (2020). Analysis of preferred codon usage in the coronavirus N genes and their implications for genome evolution and vaccine design. *Journal of virological methods,* 277, 113806.

Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D., & Schürch, A. C. (2014). Assembly of viral genomes from metagenomes. *Frontiers in microbiology,* 5, 714. https://doi.org/10.3389/fmicb.2014.00714

Sola, I., Almazán, F., & Enjuanes, L. (2015). *Continuous and Discontinuous RNA Synthesis in Coronaviruses*. https://doi.org/10.1146/annurev-virology-100114-055218

Stadler, T., Kouyos, R., VonWy, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., Günthard, H. F., Drummond, A. J., & Bonhoeffer, S. (2012). Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, *29*(1), 347–357. https://doi.org/10.1093/molbev/msr217

Tay, M. Z., Poh, C. M., Rénia, L., MacAry, P. A., & Ng, L. (2020). The trinity of COVID-19: immunity, inflammation and intervention. *Nature reviews. Immunology,* 20(6), 363–374. https://doi.org/10.1038/s41577-020-0311-8.

The Lancet. (2021). Genomic sequencing in pandemics. *The Lancet*, *397*(10273), 445. https://doi.org/10.1016/S0140-6736(21)00257-9

Theys, K., Lemey, P., Vandamme, A. M., & Baele, G. (2019). Advances in Visualization Tools for Phylogenomic and Phylodynamic Studies of Viral Diseases. *Frontiers in Public Health*, *7*(August), 1–18. https://doi.org/10.3389/fpubh.2019.00208

van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., Tamin, A., Harcourt, J. L., Thornburg, N. J., Gerber, S. I., Lloyd-

Smith, J. O., de Wit, E., & Munster, V. J. (2020). Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *The New England journal of medicine,* 382(16), 1564–1567. https://doi.org/10.1056/NEJMc2004973.

WHO Director, World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19. 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Accessed 07 April 2022.

WHO Nepal, (2021). Situation Update #89 - Coronavirus Disease 2019 (COVID-19) WHO Country Office for Nepal Reporting Date: 20 - 26 December 2021 (EPI Week 51). https://reliefweb.int/report/nepal/situation-update-89-coronavirus-disease-2019-covid-19-who-country-office-nepal

WHO, WHO Coronavirus (COVID-19) Dashboard, 1 April 2022. https://covid19.who.int/. Accessed 1 April 2022.

WHO. (2021a). *Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health*. https://www.who.int/publications/i/item/9789240018440

WHO. (2021b). *Tracking SARS-CoV-2 variants*. https://www.who.int/activities/tracking-SARS-CoV-2-variants

Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., Sheng, J., Quan, L., Xia, Z., Tan, W., Cheng, G., & Jiang, T. (2020). Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host and Microbe*, *27*(3), 325–328. https://doi.org/10.1016/j.chom.2020.02.001

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C., & Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature,* 579(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3.

You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., Pang, C. H., Zhang, Y., Chen, Z., & Zhou, X. H. (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal of Hygiene and Environmental Health*, *228*(April), 113555. https://doi.org/10.1016/j.ijheh.2020.113555

Xiao, X., Newman, C., Buesching, C. D., Macdonald, D. W., & Zhou, Z. M. (2021). Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. *Scientific reports,* 11(1), 11898. https://doi.org/10.1038/s41598-021-91470-2.

Xu, R. H., He, J. F., Evans, M. R., Peng, G. W., Field, H. E., Yu, D. W., Lee, C. K., Luo, H. M., Lin, W. S., Lin, P., Li, L. H., Liang, W. J., Lin, J. Y., & Schnur, A. (2004). Epidemiologic clues to SARS origin in China. *Emerging infectious diseases,* 10(6), 1030–1037. https://doi.org/10.3201/eid1006.030852.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). Brief Report: A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England Journal of Medicine*, *382*(8), 727. https://doi.org/10.1056/NEJMOA2001017

APPENDICES

Appendix 1a: Protocol of COVIDSEQ Library Preparation:

A. Anneal RNA:

RNA is annealed using random hexamers to prepare for cDNA synthesis.

Reagents:

- EPH3 HT (Elution Prime Fragment 3HC Mix)
  (Vortex before use) Store at -20°C, Thaw at Room temperature

Procedure:

- Label a new PCR plate/strip CDNA1
- Add 4.25µl EPH3 HT to each well
- Add 4.25µl eluted RNA sample to each well.
- Seal and shake at 1600 rpm for 1 minute.
- Centrifuge at 1000 × g for 1 minute.
- Place on the preprogrammed thermal cycler and run the COVIDSeq Anneal program

  Program:
  - Choose the preheat lid option
  - Set the reaction volume to 8.5 µl
  - 65°C for 3 minutes
  - Hold at 4°C

B. Synthesize first Strand cDNA:

Reverse transcription of RNA fragments primed with random hexamers into first strand cDNA.

Reagents:

- FSM HT (First Strand Mix HT)
  (Thaw and bring to RT, invert to mix, then keep on ice.)
- RVT HT (Reverse Transcriptase HT)
  (Invert to mic before use, keep on ice)

Procedure:

- In a 1.7 ml tube, combine the following volumes to prepare First Strand cDNA Master Mix. Multiply each volume by the number of samples.
  - FSM HT (4.5 µl)

- RVT HT (0.5 µl)

- Add 4 µl master mix to each well of the CDNA1 plate/strip.
- Seal and shake at 1600 rpm for 1 minute.
- Centrifuge at 1000 × g for 1 minute.
- Place on the preprogrammed thermal cycler and run the COVIDSeq FSS program.

  Program:
  - Choose the preheat lid option
  - Set the reaction volume to 12.5 µl
  - 25°C for 5 minutes
  - 50°C for 10 minutes
  - 80°C for 5 minutes
  - Hold at 4°C

C. <u>Amplify cDNA:</u>

Reagents:

- IPM HT (Illumina PCR Mix HT)
  (Thaw at room temperature. Keep on ice until use)
- CPP1 HT (COVIDSeq Primer Pool 1 HT)
  (Thaw at room temperature. Keep on ice until use)
- CPP2 HT (COVIDSeq Primer Pool 2 HT)
  (Thaw at room temperature. Keep on ice until use)
- Nuclease-free water

Procedure:

- Label two new PCR plates/strips COV1 and COV2.
- In a 15 ml tube, combine the following volumes to prepare COVIDSeq PCR 1 Master Mix and COVIDSeq PCR 2 Master Mix. Multiply each volume by the number of samples.
- COVIDSeq PCR 1 Master Mix:
  - IPM – 7.5 µl
  - CPP1 – 2.15 µl
  - NFW – 2.35 µl

- COVIDSeq PCR 2 Master Mix:
  - IPM – 7.5 µl
  - CPP2 – 2.15 µl
  - NFW – 2.35 µl
- Add 10 µl COVIDSeq PCR 1 Master Mix to each well of the COV1 plate/strip corresponding to each well of the CDNA1 plate/strip.

- Add 2.5 µl first strand cDNA synthesis from each well of the CDNA1 plate/strip to the corresponding well of the COV1 plate/strip.
- Similarly, add 10 µl COVIDSeq PCR 2 Master Mix to each well of the COV2 plate/strip corresponding to each well of the CDNA1 plate/strip.
- Add 2.5 µl first strand cDNA synthesis from each well of the CDNA1 plate/strip to the corresponding well of the COV2 plate/strip.
- Seal and shake at 1600 rpm for 1 minute.
- Centrifuge at 1000 x g for 1 minute.
- Place in the preprogrammed thermal cycler and run the COVIDSeq PCR program.

  Program:
  - Choose the preheat lid option
  - Set the reaction volume to 12.5 µl
  - 98°C for 3 minutes
  - 35 cycles of:
    - 98°C for 15 seconds
    - 63°C for 5 minutes
  - Hold at 4°C

SAFE POINT: seal the plate and store at -20 for up to 3 days.

D. <u>Tagment PCR Amplicons:</u>

Reagents:

- EBLTS HT (Enrichment BLT HT)
  (Store EBLTS HT upright at temperatures above 2°C. Make sure beads are always submerged in the buffer, vortex before use)
- TB1 HT (Tagmentation Buffer 1 HT)
  (Bring to room temperature. Vortex thoroughly before use)
- Nuclease-free water

Procedure:

- Label a new PCR/strip plate TAG1.
- Combine COV1 and COV2 as follows.
  a. Transfer 5µl from each well of the COV1 plate/strip to the corresponding well of the TAG1 plate/strip.
  b. Transfer 5µl from each well of the COV2 plate/strip to each well of the TAG1 plate/strip containing COV1.
- In a 15 ml tube, combine the following volumes to prepare Tagmentation Master Mix. Multiply each volume by the number of samples.
  - u TB1 HT (6 µl)

- u EBLTS HT (2 µl)
- u Nuclease-free water (10 µl)

- Add 15 µl master mix to each well in TAG1 plate/strip.
- Seal and shake at 1600 rpm for 1 minute (NOTE: No centrifugation)
- Place on the preprogrammed thermal cycler and run the COVIDSeq TAG program.

  Program:
  - Choose the preheat lid option
  - Set the reaction volume to 25 µl
  - 55°C for 5 minutes
  - Hold at 10°C

E. <u>Post Tagmentation Clean Up</u>

washes the adapter-tagged amplicons before PCR amplification.

Reagents:

- ST2 HT (Stop Tagment Buffer 2 HT)
  (Dispense ST2 HT and TWB HT slowly to minimize foaming, vortex before use)
- TWB HT (Tagmentation Wash Buffer HT)
  (Dispense TWB HT directly onto beads, vortex before use)

Procedure:

- Centrifuge the TAG1 plate/strip at 500 x g for 1 minute.
- Add 5 µl ST2 HT to each well of the TAG1 plate/strip.
- Seal and shake at 1600 rpm for 1 minute.
- Incubate at room temperature for 5 minutes.
- Centrifuge at 500 × g for 1 minute.
- Place on the magnetic stand and wait until the liquid is clear (~3 minutes). [Inspect for bubbles on the seal. If present, centrifuge at 500 x g for 1 minute, and then place on the magnetic stand (~3 minutes)].
- Remove and discard all supernatant.

- Wash beads as follows:
  - Remove from the magnetic stand.
  - Add 50 µl TWB HT to each well.
  - Seal and shake at 1600 rpm for 1 minute.
  - Centrifuge 500 × g for 1 minute.
  - Place on the magnetic stand and wait until the liquid is clear (~3 minutes).
  - For first wash only, remove and discard all supernatant from each well.

- Wash beads a second time.
  Leave supernatant in plate for second wash to prevent beads from overdrying.

F. <u>Amplify Tagmented Amplicons:</u>

This step amplifies the tagmented amplicons using a PCR program. The PCR step adds pre-paired 10 base pair Index 1 (i7) adapters, Index 2 (i5) adapters, and sequences required for sequencing cluster generation

Reagents:

- EPM HT (Enhanced PCR Mix HT)
  (Invert to mix. Keep on ice until use)
- Index adapters (IDT for Illumina-PCR Indexes Set 1, 2, 3, 4)
  (Thaw at room temperature. Vortex to mix, and then centrifuge at 1000 × g for 1
  minute)
- Nuclease-free water

Procedure:

- In a 15 ml tube, combine the following volumes to prepare PCR Master Mix. Multiply each volume by the number of samples.
  - EPM HT (12 µl)
  - Nuclease-free water (12 µl)
- Vortex PCR Master Mix to mix.
- Keep the TAG1 plate/strip on magnetic stand and remove TWB HT.
- Use a 20 µl pipette to remove any remaining TWB HT.
- Remove the TAG1 plateStrip from the magnetic stand.
- Add 20 µl PCR Master Mix to each well.
- Add 5µl index adapters to each well of the PCR plate.
- Seal and shake at 1600 rpm for 1 minute.
- If liquid is visible on the seal, centrifuge at 500 x g for 1 minute.
- Inspect to make sure beads are resuspended. To resuspend, set your pipette to 35 µl with the plunger down, and then slowly pipette to mix.
- Place on the preprogrammed thermal cycler and run the COVIDSeq TAG PCR program

  Program:
  - Choose the preheat lid option and set to 100°C
  - Set the reaction volume to 25 µl
  - 72°C for 3 minutes
  - 98°C for 3 minutes
  - 7 cycles of:

    - 98°C for 20 seconds
    - 60°C for 30 seconds
    - 72°C for 1 minute
- 72°C for 3 minutes
- Hold at 10°C

G. <u>Pool and Clean Up Libraries:</u>

This step combines libraries from each 96-well sample plate into one 1.7 ml tube. Libraries of optimal size are then bound to magnetic beads, and fragments that are too small or large are wash away.

Reagents

- ITB (Illumina Tune Beads)
  (Vortex before each use)
- RSB HT (Resuspension Buffer HT)
  (Let stand for 30 minutes to bring to room temperature. Vortex and invert to mix)
- Freshly prepared 80% ethanol (EtOH)

Procedure:

- Centrifuge the TAG1 plate/strip at 500 × g for 1 minute.
- Place on the magnetic stand and wait until the liquid is clear (~3 minutes).
- To pool libraries, do as follows. Repeat the steps for each additional sample plate.
    - Use a 20 µl eight-channel pipette to transfer 5µl library from each well of the PCR plate to a PCR 8-tube strip. Change tips after each column. These volumes result in 60 µl pooled library per row.
    - Label a new 1.7 ml tube Pooled ITB.
    - Transfer 55 µl pooled library from each well of the PCR 8-tube strip into the Pooled ITB tube.

      For each sample plate, these volumes results in 440 µl pools of pooled libraries.

- Vortex the Pooled ITB tubes to mix, and then centrifuge briefly.
- Vortex ITB to resuspend.
- Add ITB using the resulting volume of Pooled ITB tube volume multiplied by 0.9. For example, for 96 samples, add 396 µl ITB to each tube.
- Vortex to mix.
- Incubate at room temperature for 5 minutes.
- Centrifuge briefly.

- Place on the magnetic stand and wait until the liquid is clear (~5 minutes).
- Remove and discard all supernatant.
- Wash beads as follows.
    - Keep on the magnetic stand and add 1000 µl fresh 80% EtOH to each tube. (note: while doing pooling on a pcr strip we used 200µl ethanol)
    - Wait 30 seconds.
    - Remove and discard all supernatant.
- Wash beads a second time.
- Use a 20 µl pipette to remove all residual EtOH.
- Add 55 µl RSB HT.
- Vortex to mix, and then centrifuge briefly.
- Incubate at room temperature for 2 minutes.
- Place on the magnetic stand and wait until the liquid is clear (~2 minutes).
- Transfer 50 µl supernatant from each Pooled ITB tube to a new microcentrifuge tube.

SAFE STOPPING POINT
If you are stopping, cap the tube and store at -25°C to -15°C for up to 30 days.

Appendix 1b: Denaturation of Library and Loading in the MiSeq

1. Thaw the MiSeq sequence cassette and the Hyb Buffer.
2. Pool samples in equimolar amounts at 2nM. The total volume should be more than 20 µl.
3. Prepare a 2nM solution of PhiX control by combining 1 µl of 10 nM PhiX control with 4 µl of H2O.
4. Combine 1 µl of 2nM PhiX control with 19 µl of 2nM sample pool to obtain a 2nM solution of sample pool and PhiX control.
5. Prepare a 0.2 N NaOH solution by combining 2 µl of 2N NaOH with 18 µl of H2O.
6. Combine 10 µl of 0.2 N NaOH solution with 10 µl of 2nM sample pool and PhiX control solution. Incubate at room temperature for 5 min.
7. Add 980 µl of Hyb Buffer to the 20 µl. Mix by vortexing. (This is your 20 pM solution).

8. Pipet 500 µl of the sample pool/Hyb Buffer mix from the previous step into a 1.5 ml tube. Add another 500 µl of Hyb Buffer to obtain a 10pM solution. Mix by vortexing.
9. Load 600 µl of the 10 pM solution into the designated well of the thawed MiSeq sequence cassette and follow the instructions on the MiSeq to start sequencing.

Appendix 2: Ethical Approval

Government of Nepal

Nepal Health Research Council (NHRC)

Estd. 1991

Ref. No.: 143

28 July 2022

Prof. Krishna Das Manandhar

Principal Investigator, Central Department of Biotechnology, TU

Dr. David Wang

Principal Investigator, Washington University in St. Louis

Subject: **Approval of the Continuing Review Report** for the study entitled **Emerging Infections: Surveillance, Epidemiology and Pathogenesis (Reg. no. 274/2020, Approved on 1 July 2020)**

Dear Prof. Manandhar and Dr. Wang,

The meeting of the Expedited Review Sub-Committee of Nepal Health Research Council held on 26 July 2022 discussed and approved the Continuing Review Report dated 14 July 2022 and decided to approve the continuation of the study for an additional year. This approval is valid till July 2023. The date for submission of next continuing review report is June 2023. The additional budget for the extended period is USD 100762.0 and the ethical review processing fee received at NHRC is USD 3022.86 accordingly.

Note: Please adhere with the timeline mentioned in the approval letter. Any communication regarding the study will not be entertained after the completion of the timeline.

If you have any queries, please feel free to contact the Ethical Review M & E Section of NHRC.

Thanking you!

Dr. Pradip Gyanwali

Member Secretary

Tel: +977 1 4254220, Fax: +977 1 4262469, Ramshah Path, PO Box: 7626, Kathmandu, Nepal
Website: http://www.nhrc.gov.np, E-mail: nhrc@nhrc.gov.np

Government of Nepal

Nepal Health Research Council (NHRC)

Estd. 1991

Ref. No.: 2805

12 April 2021

Prof. Dr. Krishna Das Manandhar
Principal Investigator, Central Department of Biotechnology, Tribhuvan University
Dr. David Wang
Principal Investigator, Washington University in St. Louis

Subject: **Approval of requested Amendment for a research** study entitled **"Emerging Infections: Surveillance, Epidemiology, and Pathogenesis (Reg. no: 274/2020, Approved on 1 July 2020)**

Dear Prof. Dr. Manandhar and Dr. Wang,

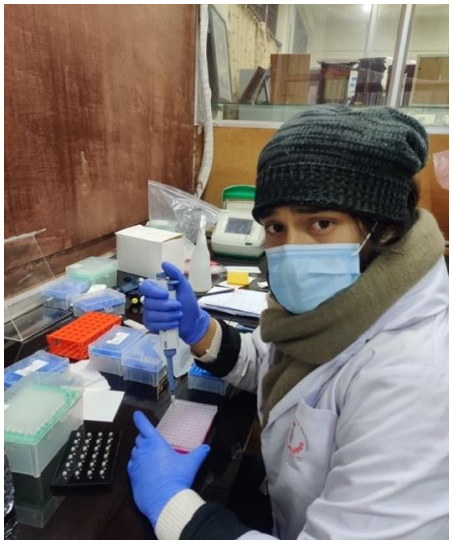
The meeting of the Ethical Review Board of Nepal Health Research Council held on 8 April 2021 discussed the amendment requested on 15 March 2021. The meeting has approved the amendment to include SARS CoV-2 variant in the above-mentioned study as the amendment is required based on the need for time to prepare for the detection of a probable outbreak of COVID-19 by new variants of SARS CoV-2 in Nepal.

If you have any queries, please feel free to contact the Ethical Review M & E section of NHRC.
Thanking You,

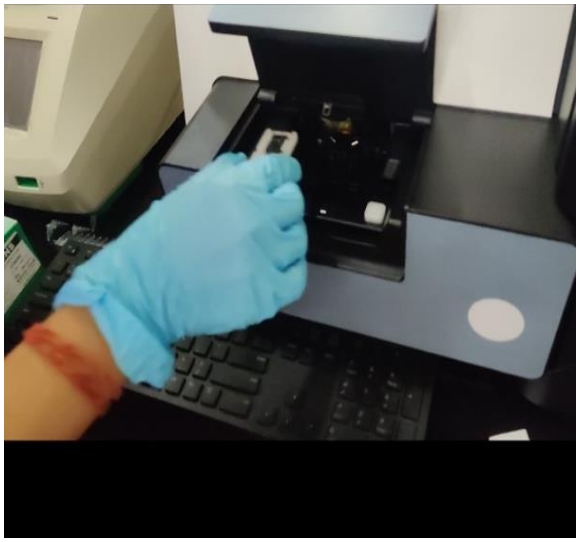
Dr. Pradip Gyanwali
Member Secretary
(Executive Chief)

Tel: +977 1 4254220, Fax: +977 1 4262469, Ramshah Path, PO Box: 7626, Kathmandu, Nepal
Website: http://www.nhrc.gov.np, E-mail: nhrc@nhrc.gov.np

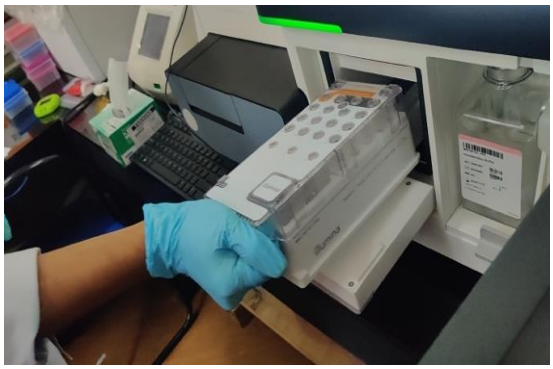
Appendix 4: Photographs



Library preparation for sequencing



Loading of Flow-cell in MiSeq



Loading of sample in MiSeq machine

Running of Sequence machine, MiSeq

GENOMIC SURVEILLANCE OF SARS-CoV-2 TO RECONSTRUCT INFECTION DYNAMICS AND PHYLODYNAMICS USING PHYLOGENETIC INFERENCE OF NEPAL