Development of bioinformatics tools for the characterization and classification of low abundant microbes at the strain level, with a study case of SARS-CoV2

By

Zahra Vafadoost

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science in Medicine

(Human Genetics and Genomics)

Division of BioMedical Sciences

Faculty of Medicine

February 2024

Memorial University of Newfoundland

St. John's
Newfoundland and Labrador

Abstract

Microbiome and metagenomics studies are key research areas to understand several diseases important for public health. The latest sequencing technologies have allowed affordable massive sequencing of specimen microenvironment and the characterization of pathological microbe strains. Strain level is the lowest level of taxonomic ranks; the characterization of strain sequences of a pathological microbe helps track new potential virulent variants and vaccine development.

My master project aimed to; 1) set a bioinformatics pipeline for specimen metagenomics analysis; and 2) characterize potential bias linked to sequencer technologies. A publicly available dataset of human tissues and SARS-CoV-2 swab specimens from the Global Initiative on Sharing All Influenza Data (GISAID)[1] database was used.

In detail, a pipeline was designed to analyze low-abundance metagenomics sequencing data from RNA samples extracted from human tissues. Furthermore, taking advantage of the worldwide effort to track the emergence of SARS-CoV-2 variants, sequencing datasets gathered via the two main sequencing platforms (Illumina and Nanopore) were analyzed to identify potential sequencing biases linked to specific sequencing protocols. Also, a descriptive analysis was generated by applying clustering techniques to the phylogenetic tree and processing and evaluating the metadata's effect on the dataset. Overall, my project guides analyzing metagenomics data for strain characterization when working with low-abundant microbiome data.

General summary

In this project, a tool was designed to help understand the available low-abundance data for microbes that are hard to detect in usual ways. In addition, it provides some information to understand if there is a bias related to sequencing technologies and the protocols used for sequencing devices. Data from publicly available datasets, GISAID, for SARS-CoV-2 were used.

This project provides crucial information about different variants, like the rate of transmissibility in every country and the availability of potential bias in two sequencing technologies, Nanopore and Illumina. Furthermore, the relationship between the sequencing technology and sequencing protocol was explored, and no apparent association between consensus nucleotides and the usage of ARTIC protocol in different continents was found. Moreover, the effect of depth of coverage on sequencing technologies was studied, and there was evidence was for proving this effect.

Acknowledgment

I would like to express my deepest gratitude to my supervisor Dr. Touati Benoukraf, whose guidance and support were invaluable throughout the research process. His insightful comments and constructive feedback helped me to improve the quality of my work and achieve the best possible results.

I would like to extend my sincere appreciation to my co-supervisor Dr. Lourdes Peña-Castillo and the committee members, Dr. Sevtap Savas and Dr. Oscar Meruvia, for their time and effort in evaluating my work and providing valuable feedback that helped me refine my research and improve the quality of my thesis. Their support and encouragement meant a lot to me and helped me to stay focused on my research goals.

I would like to express my gratitude to the School of Graduate Studies for their financial support, which has been crucial in the successful completion of my research project. I also would like to acknowledge the Canada Research Chairs Program for supporting this research project.

I would also like to express my deep appreciation to the GISAID website and the laboratories that provided data for them. Their generosity in sharing their data and resources was crucial to the success of my research, and I am grateful for the opportunities they provided me to learn and grow as a researcher.

Last but not least, I would like to express my profound gratitude to my family for their unwavering support and encouragement throughout my academic journey. Their

love, patience, and understanding have been the driving force behind my success, and I am blessed to have them in my life; I love you deeply.

Thank you all for your kindness, generosity, and support. I could not have achieved this without you.

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Abbreviations

Angiotensin Converting Enzyme 2: ACE2

Comma-separated values: CSV

Depth-first search: DFS

Global Initiative on Sharing All Influenza Data: GISAID

International Union of Pure and Applied Chemistry: IUPAC

Middle East Respiratory Syndrome Coronavirus: MERS-CoV

Multiple sequence Alignment: MSA

National Center for Biotechnology Information: NCBI

Next-generation sequencing: NGS

SARS-CoV-2 Illumina GeNome Assembly Line: COVID-19-Signal

Severe Acute Respiratory Syndrome Coronavirus: SARS-CoV

Severe Acute Respiratory Syndrome Coronavirus 2: SARS-CoV-2

Structured Query Language: SQL

Unweighted Pair Group Method with Arithmetic Mean: UPGMA

Variants of concern: VOC

World Health Organization: WHO

Whole genome sequencing: WGS

1 Introduction

1.1 Microbiome

The human body is an environment for microorganisms like viruses, bacteria, fungi, and protozoa named microbiomes. Different parts of the human body are their habitats. Microbiomes can be found on the skin, gut, mouth, lungs and other mucosal environments. This microorganism ecosystem is called microbiota. The genome of microorganisms is also called the microbiome. During the past decades, many researchers have been studying microbiomes. Recent studies showed that gut microbiomes actively impact host function, metabolism and immunity [2, 3, 4, 5]. The gut microbiota is the microbes in the human digestive tract [6, 7]. They play an essential role in the health and disease of their host, and the dysbiosis of microbiomes can cause health problems for their host [8, 9].

Microbiome composition is determined by both environment and genetics of the host [10]. Some studies showed that the environment plays a more significant role than genetics in forming microbiomes [11]. The composition of more than 20% of microbiomes inside the human body depends on factors such as diet and medicine [10]. Microbiota and the immune system have a two-sided relationship with each other. Microbiota affects the immune system's functionality, and the immune system impacts the symbiosis of microbiomes [12]. Moreover, this relationship helps the body to have a protective response to pathogens and maintain a tolerance for harmless antigens [12].

Research has indicated that the ecology of a microhabitat in humans is linked to numerous illnesses, including COVID-19 [13]. Some evidence shows changes in an increase in potentially harmful bacteria and a decrease in beneficial bacteria [13].

This study focuses on developing a tool for identifying and classifying low-abundance microbiomes. SARS-CoV-2 is chosen as a low abundant microbiome due to its importance and data availability during the pandemic.

1.2 SARS-CoV2

In December 2019, Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) appeared, which resulted in a novel coronavirus designation (COVID-19), which had a high morbidity and mortality rate [14] and threatened human health, causing a public health crisis [15, 16]. COVID-19 was declared the most common threat to humans in the past few years, and on January 30, 2020, the World Health Organization (WHO) announced a global health emergency. On March 11, 2020, the WHO declared the COVID-19 pandemic after all efforts to prevent the infection's spread were unsuccessful [17, 18].

SARS-CoV-2 belongs to the Coronavirdiae family from the Nidovirales order [19, 18], [20]. It is the second most common virus associated with common colds [21, 22]. In general, coronaviruses are divided into four genera based on their protein sequences: Alpha, Beta, Gamma, and Delta [23, 24]. Among these four genera, Alpha and Beta can infect mammals [23]. In the last decades, there have been a few Beta coronavirus outbreaks, including Middle East Respiratory Syndrome Coronavirus (MERS-CoV), Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and novel CoronaVirus-

2019 (2019-nCoV) [23]. In November 2002, SARS-CoV caused an epidemic in Guangdong Province, China [25].

COVID-19 has affected most countries all around the world. The WHO reported that as of October 2022, there were more than 623 million confirmed cases and more than 6 million deaths worldwide due to COVID-19 [26]. According to the same report, the United States of America had the most reported confirmed cases and deaths caused by COVID-19, with 95,687,463 cases and 1,054,151 deaths [26]. India was next, with 44,634,376 confirmed cases and 528,923 deaths, and France, with 35,312,935 confirmed cases and 152,499 deaths, had the third-highest numbers globally [26].

In October 2022, the Canadian government released a report stating there were 4,293,273 total confirmed COVID-19 cases (with a rate of 11,225 per 100,000 population) and 45,689 COVID-19 deaths across Canada (119 per 100,000 population) [27]. On October 8, 2022, Ontario province had the most confirmed cases, reporting 1,462,015 or 9,861 per 100,000 population [27]. Next in severity is Quebec (1,208,599 cases and a rate of 14,046 per 100,000 population), Alberta (609,465 cases and a rate of 13,718 per 100,000), and British Columbia (386,293 cases and a rate of 7,408 per 100,000 population [27]. Among the Canadian provinces, Quebec has reported the most deaths (16,761 deaths with a rate of 195 per 100,000), followed by Ontario (14,457 deaths with a rate of 98 per 100,000), Alberta (4,931 deaths with a rate of 111 per 100,000) and British Columbia (4,370 and rate of 84 per 100,000 population) [27]. In the province of Newfoundland and Labrador, there have been 52,111 confirmed cases and 246 deaths, meaning the infection rate is 10,011 cases per 100,000 and the death rate is 47 per 100,000 [27].

SARS is transmitted through respiratory droplets and close contact with diseased individuals. Its symptoms include high fever (>38.0 °C), discomfort, chills, headache, cough, and difficulty in breathing. Although the known SARS cases were around 8000, it caused more than 700 death (10 percent mortality rate) [21, 28].

MERS-CoV was first reported in 2012 in Saudi Arabia. It also causes acute respiratory syndrome with symptoms such as cough, fever, and shortness of breath, spreads through close contact, and it has a low human-to-human transmission. The total confirmed cases were more than 2500, and it caused around 900 deaths (Mortality rate of 34.5%) [21, 29, 30, 31, 32].

The current Coronavirus (COVID-19) is caused by SARS-CoV-2 and belongs to the genus Beta. It is considered the third major coronavirus outbreak in the last 20 years after SARS and MERS [33]. Patients with COVID-19 have symptoms similar to viral pneumonia, Influenza, rhinovirus, and adenovirus [34]. Symptoms depend on the severity of the illness and include cough, fever, chest pain and discomfort, dyspnea, and pulmonary infiltrate [35, 36, 37]. At first, sick animals were thought to be the primary source and person-to-person transmission was considered unlikely; however, further evidence rejects this opinion [36]. The source and natural reservoir of the 2019 novel coronavirus are still unknown[38].

Origin

Most of the first documented cases were epidemiologically linked to a market selling seafood and live animals in downtown Wuhan [39]. However, according to the

Joint WHO-China Study document, there is conflicting evidence for this assumption, and therefore there is no firm conclusion about the outbreak's origin [40].

Natural reservoirs

In general, bats are considered to be the natural host reservoir for SARS-like coronavirus [41]. Evidence suggests bats and pangolins as the natural reservoir of SARS-CoV-2. However, neither of the viruses is similar enough to be the direct ancestor of SARS-CoV-2. Also, cats and mink are suggested as possible SARS-CoV-2 reservoirs [40].

According to WHO, in December 2021, SARS-CoV-2 was spread by close contact with an infected person [42]. Both symptomatic and asymptomatic infected people can be contagious [42]. When an infected person coughs, sneezes, speaks or breathes, virus particles can spread from their mouth or nose [42]. Another person can inhale the particles, or the droplets can enter their eyes, mouth, or nose. Also, the virus can be spread if a person touches a contaminated surface and then touches their nose, eyes, or mouth [42]. Evidence suggests the virus is likely to spread in crowded places with poor ventilation [42].

COVID -19 structure

Coronavirus is an enveloped virus, and its size is in the range of 70-90 nm [43], [44, 45]; it contains a single-stranded positive-sense RNA [21, 46]. The genome size of SARS-CoV-2, with approximately 30 kilobase pairs, is the largest genome among known RNA viruses [1, 36, 43, 44, 46, 47, 48, 49, 50].

The coronavirus' genome encodes four structural proteins. Nucleocapsid protein (N protein) forms a complex capsid with COVID-19 RNA [51]. There are also Spike protein (S), envelope (E) and membrane (M) proteins [51].

Coronaviruses are named after their crown-shaped spike proteins [52]. Spike proteins are multifunctional proteins which help the binding of the virus for its entry to the host cells [53]. Spike protein interacts with host angiotensin-converting enzyme-2 (ACE2) as its receptor [54, 55]. ACE2 receptors are distributed in human tissues such as the stomach, colon, liver, lungs and kidney [56, 57, 58]. This interaction between the host ACE2 receptor and Spike protein causes a sequence of events resulting in the fusion of the envelope protein (E protein) with the host membrane or endocytosis [56, 57, 58].

Mutations

SARS-CoV2 shares more than 82% of its genome with SARS-CoV and MERS-CoV, as well as over 90% of structural proteins and enzymes [1, 59, 60]. A few coding genes appear to contradict the basic features of the viral genome and the minimal grouping of hereditary data [1, 59, 60]. Adaptive mutations in multiple regions in the SARS-CoV-2 genome can modify its pathogenic potential while complicating treatment and vaccine development [61, 62]. However, the virus must spread efficiently to result in large-scale person-to-person transmission, as seen in the previous SARS pandemic. Infected individuals commonly develop symptoms 3–7 days after exposure. Although the symptoms can occur between 1 to 14 days after exposure to the virus, some individuals are asymptomatic or pre-symptomatic [63, 64].

Acute COVID symptoms include flu-like symptoms such as sore throat, congestion or runny nose, sneezing, fever or chills; respiratory symptoms including cough, throat pain, and shortness of breath; musculoskeletal symptoms such as myalgia and fatigue or weakness [63, 64, 65, 66]. Other symptoms include abdominal pain, nausea or vomiting, loss of taste or smell and diarrhea, headache, chest pain, loss of appetite, and confusion [63, 64, 65, 66]. Symptoms such as seizures, meningoencephalitis, and immune-mediated neurological diseases have been reported on rare occasions [63, 64], [65, 66]. In mild cases, the person recovers 7–10 days after the onset of the symptoms [67 ,[68, 69]. However, in more severe cases, the recovery could be delayed for weeks, and the person could experience symptoms for months [67, 68, 69]. Long COVID-19 symptoms include fatigue, headache, attention disorder, hair loss, and dyspnea [67, 68, 69]. Other symptoms such as myalgia, palpitations, chest pain, and arthralgia are also common in long COVID-19 [67, 68, 69], which can ultimately lead to death, depending on patients' comorbidity.

Effect on public health

Characterizing mutation is very important because this is a public health issue, and vaccinations may not target these variants. Variants of concern (VOCs) in SARS-CoV-2, such as Alpha, Beta, Gamma, Delta, and Omicron, exhibit increased transmissibility and impact public health [70]. Notably, the Delta variant surpassed the Alpha variant and was later overshadowed by the highly infectious Omicron variant [70]. These variants, marked by mutations in spike proteins, affect disease severity, vaccine efficacy, and immune response [70].

1.3 Sequencing technologies

Two generally employed sequencing technologies during the pandemic were Nanopore and Illumina [71], and the focus of this study was these sequencing technologies.

Illumina

SARS-CoV-2 was first identified as a novel coronavirus by using metagenomics next-generation sequencing technology [72]. Next (or second) generation sequencing (NGS) is the most popular sequencing technique around the world. It can be divided into two subgroups, sequencing by hybridization and sequencing by synthesis (SBS) [73], [74]. Illumina by synthesis (SBS) is one of the next-generation technologies [75]. It performs both single-read and paired-end runs [76]. Illumina offers various sequencing platforms such as MiSeq, iSeq, HiSeq, MiniSeq, NextSeq and NovaSeq [77]. The main steps in next-generation sequencings, such as Illumina, include preparation, amplification, sequencing and analysis [78]. The preparation step rule is to make the sample compatible with the sequencer; this is usually done by adding some special adaptor [78]. Illumina platforms use a bridge amplification strategy for clonal amplification [79, 80]. For Bridge amplification DNA molecules (approximately 500 bp) attach to a flow cell and then amplified locally, which result in randomly scattered thousands of copies of the sequence [81, 82, 83]. Then sequencing by synthesis happens, and each fluorescent tag nucleotide binds to its natural complementary nucleotide in the DNA template [78, 84]. Then in the base calling process (in the analysis step), nucleotides are identified [78, 84]. MiSeq is

one of the most common second-generation sequencing platforms [85]. It uses SBS technology and has a rapid turnaround time, as it can produce 25M single reads and 50M paired-end reads in one runtime of 4 hours [85].

Nanopore

Another sequencing technique (third generation) is Nanopore sequencing [86, 87]. The first commercial Nanopore sequencer was MinIO, which was released in 2014 by Oxford Nanopore Technologies Inc. [88]. Oxford Nanopore Technologies provides direct sequencing of DNA and RNA molecules [88, 89]. The sequencer uses electrophoresis to transport nucleic acids through nanometer-sized protein channels [86]. It then determines the nucleotide sequence by monitoring the changes in electrical conductivity and decoding them using base-calling algorithms [86, 88, 90, 91].

Nanopore sequencers such as MinIO are portable and affordable, so they are ideal for real-time sequencing [92]. Moreover, because they provide real-time direct analysis and immediate access to results, they are suitable for analyzing critical information, such as identifying pathogens [90].

Illumina vs Nanopore

Nanopore and Illumina sequencing are both powerful technologies [93]. A drawback of the Nanopore platform is its low accuracy [94]. The error rate in Nanopore sequencing is significantly higher compared to Illumina platforms [95]. Comparing these two sequencing technologies shows that Illumina has a higher accuracy with an error rate of less than 1% compared to Nanopore, with an error rate between 5 to 15% [96]. The

read length in Nanopore is higher, with up to 900 kilobase pairs, compared to Illumina, with a read length of up to 150 base pairs [96]. The Nanopore sequencing platform requires a shorter time than short-read sequencing platforms like Illumina [97]. Other advantages of Nanopore technology over short-read technologies are real-time analysis and data access [91]. To better understand the comparison between two sequencing technologies, it is shown in Table 1-1.

Table 1-1. Comparison of Nanopore and Illumina.

Nanopore	Illumina
Low accuracy [94]	High accuracy
Higher error rate [95]	Lower error rate
Lower accuracy with an error rate between 5% to 15% [96]	Higher accuracy with error rate less than 1% [96]
Higher read length, up to 900 kilobase pairs [96]	Low read length, up to 150 base pairs [96]
Long read	Short read
Real-time analysis [91]	

The way of library preparation for Illumina and Nanopore is shown in Figure 1-1. the amplicons will match the regions of interest in the RNA, which will then be amplified. Further, the amplified regions will be sequenced by sequencing technologies, Nanopore and Illumina.

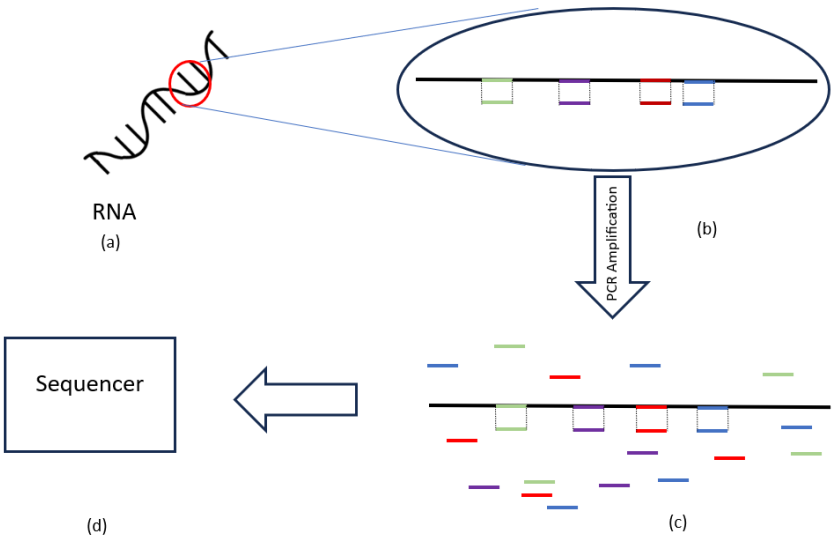


Figure 1-1. This figure shows (a) a fragment of RNA, then (b) the amplicons will match the regions of interest, (c) these regions will amplify, and (d) sequenced by Illumina and Nanopore.

1.4 ARTIC protocol

In January 2019, Artic Network designed and distributed a protocol for COVID-19 (ARTIC). This protocol was initially released for the Nanopore sequencer, but other sequencing technologies also used some part of the protocol [98]. The ARTIC protocol is used broadly worldwide. This broad usage results in nearly the entire whole genome database presently accessible on the GISAID site [99]. The first version of the protocol had almost 98% genome coverage and dropped out some amplicons, resulting in versions two and three [99]. In version three, the ARTIC protocol, for improving the genome coverage, used alternative primers, resulting in almost complete coverage [99].

The primary source of the data used in this project was GISAID[1]. This protocol was mainly used in GISAID; it had an essential role in this project, and the relationship

between ARTIC protocol and sequencing technologies in cases where there is suspicion of a patient having two different strains or an error in sequencing is being studied.

Benchmarking study

Benchmarking is a process that compares products or functions with a standard and indicates a standard of excellence [100]. This method helped identify the strengths and weaknesses and provided guidance for choosing the best protocol for other studies [101]. No whole genome sequencing (WGS) benchmarking of SARS-CoV-2 protocols was found before Liu et al. [101]. Other researchers studied the characterization of one WGS protocol, but no comprehensive study compared all protocols and platforms using the same patient samples[101]. They compared seven library protocols of SARS-CoV-2 on a nasopharyngeal swab of 8 patients [101].

Liu and their team's research have provided information that helps choose an efficient protocol [101]. They compared some characteristics such as mappability, genome coverage, the effect of read depth, and reproducibility [101]. Two of these seven protocols used amplicon method methods, p1 and p7, and they had an ARTIC v3 primer set [101]. Their study showed that some comparisons like sequence mapping and viral genome coverage and the effect of read depth amplicon-based protocols had a better result than other protocols [101]. In general, one amplicon-based protocol that used ARTIC v3 primer set, along with another protocol-QIAseq FX Single-cell RNA-seq library kit coupled with human rRNA depletion, performs best [101].

1.5 Thesis objectives

Understanding the variant characterization better is essential for public health, and an adequate methodology is crucial to avoid spreading viruses. So that I have these objectives

- To describe the COVID-19 and the way it spreads around the world.
- To evaluate the accuracy of each sequencing technology and ARTIC protocol. which can be split into some sections such as:
    - Employing the UPGMA tree for finding bias in sequencing technologies
    - Studying the effect of confounding factors
    - Studying the gap in the quality of sequencing technologies
    - observing the relationship between sequencing technologies, consensus nucleotides and ARTIC protocol
- To explore the mutation's effect on depth of coverage by sequencing technology

2 Methods

2.1 Descriptive analysis

2.1.1 Analyzing a phylogenetic tree from GISAID

In the first step, the pipeline generates descriptive analysis for given phylogenetic tree. A *phylogenetic tree* is a binary tree that shows the evolutionary relationship between species based on the similarity or differences in their features [102, 103]. A phylogenetic tree in Newick format is the required input for this part. A *Newick format* is a format that uses parentheses and commas for edge length and also contains leaf names (accession id) [104]. The edge length in the phylogenetic tree is their evolutionary distance from their parents in a clade. Also, it is constructed based on the distance of each sequence from its parents.

The Newick format phylogenetic tree collected from the GISAID [1] website was from Jan 2020 to Jun 22nd, 2022; a total of 8,918,723 high-quality genomes are available in this phylogenetic tree. The phylogenetic tree came with a Metadata file containing the accession id, virus name, and collection date. Table 2-1 shows ten random rows from 1048576 total rows.

The pipeline generates different clusters using depth first search (DFS) algorithm, and for each cluster, it generates a pie chart which groups of data in proportion to the entire data [105], a cumulative time chart which shows the total amount of data in a period [106], and an information file containing accession id, collection time and country. The pie chart summarizes the number of COVID-19-infected cases in different countries,

and the cumulative time chart shows the total growth of infected cases in a specific period. The process of generating output from the Newick format input file is available in Figure 2-1.

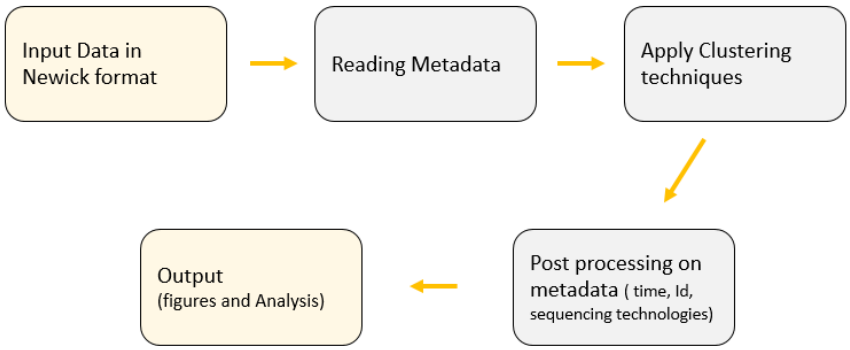


Figure 2-1.Process of generating output from Newick format input file. In the first step, data is given to the pipeline; then, metadata is read from the input. After applying clustering techniques on input files, some post-processing is applied to the metadata output, and some data related to each cluster, such as sequencing technology, id, and collection date, is stored for further analysis. As a result, it provides some features for analyzing data, like calculating the mutation rate and distribution of variants in each cluster.

Table 2-1 Ten random rows from 1,048,576 total rows, extracted from GISAID website.

Accession Id	Virus name	Collection date
EPI_ISL_11281549	hCoV-19/England/MILK-3ACC0FE/2022	2022-03-16
EPI_ISL_11724649	hCoV-19/Japan/PG-208147/2022	2022-03-04
EPI_ISL_12626187	hCoV-19/USA/OR_UO_MW001058_S94_L001/2022	2022-03-04
EPI_ISL_800403	hCoV-19/England/CAMC-D11931/2020	2020-12-26
EPI_ISL_8957430	hCoV-19/USA/VT-CDCBI-CRSP_FOZ6BCMXRAFXLBW4/2022	2022-01-09
EPI_ISL_798905	hCoV-19/England/ALDP-D40623/2020	2020-12-26
EPI_ISL_2095036	hCoV-19/Germany/BY-MVP-000002546/2021	2021-01-29
EPI_ISL_1402096	hCoV-19/Slovenia/21-4818/2021	2021-03-02
EPI_ISL_10657762	hCoV-19/USA/PA-0027/2021	2021-04-14

EPI_ISL_13146341	hCoV-19/USA/IL-LCH-1327/2022	2022-03-21

The metadata file was parsed, and each collection date and accession id for sequences were extracted. The pipeline ignores the data without a collection date. As a result, a dictionary of accession id and collection dates for all valid data is being made. Then clustering techniques are applied to the data. *Clustering* is a machine learning technique for grouping objects based on similarity [107] for better analysis [108], and that helps with identifying new objects, in this case, new variants and finding better treatment or screening methods [108].

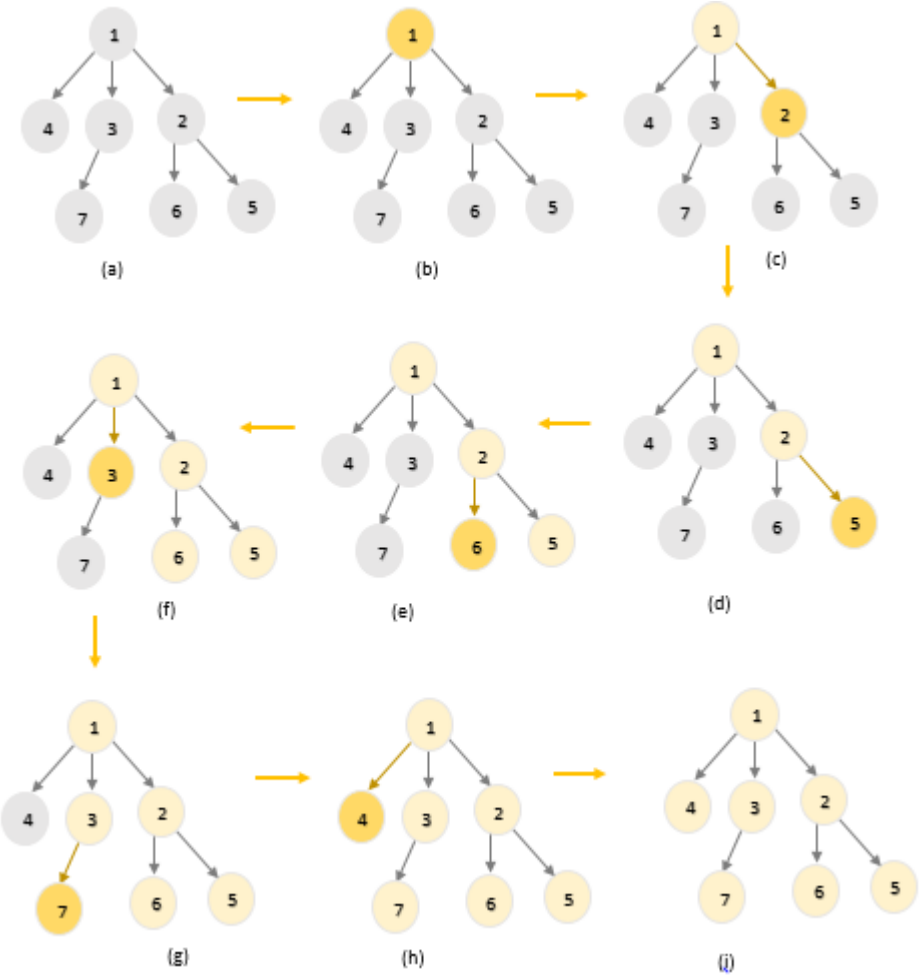


Figure 2-2. Different steps of the DFS algorithm. The DFS algorithm in this project was written using the backtracking idea, meaning it calls itself with different inputs. In each step, if the node is not terminal (b, c, f), the method invokes itself with its child clades; otherwise (d, e, g, h), it backs to the parent clade and continues.

Then, The DFS algorithm was used to cluster data for groups of the various variants from the phylogenetic tree. The *DFS Algorithm* is a technique for traversing a graph or tree. It starts by visiting nodes from the root, exploring the branch's deepest node, and then backtracking [109]. Figure 2.2 shows the steps of the DFS algorithm for a small tree. The DFS starts from the tree's root to the next clade to explore its distance

from the parent. A threshold[1] based on the nodes' distance from their parents was set for clustering; so that the nodes that were not similar categorize into different groups. If the distance from the parent is less than a specific amount, the DFS explores that node. If the clade that the DFS tries to explore is terminal, the DFS goes back one step and explores the sibling clade of that clade.

As a result, the data that their distance from their clades was more than the threshold was employed, and the outliers were ignored. For each cluster, the pipeline generates one pie chart, time chart and a comma-separated values (CSV) analysis file.

The pseudocode for the algorithm of the clustering is available in the following (Algorithm 2.1). The algorithm was implemented in python.

Retrieve phylogenetic tree into variable *treeData* ← input ()
Retrieve metadata from the CSV file into variable *CSVInfo* ← input ()
Parse and return precisely one tree in the Newick format; data is loaded in string format
Perform a for loop for clades in the tree
The DFS method was performed for clustering the tree (colouring the tree)
Save the coloured tree into the output file ← output ()
Perform AnalyzeTree method on the tree to calculate the Minimum and the Maximum time and number of leaves in the Cluster in a dictionary variable
Save the dictionary into a file ← output ()
Generate cluster information based on countries' dictionary
Save the cluster dictionary into a file ← output ()
Plot pie chart for cluster-countries dictionary ← output ()
Plot cumulative time chart ← output ()

DFS method
DFS method with three inputs colour, clade, and distance measure
If the given clade is terminal

[1] The user sets this threshold; branches with more distance from their parent than that threshold are categorized in different clusters. Therefore it adjusts the distance between nodes in the cluster, meaning variants that are more similar to each other are grouped in one category.

```
   Give the leaf the same colour as the father.
Else if the clade is not terminal
   Perform a for loop for nodes in the clade
         If the distance from the parent is less than the distance measure
               Call DFS Method with the same colour, node, and distance measure
         Else if the distance from the parent is more than the distance measure
               Increase colour number
               Call DFS Method with the new colour, node, and distance measure


AnalyzeTree method
AnalyseTree method with two input DFSTreeDictionary, CSVDictionary
Perform a for loop for leaves in the DFSTreeDictionary
If CSVDictionary contains the leaf
   Set minDate to now
   Set maxDate to a min time (start of the pandemic)
   If the collectionDateDictionary was not empty and contained the cluster
number:
         Retrieve minDate for the cluster from collectionDateDictionary
         Retrieve maxDate for the cluster from collectionDateDictionary
   Else
         Save 0 for cluster and minDate and MaxDate into the
collectionDateDictionary dictionary.

   Load collection date into a variable
   If minDate is greater than the collection date
         Save collection date instead of minDate into collectionDateDictionary
   If maxDate is less than the collection date
         Save the collection date instead of maxDate into collectionDateDictionary
```

Algorithm 2-1. pseudocode of algorithm for clustering technique.

2.1.1.1 Rate of mutation in the different parts of the world

A report file in txt format is generated as a result of the Phylogenetic tree analysis, and it contains the cluster number, start time, end time, and the number of variants on that cluster. From the report files, the rate of mutation in each cluster was calculated by

finding the time difference between the first variant[2] and the last one in the cluster and dividing that by the number of total variants.

2.1.1.2 Distribution in different cities/countries

In the file generated during the process, the distribution of variants in different countries in each cluster was reported. This reported distribution helps with reporting the transmission of COVID-19 variants between countries.

2.1.2 Cumulative diagram

The worldwide cumulative time chart is another pipeline output. It assists with understanding the transmission of different variances and comparison variation growth in different countries, which helps with anticipating the behaviour of the SARA-CoV-2 virus worldwide, by showing the starting country of a variant and the way it travelled around the world and infected other countries. The Pyplot library from Matplotlib in the python programming language was used to generate this chart.

2.2 Studying the effect of Sequence technology in providing bias in variants

2.2.1 Aligning the sequences

A Multiple Sequence Alignment (MSA) file is one of the most important outputs of this pipeline because it is the input for the rest of the pipeline. Sequences, metadata containing sequence technology, and collection date are necessary for generating the

[2] Every sequence in the phylogenetic tree is a variant.

MSA file. All sequences for SARS-CoV-2 were extracted from the GISAID browser. Afterwards, the metadata in some files with tab-separated values format was downloaded from the GISAID browser. From all the TSV files, pipelines generate a Structured Query Language (SQL) file for the metadata database. The Muscle[110] Command line package from Biopython (version 1.80) [111] was utilized for aligning the sequences, and it was found to have average accuracy compared to other MSA tools [112]. Although other MSA tools like ProbCons, SATe, and MAFFT(L-INS-i) are more accurate, Muscle was selected as the fastest option for our project [113]. Furthermore, Muscle is one of the most promising approaches for aligning large-scale datasets containing more than 1000 sequences [114]. Muscle has been found to perform well in this project, delivering both fast and accurate results, making it a suitable choice for the task. In the next step, the pipeline parses the aligned sequences for modifying the headers of each sequence to add sequencing technology used for each sequence to their header and generate a new multiple-sequence alignment file.

2.2.2 Sequence alignment viewer

The pipeline's next step was to visualize the MSA to ensure its correctness. Therefore, the sequence alignment viewer was used for this purpose, which uses an interactive plotting library, Bokeh, for rendering the dashboards and graphics [115].

Headers of sequences in multiple sequences alignments (MSA) file has some spaces. First, the space is removed then the MSA result is passed to the clustalW[116] to make a file with multiple alignment format. Further, the generated file is passed to the remote sequence alignment viewer. Also, since the data had IUPAC (International Union of Pure

and Applied Chemistry) codes, the sequence alignment viewer was slightly modified to support them. IUPAC codes are annotations that allow ambiguity in the consensus sequences. They are 16 characters which help with representing states for single nucleic acids or ambiguity with nucleic acids among 2, 3 or 4 potential nucleic acid states. All these 16 characters are available in Table 2-2 [117, 118].

The modified code for the sequence alignment viewer is available in Appendix A.

Table 2-2 IUPAC codes [11] -International Union of Pure and Applied Chemistry codes

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
M	A or C
R	A or G
W	A or T
S	G or C
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	any base
. or -	Gap

2.2.3 Phylogenetic Tree

The next step was generating a phylogenetic tree for all the data to study the effect of sequencing technology. In this regard, hierarchical clustering was used to cluster the available data. A distance matrix was made for COVID-19 data from Canada using the

Phylo library[3]. The distance matrix generated for the phylogenetic tree is a genetic distance between sequences. This matrix was calculated using Phylo.TreeConstruction library from Bio python library (version1.80). The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clustering was used for the hierarchical clustering method. UPGMA is an unweighted pair group method with arithmetic mean, a simple bottom-up hierarchical clustering [119].

There are two types of Hierarchical clustering:

1. Agglomerative Hierarchical Clustering (bottom-up)
2. Divisive Hierarchical Clustering (top-down)

Agglomerative hierarchical clustering starts from n clusters, and in each step, clusters join until they make one cluster [120]. On the other hand, divisive hierarchical clustering starts from one cluster, and in each step, they divide, and at the end, there will be n clusters [120]. Agglomerative clustering is more common [120] and we used one kind of agglomerative clustering because it serves our need in making a bottom-up clustering.

There are various definitions for distance in clustering; single-linkage, complete linkage, and average linkage are the three most common. In the single linkage, the distance between two clusters is represented by the distance between their two closest points [121]. In complete linkage, the distance between two clusters is represented by the distance between the two farthest points [121]. There are two types of average linkage. In one type of average linkage (UPGMA), the distance between two clusters is represented by the average distance between all two points of individuals in the two clusters [120], and in another type (WPGMA), the distance is calculated as a simple average [120].

[3] Phylo library is available in Biophyton; I used Biopython 1.80 in this project.

Given its capability to effectively merge clusters with low variances and its intermediary position between the single and complete linkage methods [120], the UPGMA algorithm closely aligns with our requirements. Moreover, it considers the clusters' structure and demonstrates a considerable degree of robustness [120], making it a superior choice among the available options.

For generating UPGMA, two sequences with more similarity are clustered, and a new distance matrix is made by adding the mean of these two sequences instead of the two closest sequences [122]. This procedure continues until only two sequences remain. These two final sequences, which are the tree's roots, have a far distance compared to other sequences [122].

Then, the DFS algorithm was used to specify colours for the branches and clustering data and calculate the ratio for each clade in the phylogenetic tree. The colouring of the UPGMA tree is based on the sequencing technology used for each sequence. The DFS algorithm is available in Figure 2.2. Figure 2.3 shows the procedure for producing the phylogenetic tree.

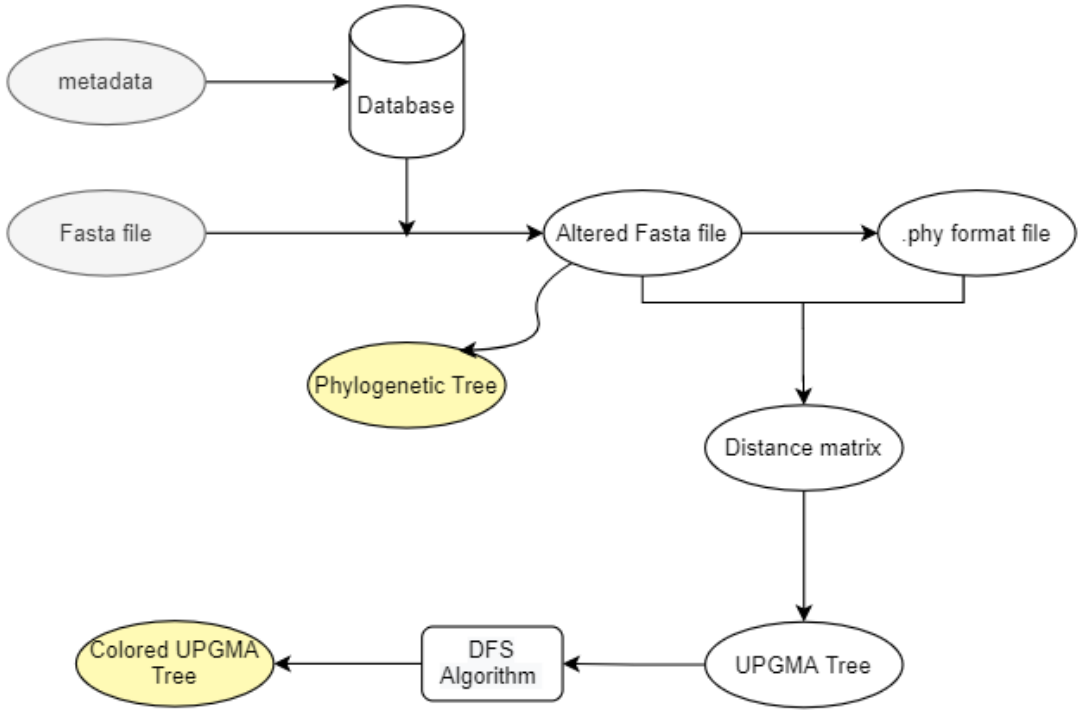


Figure 2-3. The procedure for producing the phylogenetic tree.[4] A metadata file with TSV format was extracted from the GISAID website to generate the database. Then the Altered fasta file is generated using the database, sequencing technology and some other information added to the fasta file. After generating a distance matrix, the UPMGA clustering (Unweighted Pair Group Method with Arithmetic mean), a hierarchical clustering, was applied to the data. Then DFS algorithm was used for colouring the tree based on the sequencing technology. Then UPGMA clustering was applied again to the data, and the coloured phylogenetic tree was generated.

2.2.4 Confounding factors

Confounding factors are variables that influence research; they might indicate a relationship or bias in an experiment, or they can generate outliers; this makes understanding confounding factors vital to have a valid result [123, 124, 125].

One step to better understand the effect of sequencing technology was to remove the effect of confounding factors. In this research, the independent factor is different

[4] This figure was generated using the https://app.diagrams.net/ website

sequencing technology, and the dependent variable is the biases in determining SARS-CoV-2 sequences, while the confounding factors are time and location.

For the time, data from the three most significant waves of infection has been incorporated, focusing on Canadian COVID-19 data. For location, two provinces, Quebec and British Columbia, each utilizing distinct sequencing technologies, were selected for analysis.

Fisher exact test was used to check the potential bias linked to the different sequencing technologies. First, all the nucleotides in every vertical cut were counted for implementing the Fisher exact test. Afterwards, IUPAC codes were changed according to the highest score for the nucleotide to reduce ambiguity. All these processes were done separately for sequences that used Illumina and Nanopore.

Fisher exact test

Fisher exact test is a type of statistical significance test that is utilized for analyzing contingency tables [126] (which is not assume normal distribution). It was used because there were two categorical variables, and the aim was to examine whether the proportions of one variable vary based on the values of the other variable [127]. In this Fisher exact test, the significance between two classifications was studied: First, nucleotides belong to sequences using the Nanopore platform and second, nucleotides sequenced by Illumina. Nanopore's sequencing bias was examined compared to Illumina, which is standard and, in this study, considered as control. Also, the difference between Nanopore and Illumina in comparison to each other was studied.

Table 2-3 and the following formula show the contingency table and formula used to calculate the Fisher exact test for each location. The variables in this study are independent since other variables do not influence them and cannot be manipulated during the research [128].

Table 2-3 Fisher exact test contingency table. The "n" in this table is the specific nucleotide on that location, for example, A, C, G or T in specific loci. The "! Nucleotide" means a nucleotide is not in the collection.

	Nanopore	Illumina	Total
Nucleotide	n in Nanopore	n in Illumina	n in total
!Nucleotide	Not n in Nanopore	Not n in Illumina	Not n in total
Total	Total for Nanopore	Total in Illumina	Total

For Fisher's exact test, library stats for Scipy 1.9.3 in python were used. The following formula is Fisher's exact test formula:

$$p = \frac{(\,(n\ in\ total\,)\,!\,(Not\ n\ in\ total\,)\,!\,(\,Total\ for\ Nanopore\,)\,!\,(Total\ for\ Illumina)\,!\,)}{(n\ in\ Nanopore)!\ (n\ in\ Illumina)!\ (Not\ n\ in\ Nanopore)!\ (Not\ n\ in\ Illumina\,)!\,Total\,!}$$

2.2.4.1 Time

Time was studied as the first confounding factor related to the project. Since the start of the pandemic, there have been different variants available. Data for three highest waves of COVID-19 in Canada were the research subject in this section. Then for each nucleotide location in the sequence, statistical analyses were performed to study the effect of time as a confounding factor. The nucleic acids frequency was calculated for statistical analysis for each position. Then the percentage of data was calculated and used to remove

the effect of sequencing technology. Three major outbreaks from January 2020 until November 2021 were used to remove the effect of time. The first outbreak was from November 2020 to February 2021, the second wave was from March 2021 to May 2021, and the third was from August 2021 to October 2021. Then the effect of time as a confounding factor was studied by the Fisher exact test.

2.2.4.2 Location

The Fisher exact test was used in the pipeline to evaluate the effect of location, one of the confounding variables, in whether sequencing technology provides bias in determining SARS-CoV-2 sequencing. The pipeline needs one location with different sequencing technology as input. The contingency table for calculating the Fisher exact test are available in Table 2-2.

This study focused on the impact of location by selecting the Canadian province of British Columbia, where two sequencing technologies were used.

2.2.5 Gap in quality of sequencing in two sequencing technologies

FastQC was used to compare the sequencing quality in Illumina and Nanopore in the offline mode. The offline mode allows the generation of reports without running the interactive application [129]. FastQC is a tool used to check the quality of raw data [129]. Sequences from one Canadian province, British Columbia, were used for quality control checks. Only British Columbia used both sequencing methods in the SRA database, and the difference in the number of sequences produced by Nanopore and Illumina was smaller than in other provinces. Afterwards, all reports generated with FastQC were

collected, and one thorough report was generated using MultiQC. MultiQC is a tool for collecting and summarizing the result of multiple FastQC reports [130]. In total, there were 106 sequences available; 86 of them were sequenced by Nanopore, and 20 of them were sequenced by Illumina. These 106 raw data were extracted from the SRA database of the National Center for Biotechnology Information (NCBI). The following command was used to download data to make a concatenate read.

$$fasterq - dump\ --concatenate - reads --include - technical < sra_{id} >$$

For downloading multiple SRA following script is being used.

```
import subprocess
SRA_numbers = ["SRA_id","SRA_id","SRA_id"]

for SRA_id in SRA_numbers:
   prefetch = "prefetch " + SRA_id
   subprocess.call(prefetch, shell=True)
   fastq_dump = "fastq-dump --gzip " + SRA_id
   subprocess.call(fastq_dump, shell=True)
```

The report provided by MultiQC is available in the result section (Section 3.2.5).

2.2.6 Relationship between sequencing technologies, consensus nucleotides and ARTIC protocol

After studying sequencing quality in Nanopore and Illumina using FastQC, it is time to examine the consensus nucleotides in different sequences. As an input of the pipeline, some sequences for infection are needed. For calculating the consensus sequences of the SARS-CoV-2, the sequences were downloaded from the GISAID

database. Some sequences in the database have IUPAC (International Union of Pure and Applied Chemistry) codes, which are multiple codons coded as one amino acid [131] and allow ambiguity in the consensus sequences [132]. In Table 2-3, IUPAC codes are available.

Having IUPAC codes in the sequence have two meanings:

1- Error in sequencing

2- Sequences where the patient had at least two strains of the virus, one with mutation and the other with/without mutation.

As the next step, the regions and the sequencer with the most differences were studied. The interest in this part was studying the sequences with two strains. In parts of sequences where there were IUPAC codes, the reference genome was compared to the sequences. Then the IUPAC codes and the percentage of other nucleotides were monitored. Afterwards, the study explored whether there was a connection between sequencing protocols and the abundance of IUPAC codes in sequences across different parts of the world.

Consensus sequences

Calculating consensus sequences is essential for preparing data to study the relationship between sequencing technologies. A consensus sequence is a sequence of aligned related DNA, RNA or protein sequences [133]. In the consensus sequences, every position represents the most common nucleotide or amino acid residues in the vertical cut [133]. For calculating the consensus sequences of the SARS-CoV-2, the sequences were downloaded from the GISAID database.

2.2.7 COVID-19 Signal pipeline [134]

For this part of analyzing data, first, the SRA toolkit was used to download 20 sequences with the maximum number of IUPAC codes from section 2.2.6 and 20 sequences without IUPAC code from the MSA source file available on the NCBI SRA database.

Then SARS-CoV-2 Illumina Genome Assembly Line (Signal) was used for analyzing and quality control of data. The goal was to identify individuals with evidence of infections by two variants. From 20 raw data with IUPAC code available on GISAID, 19 were matched to sequences from the SRA database based on their metadata, and all of them had an excellent quality to be given to the COVID-19-Signal pipeline as input. Phred quality scores for data were greater than 30; therefore, the error rates were less than 0.1%. The *Phred quality score* is a measure to evaluate the accuracy of sequencing technology [135]. The other 20 sequences were randomly selected from 646,783 sequences without IUPAC codes available on the MSA file. These twenty sequences with a match in the SRA database were considered a positive control for the COVID-19-signal pipeline.

COVID-19-SIGNAL was used to study the average read coverage, percentage of fraction and the percentage of each read mapped to the reference genome [136].

2.3 Exploring the mutation's effect on depth of coverage by sequencing technology

The *depth of coverage* is the number of times a single nucleotide in the reference genome has been sequenced and thus included in a read [137]. Both sequencing technologies considered in this study, Nanopore and Illumina, use amplicon methods, which help analyze low abundance DNA inputs[5]. An *amplicon* is a DNA or RNA fragment which has been amplified or replicated [138]. The process of creating two equivalent duplicates of DNA from a DNA molecule is known as DNA replication [138].

Primers used were designed for the Wuhan strain; therefore, it is predicted that the depth of coverage has some gaps and decreases over time because of introduction of numerous mutations. The aim was to investigate the mutation's effect on the depth of coverage of different sequencing technologies, where in all amplicon methods if a mutation occurs in a region that matches the amplicon, the region will not be captured and cannot be appropriately sequenced.

To discover the mutation's effect on the depth of coverage of different sequencing technologies, we calculated the depth of coverage of 50 sequences randomly selected per collection month from January 2020 to September 2022 from the GISAID database.
All the implementation for the pipeline is available on the following GitHub link:
https://github.com/zahraav/covid.git

[5] Nanopore does not require amplification, but ARTIC protocol, which was the primary protocol for COVID-19 data, is an applicant-based method.

3 Results

The previous chapter provided an overview of the methodologies employed in different thesis sections. This chapter examines the results produced in each section by applying those techniques. This chapter is structured into three primary sections, further divided into subsections. Data for this thesis are extracted from the GISAID [1] website.

The first section of this chapter is a descriptive analysis using a dataset of 8,918,723 sequences collected between January 1st, 2020, and June 22nd, 2022. The sequences were divided into 63 clusters based on their distance from their parents using DFS algorithm. Subsequently, each cluster was represented using a pie chart to summarize the distribution of infected cases across different countries, while a cumulative time chart was generated to show the overall growth of infections within a specific period.

The second part of the study investigated how sequencing technologies like Nanopore and Illumina might introduce bias in variants. This objective was explored in several subsections.

Due to certain limitations[6], the first subsection was restricted to a dataset containing only 4,000 nucleotides derived from 1,000 sequences featured in the dataset used to illustrate sequence alignment in the first section of chapter three. In the second subsection, a phylogenetic tree was employed from the pandemic's beginning until August 2020. The tree was then used to generate a hierarchical clustering tree to evaluate whether any bias existed in the sequencing technologies.

[6] For example, we had to use a strong computer or supercomputer for large amounts of data.

The third subsection focused on investigating the influence of confounding factors on the study. Time and location are two confounding factors that were studied. Three outbreaks from the start of the pandemic until October 2021 were used, and location data was limited to one province of Canada, British Columbia, with both sequencing technologies, and the dataset is from January 1st, 2020, to November 30th, 2021. The other provinces had data for only one sequencing technology or did not have enough data for both.

Further, the fourth subsection focused on comparing the gap between the qualities of sequencing technologies. This was performed using MultiQC tools, with data for British Columbia extracted from the SRA toolkit.

The next subsection investigated the relationship between sequencing technologies, ARTIC protocol, and consensus nucleotides. The dataset of this subsection is worldwide data from January 2020 to February 18th, 2021. some of the subsection's results are the distribution of consensus nucleotides across each continent, the proportion of nucleotides per sequencing technology, distribution of sequencing protocol used in each continent.

In the last subsection of the second section, a relationship between the existence of the IUPAC codes with an average depth of coverage and genome fraction was studied using the COVID-19 Signal pipeline. The dataset for this subsection contains 20 sequences with the highest number of IUPAC codes and 20 sequences without IUPAC codes.

The last section of this chapter studied the result of mutation's effect on the depth of coverage of sequencing technologies. The dataset for this section was from the start of the pandemic until September 2022.

3.1 Descriptive analysis

3.1.1 Analyzing a phylogenetic tree from GISAID

A phylogenetic tree in Newick format and its metadata for worldwide data from January 1st, 2020, to Jun 22nd, 2022, is given to the pipeline as an input; a total of 8,918,723 high-quality genomes[7] [1] are available in this phylogenetic tree. My pipeline (described in section 2.1.1) identified 63 clusters in this phylogenetic tree. Figure 3-1. shows the number of variants in each cluster. As is apparent in Figure 3-1, clusters 18, 17, and 19 have the highest number of variants compared to other clusters. The pie chart and cumulative time chart for these three clusters are shown in the following sections. Table 3.1.1 shows the four first countries with the highest number of variants for the three clusters with the highest number of variants. All charts are available in Appendix A. Table 3.1 shows that the USA and England have the highest number of variants per cluster.

[7] Quality checks was done by the GISAID website for all the provided data.

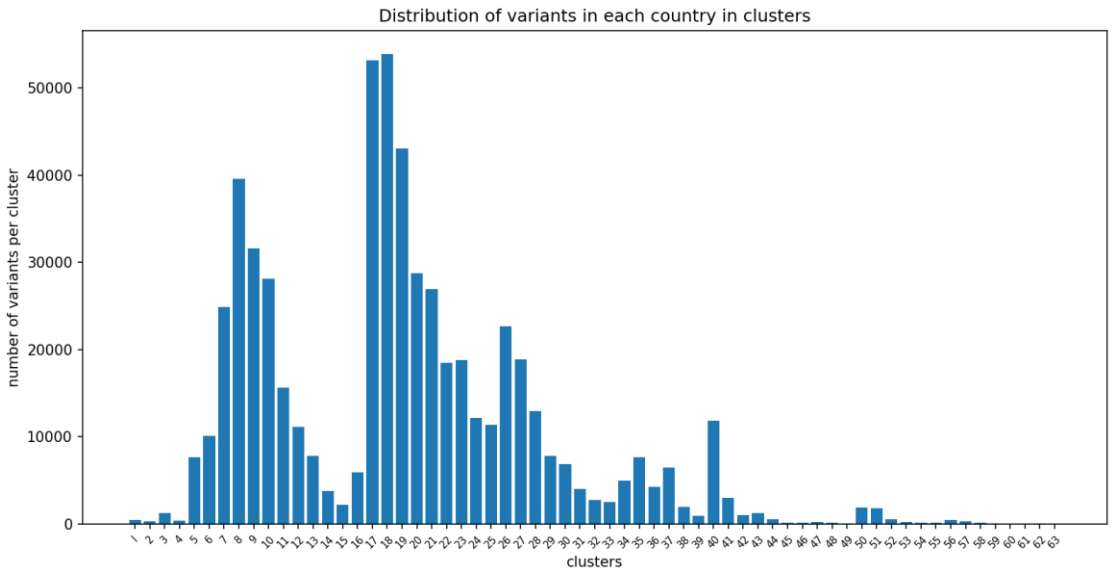


Figure 3-1. Shows the number of variants per cluster. Clusters 18, 17, and 19 have the highest number of variants among all the clusters.

Table 3-1 Three clusters with the highest number of variants. Four countries with the highest number of variants are shown per cluster.

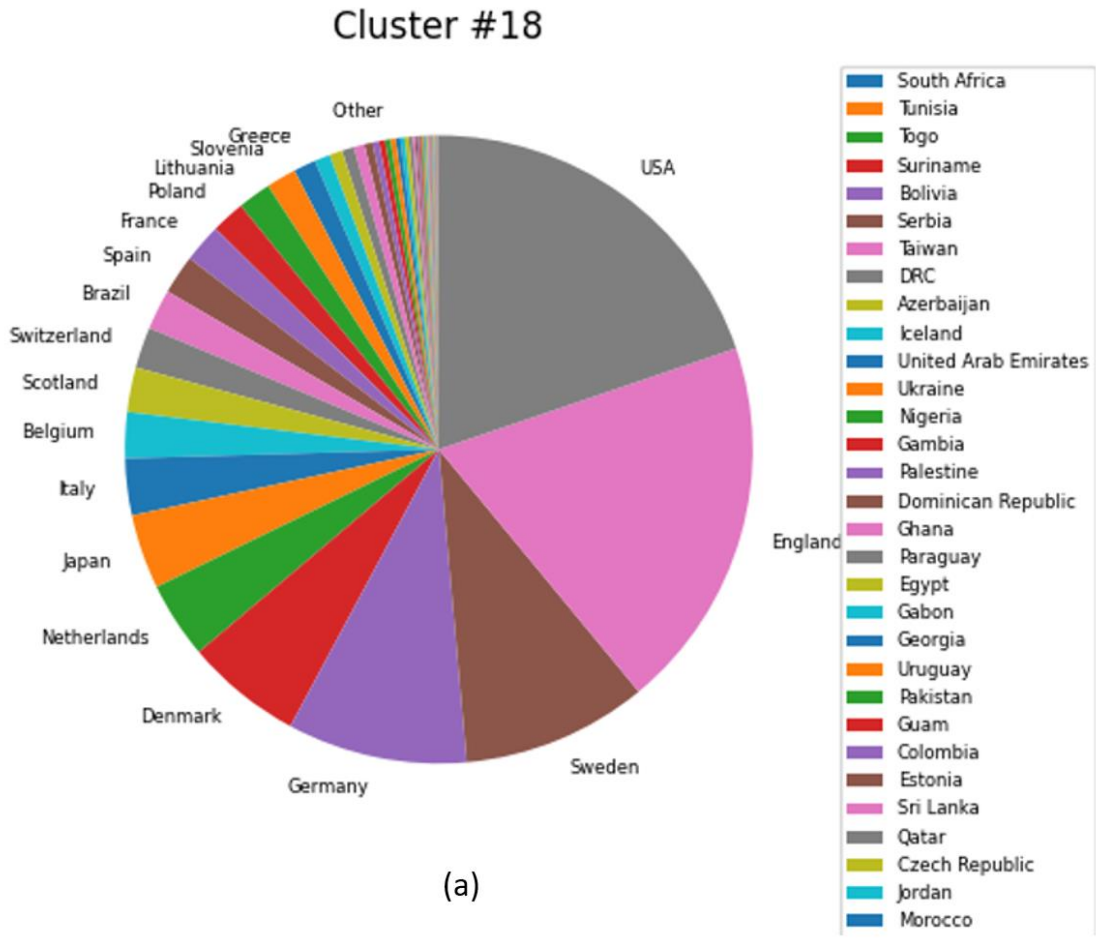
Cluster Number	Total	Country	Number of variants	Country	Number of variants	Country	Number of variants	country	Number of variants
18	53916	USA	10665	England	10356	Sweden	5152	Germany	4986
17	53200	USA	10818	England	9328	Sweden	6204	Germany	5457
19	43084	USA	7457	England	7537	Germany	6944	Denmark	3769

3.1.1.1 Distribution in different cities/countries

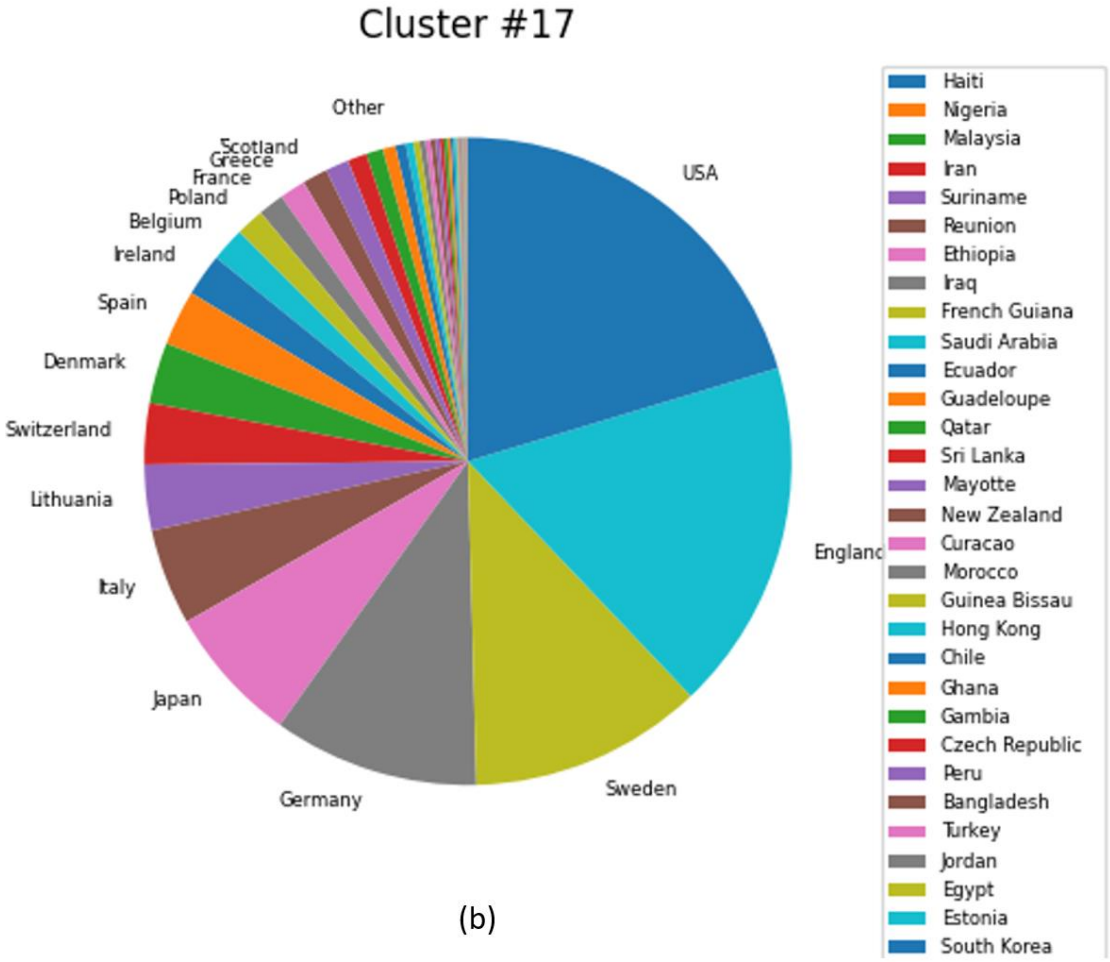
The pipeline generates a CSV report containing information for each cluster. The pie chart and CSV report indicate noticeable clusters with the greatest number of variants and

the highest variant counts for countries within each cluster. For the data that was used from the pie chart of three clusters with the highest number of variants out of 63 generated clusters, USA, England, Sweden, and Germany for the first two clusters and in the third cluster USA, England, Germany, and Denmark are the countries with the highest number of variants.

The distribution of variants in different countries for three clusters with the highest number of variants is available in Figure 3.2. Appendix A shows pie charts for all clusters.



(a)


Cluster #17
Other
Scotland
Greece
France
Poland
Belgium
Ireland
Spain
Denmark
Switzerland
Lithuania
Italy
Japan
Germany
Sweden
Englanc
USA
Haiti
Nigeria
Malaysia
Iran
Suriname
Reunion
Ethiopia
Iraq
French Guiana
Saudi Arabia
Ecuador
Guadeloupe
Qatar
Sri Lanka
Mayotte
New Zealand
Curacao
Morocco
Guinea Bissau
Hong Kong
Chile
Ghana
Gambia
Czech Republic
Peru
Bangladesh
Turkey
Jordan
Egypt
Estonia
South Korea

(b)

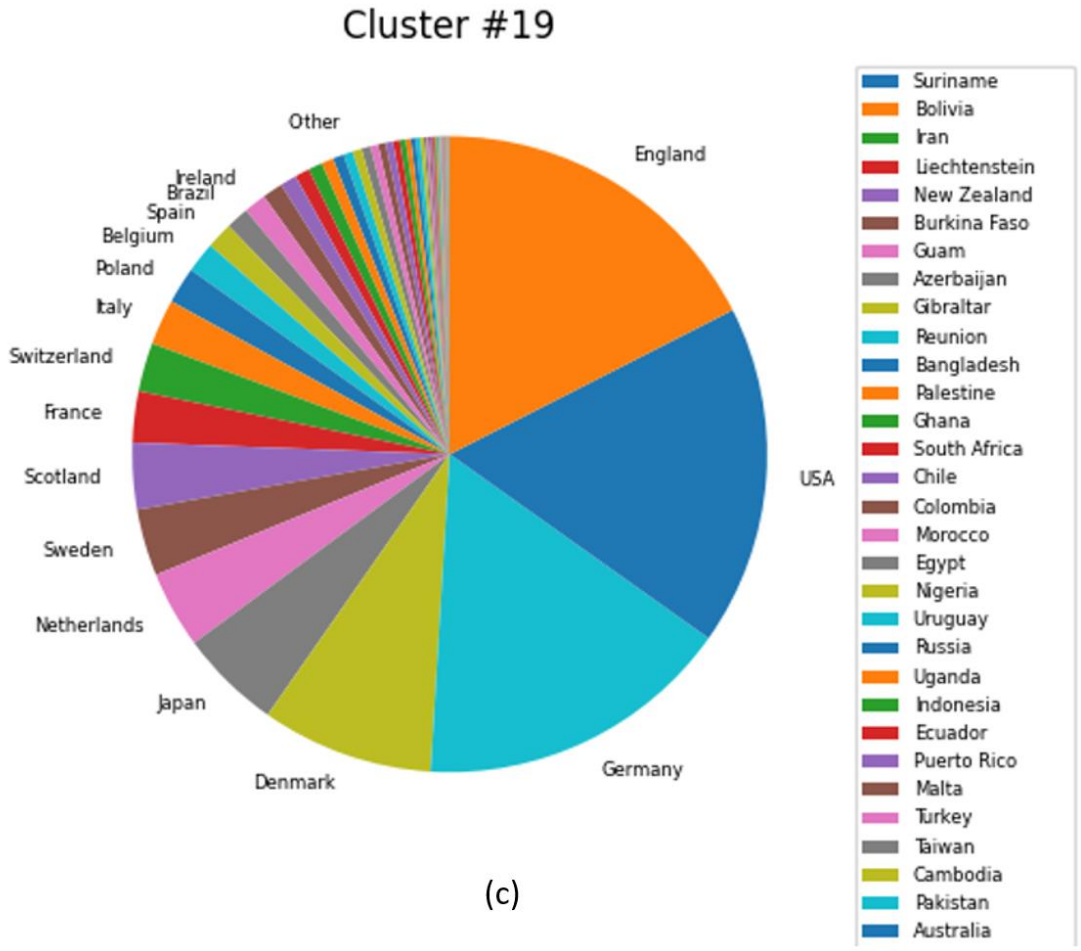


Figure 3-2. Pie-charts for three clusters with the highest number of variants. (a) cluster number 18 (b) cluster number 17 (c) cluster number 19.

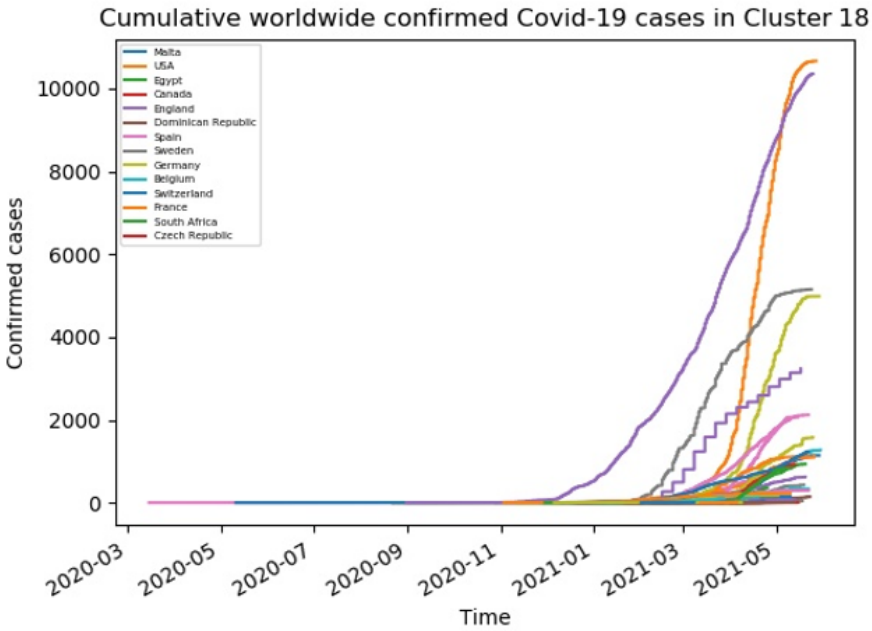
3.1.1.2 Rate of mutation in different part of world

Another piece of information that the pipeline provides is the mutation rate[8] in the different parts of the words for each cluster. The mutation rate for cluster 18 was a new variant every 11.46 minutes, and for cluster 17 mutation rate

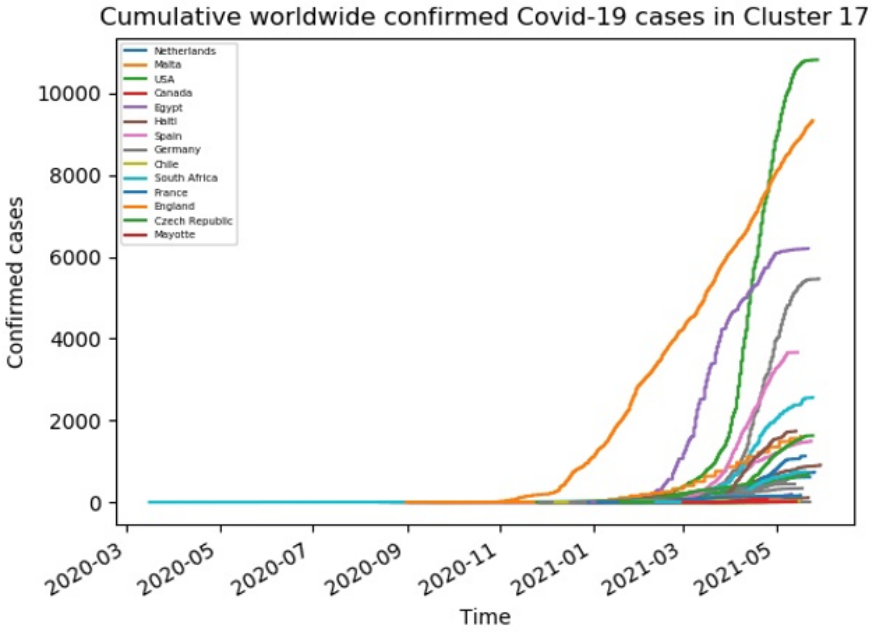
[8] Mutation rate is the probability of a specific base pair or a more DNA segment alteration over time [159].

was a new variant every 11.54 minutes. For cluster 19, the mutation rate was a new variant every 10.27 minutes.

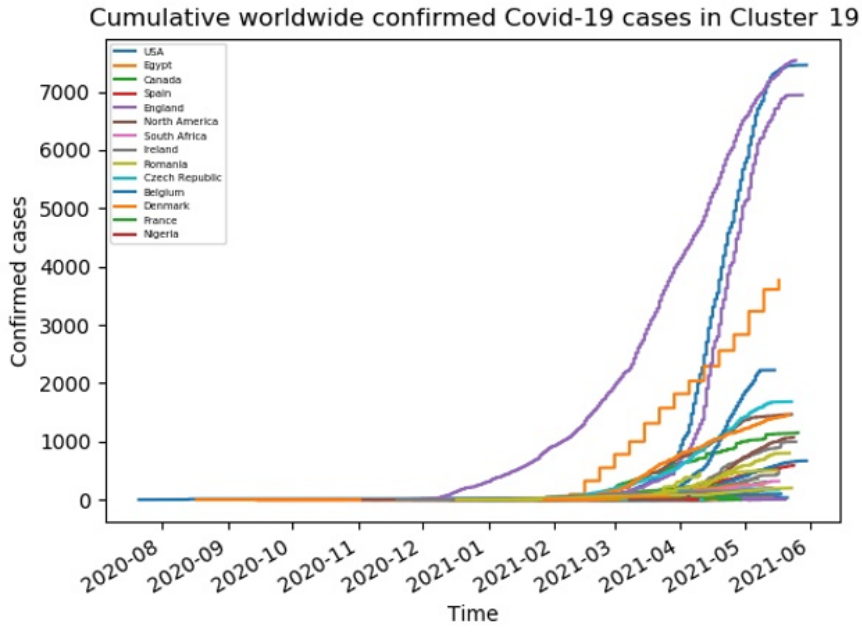
To visualize the increase in cases of infections per country, we created cumulative time charts. There is a positive relationship between the mutation rate and the number of reported infections. Cumulative time charts for these three clusters with the highest number of variants are available in Figure 3-3. All the cumulative time charts are available in Appendix B.



(a)



(b)



(c)

Figure 3-3. Cumulative worldwide confirmed cases of COVID-19 for cluster 18(a), 17(b), and 19(c)

High mutation rate in countries shows the availability of the virus in these areas, and, because of the rapid mutation, it raises concerns about the effectiveness of the vaccines in these areas.

3.1.2 Cumulative diagram

Another diagram that the pipeline generates is the cumulative time chart. This chart and the CSV report show the mutation rate worldwide and help determine the start time, mutation rate and the probability of travelling variants from different countries. For example, the following chart shows that the number of confirmed cases in England, in general, was higher in comparison to other countries and from 09-2020 to 01-2021, the growth in cases was higher than in other countries. Figure 3.4 shows the cumulative time chart for worldwide data.

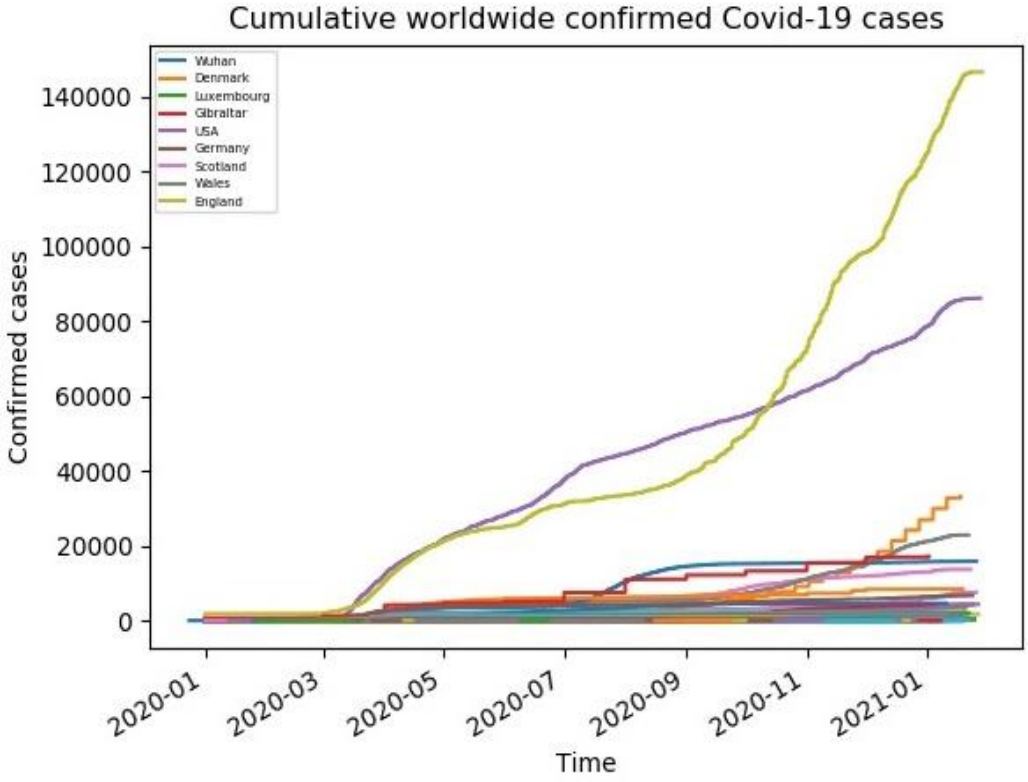


Figure 3-4. Cumulative worldwide plot for confirmed cases of COVID-19.

3.2 Studying the effect of Sequence technology in providing bias in variants

3.2.1 Aligning the sequences

This part of the pipeline generates an MSA (Multiple Sequence Alignment) file from unaligned sequences. The codes for aligning the fasta files are available in the project's GitHub[9].

[9] https://github.com/zahraav/covid.git

3.2.2 Sequence alignment viewer

Every nucleotide has a different colour in the following figure; as the figure shows, the sequence alignment viewer showed that the data generated from the previous step was aligned correctly. Therefore, the pipeline can proceed to the next step. Because the worldwide data was so large, more than 8 million sequences, Canadian data was used as a sample to check the alignment in this part. Four thousand nucleotides out of 32919 from 1000 sequences from 8432 Canadian sequences were used as a sample for checking the alignment with the sequence alignment viewer. The starting point for selecting nucleotide and sequences were chosen randomly. The sequence alignment viewer result is also available in Appendix D.

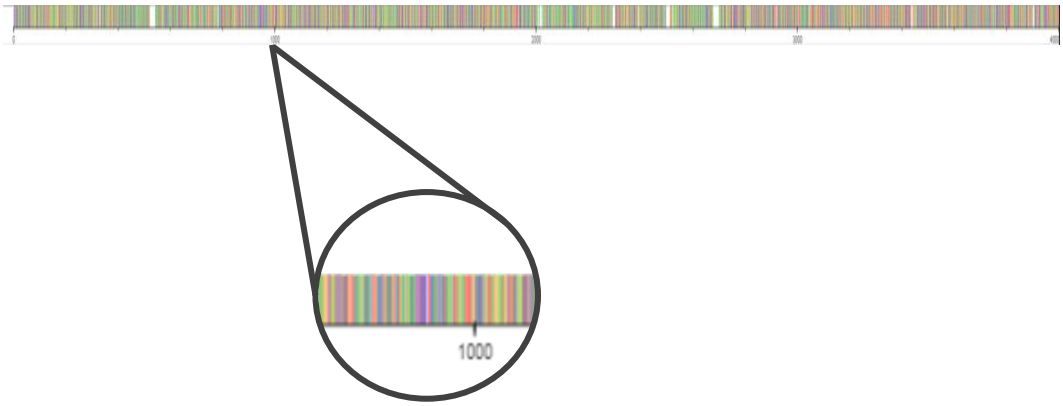


Figure 3-5. Sequence Alignment viewer for 4000 nucleotides for 1000 sequences.

Parallel lines in the magnified part of Figure 3-5 show good alignment. Each colour represents a consensus nucleotide in the consensus sequences. Modified parts for the helper functions from the sequence alignment viewer are available in Appendix A.

3.2.3 Phylogenetic tree

Other pipeline outputs are the phylogenetic tree and hierarchical clustering for the phylogenetic tree. The result for a small data set is available in Figure 3.6. Figure 3.7 presents Canada's phylogenetic tree for August 2020, and Figure 3.8 shows the hierarchical clustering for these data. The following figure uses colours to make differences in recognizing sequencing technologies.

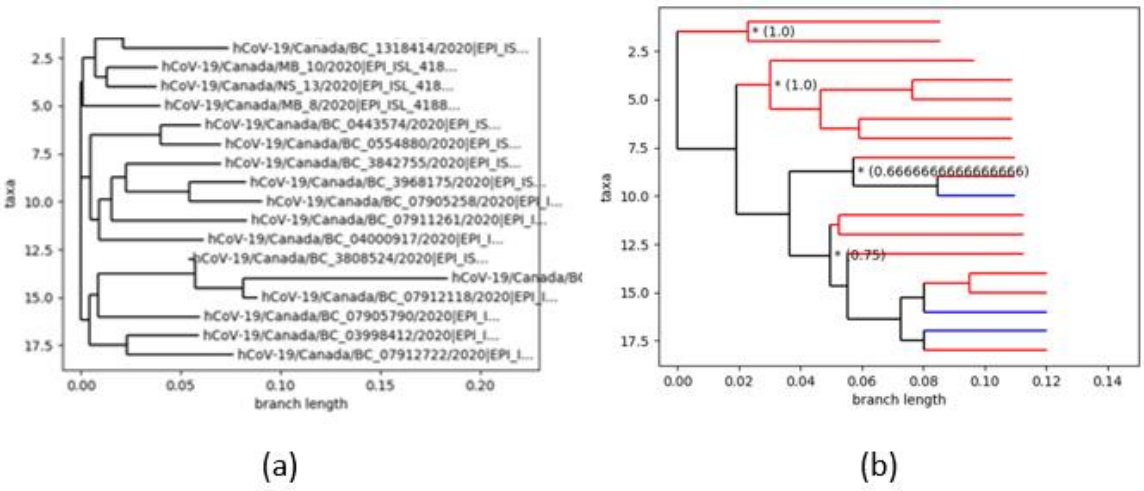


Figure 3-6. Phylogenetic tree (a) and hierarchical clustering (b) for a small sample dataset. The red lines in the hierarchical clustering are sequences sequenced by Nanopore, and the blue lines are for Illumina. When a branch has both sequencing technologies as a sub-branch, the colour of that branch is black.

In Figure 3.8 the overall ratio for Nanopore to Illumina in Canada is 2.02581755. A threshold was set to remove the branches with high distance, and this threshold was set to 100. As a result, the data was distributed in some subtrees, and only the first and second subtrees had Illumina. In Section 2.2.3, the more similar branches are connected and clustered in one group with similar features. Therefore, both sequencing technologies are expected to identify sequences belonging to one cluster. Moreover, if the sequences only

have one colour, it means that only one sequencing technology can identify that variant. If there is no bias, the data should be distributed randomly; in every subtree, we should see both red and blue, but now in the third subtree, there is no blue and the distribution in the first and second subtrees is not equal. We can conclude biased and unequal distribution results from sequencing technology.

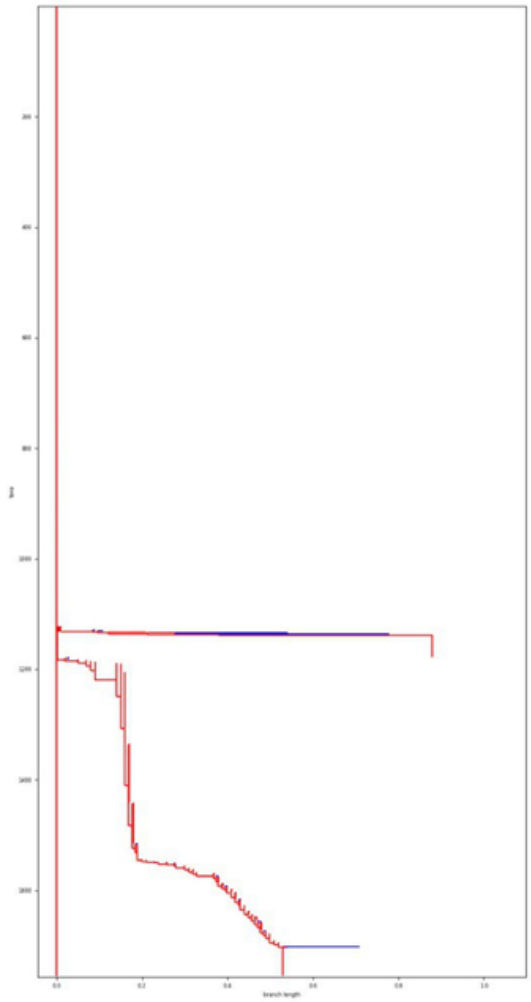


Figure 3-7. phylogenetic tree without accession ids. In these figures, the red lines are sequences sequenced by Nanopore, and the blue lines are for Illumina.

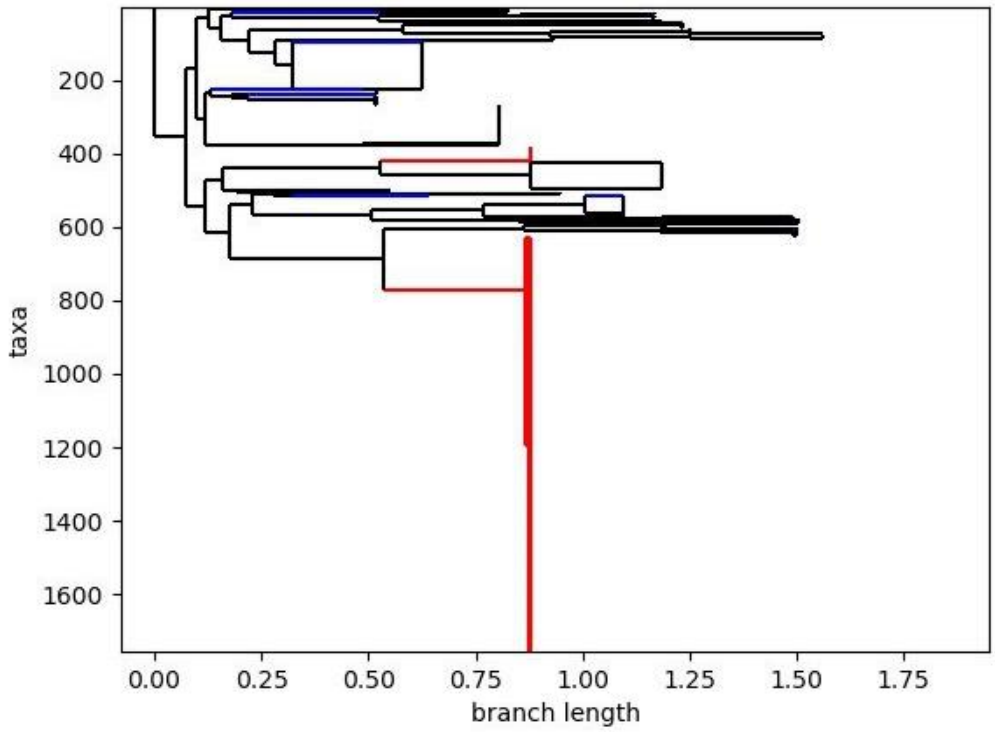


Figure 3-8. Hierarchical clustering for the phylogenetic tree. In these figures, the red lines are sequences sequenced by Nanopore, and the blue lines are for Illumina. The lower part of the tree has only Nanopore, which means Nanopore is the sequencer that only identifies these sequences, and on the upper part of the tree, the blue lines, some groups of variants only identified by Illumina.

3.2.4 Confounding factors

In addition to the previous section, the pipeline helps understand the effect of confounding factors in the study. Time and location are two confounding factors that this pipeline considers.

3.2.4.1 Time

Fisher exact test

The result of the Fisher exact test for each of the three peaks of COVID-19 is available. Data given to the pipeline are for three outbreaks from Jan 2020 until November 2021. The first outbreak was from November 2020 to February 2021, the second wave was from March 2021 to May 2021, and the third was from August 2021 to October 2021. Using Table 2.2 and the formula from Section 2.2.4 and setting the significance level to 0.01 for all data, the result for each peak is available in the following tables. Eight p-values from the results are available in Tables 3-2, 3-3, and 3-4 for each peak.

Moreover, the difference between the consensus sequences of two sequencing technologies and from the reference genome in different outbreaks was investigated to see if the time should be considered a confounding factor.

For the first outbreak, there were only 161 differences out of 33669 nucleotides; in the second outbreak, there were 207 differences out of 33669; in the third outbreak, there were 236 differences out of 33669. As we can see, there is a consistent difference in all outbreaks, which is the issue of sequencing technologies.

From the comparison of outbreaks and the p-value for nucleotides for each row, we can conclude that the result is not statistically significant, and time does not have an effect as a confounding factor.

*Table 3-2 Significance of base calling bias between Illumina and Nanopore sequencing for the **first** outbreak (**November 2020 to February 2021**). In the table, cells in bold show the significant by p-values on a specific position. For example, the first row highlights a significant over representation of A where C was expected. Indeed, the ratio of having A in position 6151 in Nanopore is 20.70%, while it is 1.272% in Illumina, resulting to a significant p-value of 4.13E-273 when comparing proportion using the Fisher exact test.*

	Count of Nucleotide in Nanopore				Count of Nucleotide in Illumina				P-Value				Reference	
position	A	C	G	T	A	C	G	T	A	C	G	T	nucleotide	Count
6151	445	1,703	1	0	221	17,141	1	1	**4.13E-273**	Ref	0.20808	1	C	18,844
31742	1	2,099	0	0	1,116	16,252	0	0	**1.48E-56**	Ref	1	1	C	18,351
7706	0	447	0	1,695	0	220	0	17,073	1	**2.74E-275**	1	Ref	T	18,768
12337	0	174	0	1,976	0	23	0	17,353	1	**1.55E-141**	1	Ref	T	19329
18785	0	0	167	1,983	0	0	58	17,319	1	1	**1.80E-111**	Ref	T	19302
28013	0	0	1,287	862	0	0	7,044	10,325	1	1	**1.14E-64**	Ref	T	11187
14865	0	0	1,702	448	0	0	17,156	217	1	1	Ref	**1.73E-277**	G	18,858
2129	0	1,702	0	447	0	17,145	0	220	1	Ref	1	**2.74E-275**	C	18,847

*Table 3-3 Significance of base calling bias between Illumina and Nanopore sequencing for the **second** outbreak (**March 2021 to May 2021).** In the table, cells in bold show the significant p-value on that position. For example, the first row highlights a significant over representation of A where C was expected. Indeed, the ratio of having A in position 23014 in Nanopore is 12.36%, while it is 2.56% in Illumina, resulting to a significant p-value of 4.71E-215 when comparing proportion using the Fisher exact test.*

	Nanopore				Illumina				P-Value				Reference	
index	A	C	G	T	A	C	G	T	A	C	G	T	nucleotide	Count
23014	794	5,611	0	17	828	30,592	0	897	**4.71 E-215**	Ref	1	4.41 E-48	C	36,203
731	13	0	6,624	0	661	0	33355	2	**3.07 E-35**	1	Ref	1	G	39,979
33156	0	869	0	5,762	0	913	0	32,949	1	**5.88 E-235**	1	Ref	T	38,711
8853	0	808	0	5,823	0	813	0	33,209	1	**1.04 E-225**	1	Ref	T	39,032
21780	0	0	1	6,618	0	3	382	33,556	1	1	**2.51 E-28**	Ref	T	40,174
29450	0	283	2	6,351	0	2,352	379	31,297	1	4.64 E-17	**1.58 E-26**	Ref	T	37,648
1423	0	6,632	0	3	0	32,428	0	1,592	1	Ref	1	**9.43 E-120**	C	39060
7571	0	0	5,891	711	0	0	32,532	1,454	1	1	Ref	**1.57 E-84**	G	38,423

*Table 3-4 Significance of base calling bias between Illumina and Nanopore sequencing for the **third** outbreak **(August 2021 to October 2021)**. In the table, cells in bold show the significant p-value on that position. For example, the first row highlights a significant over representation of A where C was expected. Indeed, the ratio of having A in position 14508 in Nanopore is 1.26%, while it is 0.139% in Illumina, resulting to a significant p-value of 2.53E-06 when comparing proportion using the Fisher exact test.*

	Nanopore				Illumina				P-Value				Reference	
Index	A	C	G	T	A	C	G	T	A	C	G	T	Nucleotide	Count
14508	12	934	0	0	8	5730	0	0	**2.53E-06**	ref	1	1	C	6664
11954	17	0	929	0	14	0	5,726	0	**1.19E-07**	1	Ref	1	G	6,655
6621	0	0	0	946	0	132	0	5,606	1	**2.74E-09**	1	Ref	T	6,552
11917	0	8	0	938	0	0	0	5,739	1	**1.57E-07**	1	Ref	T	6,677
27524	883	0	63	0	5,679	0	55	0	ref	1	**4.02E-24**	1	A	6,562
25941	0	0	16	929	0	0	5	5735	1	1	**2.33E-10**	Ref	T	6,664
19099	0	0	884	61	0	0	5,687	51	1	1	Ref	**5.30E-24**	G	6,571
32100	0	906	0	40	0	5,715	0	20	1	Ref	1	**1.26E-20**	C	6621

This global analysis highlights potential bias in the characterisations of mutations that are potentially due to sequencing quality difference in base calling across both technology. However, although we tried to track all potential confounding factor such as time and location the lack of accessible data at the time of study makes a comprehensive confounding analysis not feasible. Therefore, further analysis is required to validate this observation.

3.2.4.2 Location

For this section, data for one location are needed as input. Here data from one Canadian province, British Columbia, with each sequencing technology was chosen. The Fisher exact test was used to study the effect of location, one of the potential confounding variables, in whether sequencing technology provides bias in determining SARS-CoV-2 sequencing. The formula and table for calculating the following table are available in section 2.2.4.

The result of the pipeline in this part is one CSV report and one.txt file. The nucleotide counts for each location for both sequencing technologies and the p-value were calculated. Moreover, in the txt file for each site, data for calculating the p-value is available. The input file for this section is one fasta file for British Columbia.

British Columbia is a Canadian province with different sequencing technology for studying the effect of location. The primary sequencing technology used in British Columbia was Nanopore. The data used in this part was Canadian data from 1st January 2020 to 30th November 2021 from GISAID. Table 3.5 shows the nucleotide distribution for Nanopore and Illumina and the p-value of ten random locations for British Columbia.

*Table 3-5 Nucleotide distribution for Nanopore and Illumina and the p-value of ten random locations with a p-value for **British Columbia**. Also, this table shows the nucleotide in the reference genome on that loci.*[10]

	Nanopore				Illumina				P-Value			
index	A	C	G	T	A	C	G	T	A	C	G	T
18872	517	0	0	0	313	0	0	0	1.73E-09	1	1	1
18883	517	0	0	0	313	0	0	0	1.73E-09	1	1	1
27835	562	0	0	0	302	0	0	0	1.09E-05	1	1	1
28258	0	562	0	0	0	303	0	0	1	3.12E-05	1	1
8695	0	539	0	0	0	313	0	0	1	6.33E-05	1	1
27991	0	0	561	0	0	0	303	0	1	1	0.000234	1
10167	0	0	543	0	0	0	313	0	1	1	0.000329	1
28238	0	0	0	562	0	0	0	303	1	1	1	3.12E-05
10386	0	0	0	544	0	0	0	313	1	1	1	0.000611
202	0	0	0	546	0	0	0	313	1	1	1	0.001008313

Both sequencing technologies, Illumina and Nanopore, were used in British Columbia, but the primary sequencing technology was Nanopore. The results from 33,665 in 3,278 locations show that the p-value was less than 0.05, which means they are

[10] Because the data in this table is randomly selected, and the p-value for 91 percent of the data is one, this table was expected to show only data with a p-value equal to 1.

statistically significant (32.78 percent). The rest, in 30,387 loci, was not statistically significant (67.22 percent).

3.2.5 Gap in quality of sequencing in two sequencing technologies

MultiQC was used to summarize and compare the gap between the qualities of sequencing technologies. The result of FastQC for comparing the quality of sequencing in Illumina and Nanopore for British Columbia sequences is available in Appendix E. All 106 sequences available in the SRA database were downloaded using the SRA toolkit. Nanopore sequenced 86 sequences from these 106 sequences, and Illumina sequenced 20. The quality control checks summary and accession numbers for all the Nanopore and Illumina sequences is available in Appendix E.

The summary and results in the following show that the quality of sequences by Nanopore is lower than the quality of sequences by Illumina. Nanopore and Illumina are two widely used sequencing technologies during the pandemic. The quality score for sequencing reads is a measurement that indicates the accuracy of the base calls in sequences [139, 140] The quality score for Illumina is higher than Nanopore's since Illumina the error rate in Illumina is lower than Nanopore. The quality score range for Nanopore is between 10-20 [141], and for Illumina is 30 and above [142]. Row data was unavailable for the sequencing, so I worked on the fasta file.

Mean quality scores

The sequence quality histogram shows the mean quality read for each base position in the read [130]. A quality score above 20 is considered good quality. Among 86 sequences from Nanopore, 64 of them had good quality, and 20 of them had low quality. The mean quality read for these 20 sequences was less than 5. The quality score for Illumina was high for the 20 sequences, with a minimum of 26. The quality read for Nanopore and Illumina is shown in Figures 3-9 and 3-10.

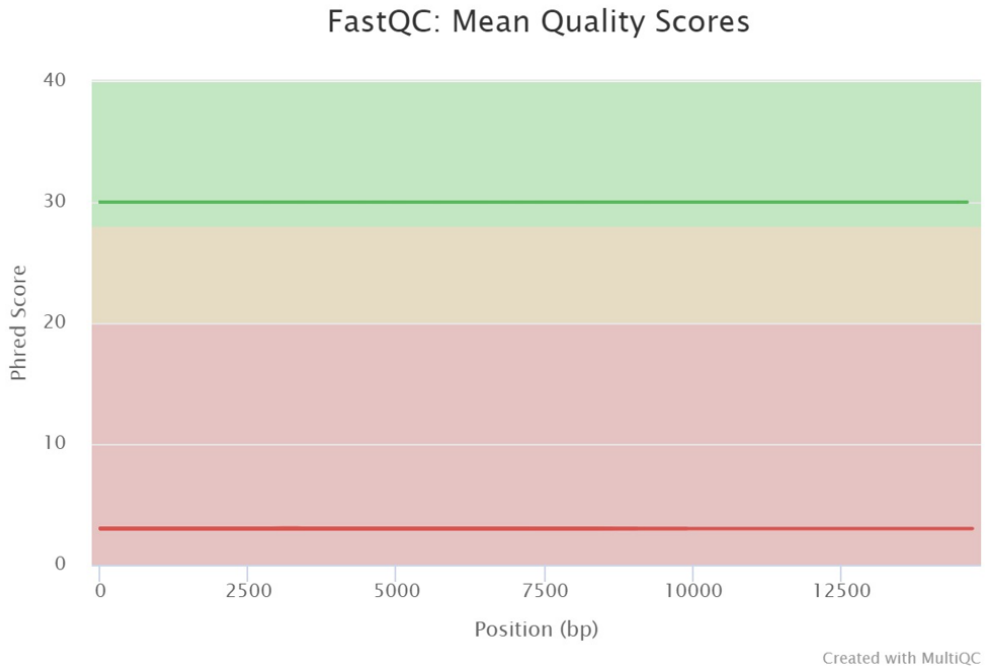


Figure 3-9. Mean quality score for 86 sequences from British Columbia sequenced by Nanopore. The red lines represent 22 sequences, all of which have a mean quality score of 3, and the green lines represent 66 sequences, which all have a mean quality score of 30.

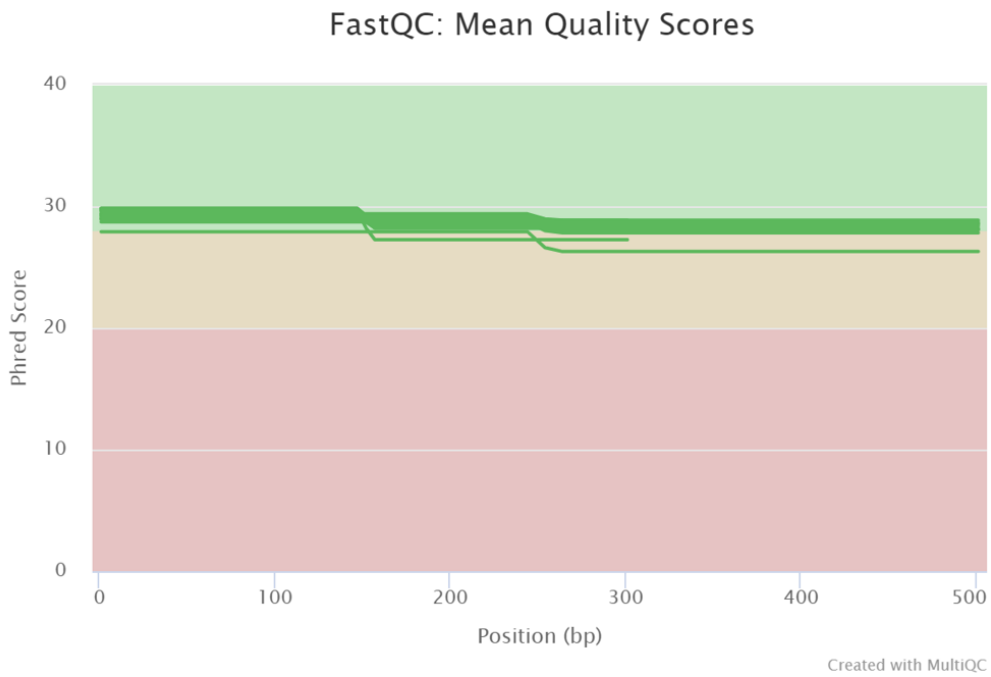


Figure 3-10. Mean quality score for 20 sequences from British Columbia sequenced by Illumina. All the sequences have a good quality score, and each green line in the figure shows the mean quality score for one sequence. The minimum mean quality score among all 20 sequences was 26.28.

Per base sequence content

In a random library, it is expected to have an equal portion of each nucleotide; but the results for the sequences do not follow the pattern because amplicon methods for SARS-CoV-2 were used for both sequencing technologies. Figures 3-11 and 3-12 show the per-base sequence content for Nanopore and Illumina.

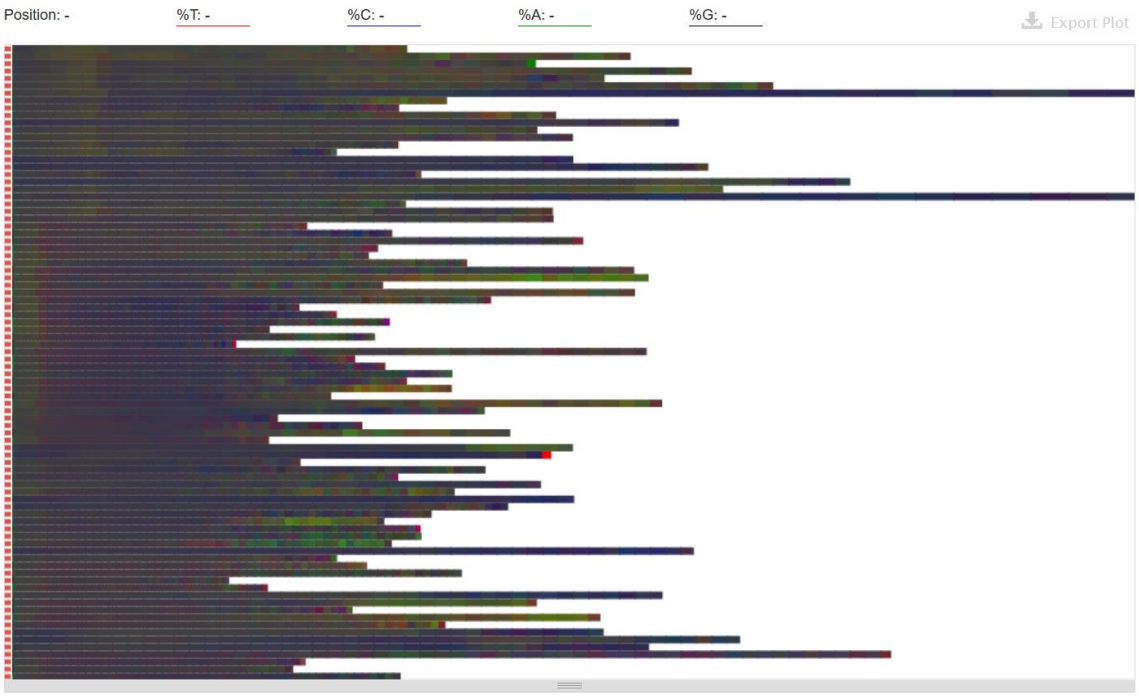


Figure 3-11. Per base sequence content for 86 sequences from British Columbia sequenced by Nanopore. In this figure, nucleotide T showed with the colour red, C with blue, A with green, and G with black.

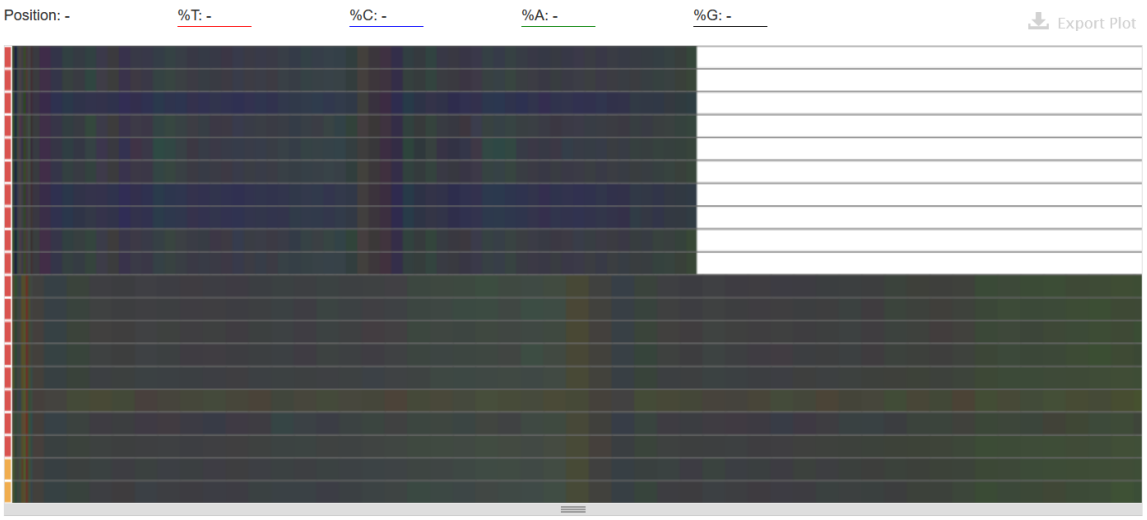


Figure 3-12. Per base sequence content for 20 sequences from British Columbia sequenced by Illumina. In this figure, nucleotide T showed with the colour red, C with blue, A with green, and G with black.

Per sequence GC content

This plot shows the distribution of GC content in the length of all sequences. GC-content measures the proportion of G and C bases in four total bases: adenine, guanine, cytosine, and thymine in DNA and uracil in RNA. The red lines show the GC content distribution in the samples. The sharp peaks show contamination, highly over-expressed genes, or biased subset [129, 143, 144, 145, 146, 147]. The module was predicted to fail due to using amplicon methods for SARS-CoV-2. The result per sequence GC content for Nanopore and Illumina are shown in Figure 3-13 and 3-14.

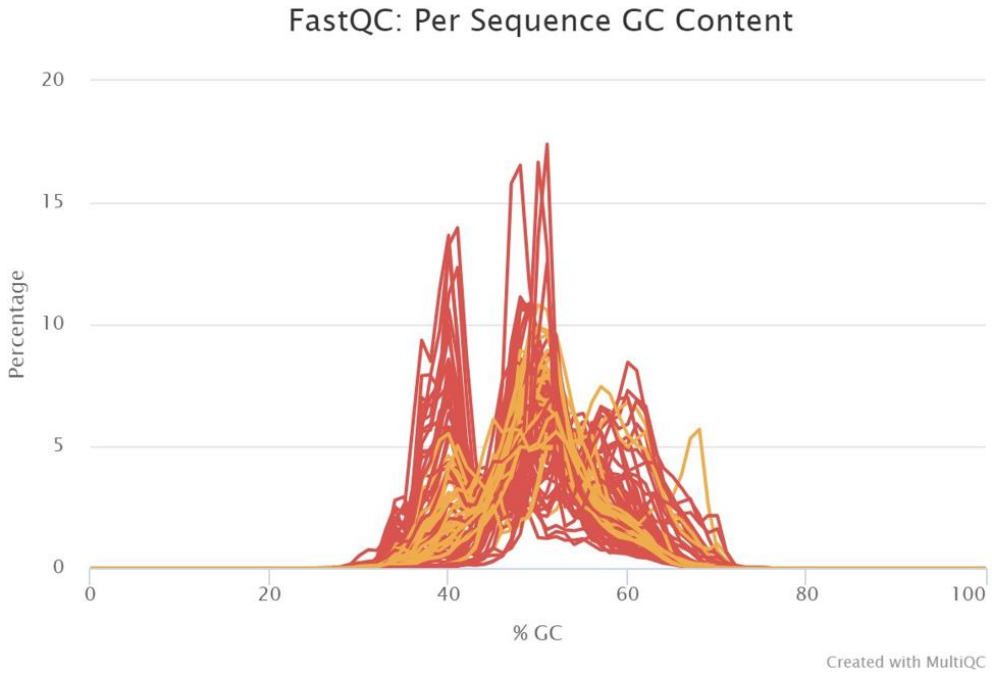


Figure 3-13. Per sequence GC content for 86 sequences from British Columbia sequenced by Nanopore. The red lines represent the sequences that failed this module, and the orange line represents sequences with warnings. 28 out of 86 sequences showed a warning, and 58 failed the module.

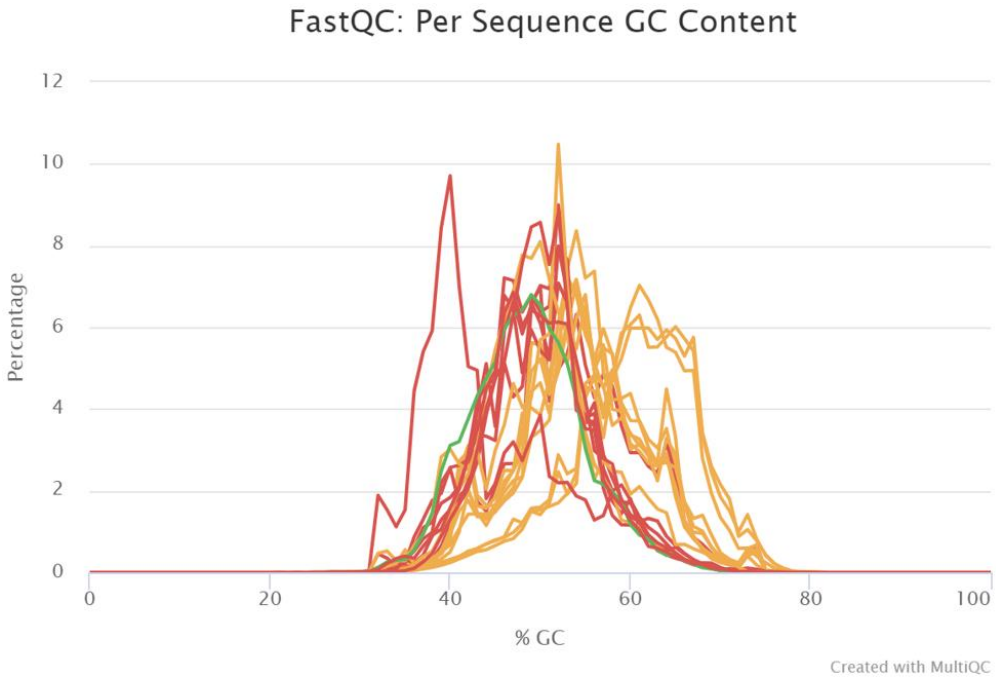


Figure 3-14. Per sequence GC content for 20 sequences from British Columbia sequenced by Illumina. The green lines represent the sequence that passed the module, the red lines show failure, and the orange line represents sequences with warnings. One out of 20 sequences passed this module, 11 showed warnings, and eight failed.

Per base N content

This plot shows the ability of the pipeline to interpret base calls. Based on the following plots, there were no uncalled bases, and both sequencing technologies passed the module [129, 143, 144, 145, 146, 147]. Figures 3-15 and 3-16 show the per base N content for Nanopore and Illumina.

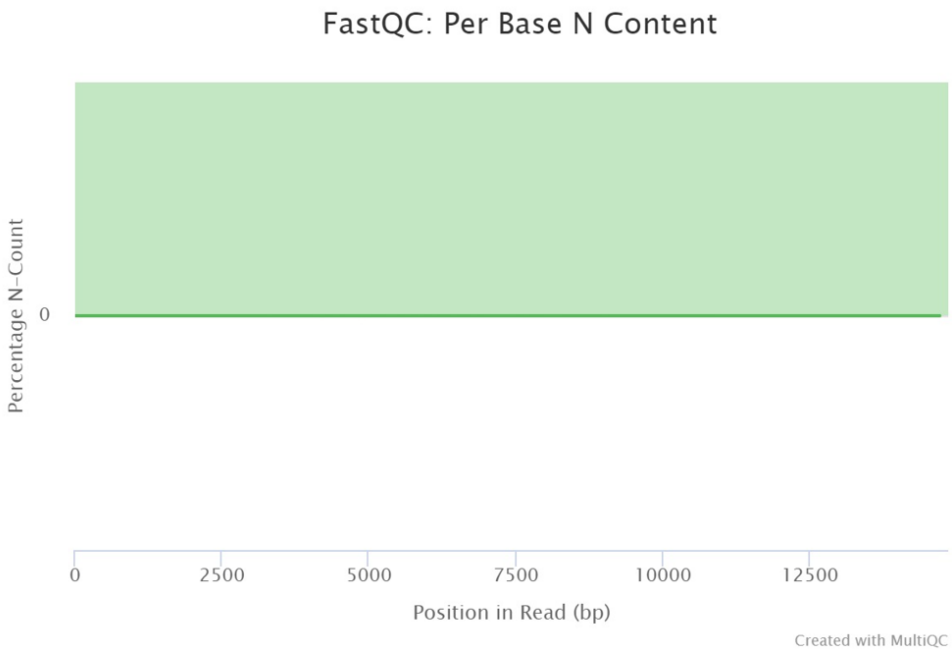


Figure 3-15. Per base N content for 86 sequences from British Columbia sequenced by Nanopore. All 86 sequences passed the module. The green line, which shows the N content, is near zero.

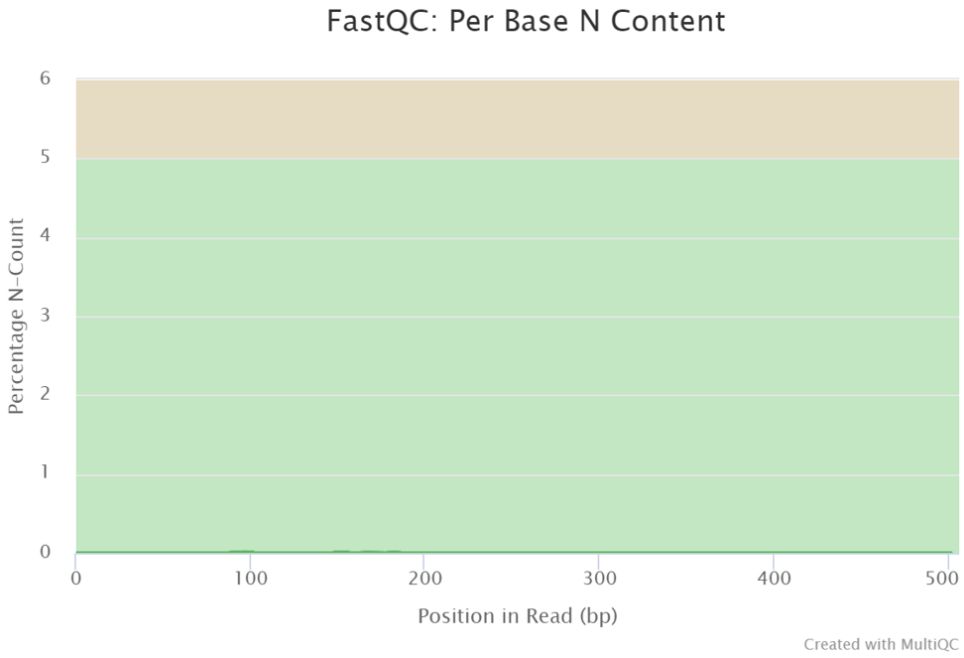


Figure 3-16. Per base N content for 20 sequences from British Columbia sequenced by Illumina. All 20 sequences passed the module. The green line, which shows the N content, is near zero.

Sequence length distribution

The distribution of component sizes used in the analysis is displayed in this module. Figure 3-17 shows the sequence length distribution for Nanopore. As it is apparent in Figure 3.18, All the sequences Illumina sequenced had the same length.

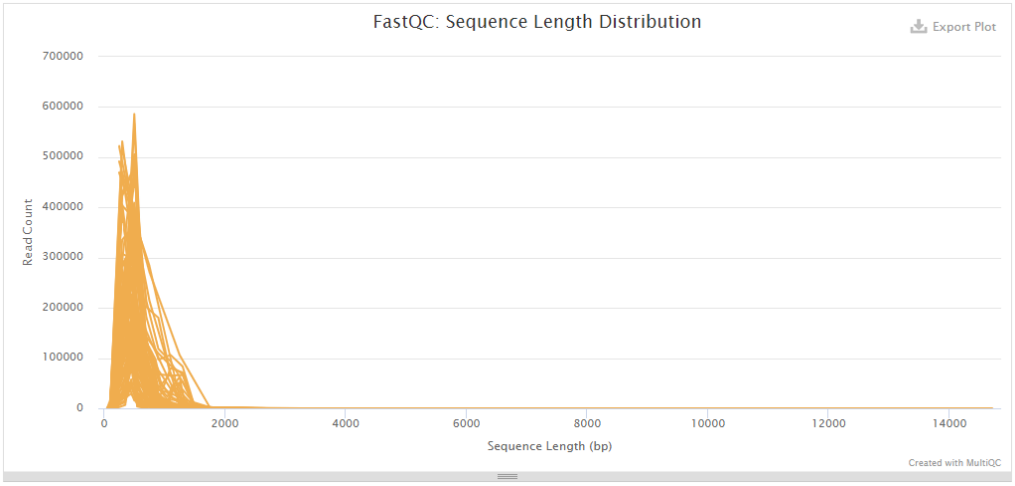


Figure 3-17. Sequence Length Distribution for 86 sequences from British Columbia sequenced by Nanopore.

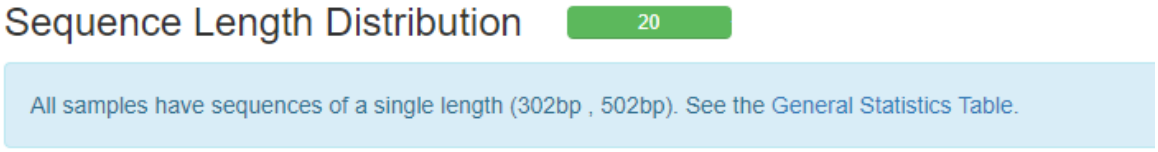


Figure 3-18. Sequence Length Distribution for 20 sequences from British Columbia sequenced by Illumina.

Sequence duplication levels

Sequence duplication levels measure how much the sequences are unique in the library. A low duplication level shows a high level of coverage; however, a high duplication level can be a sign of PCR amplification or a low starting material. In these sequences, the sequence duplication level was expected to be high because of the usage of amplicon sequencing methods [129, 143, 144, 145, 146, 147]. Figures 3-19 and 3-20 show the sequence duplication levels in Nanopore and Illumina.

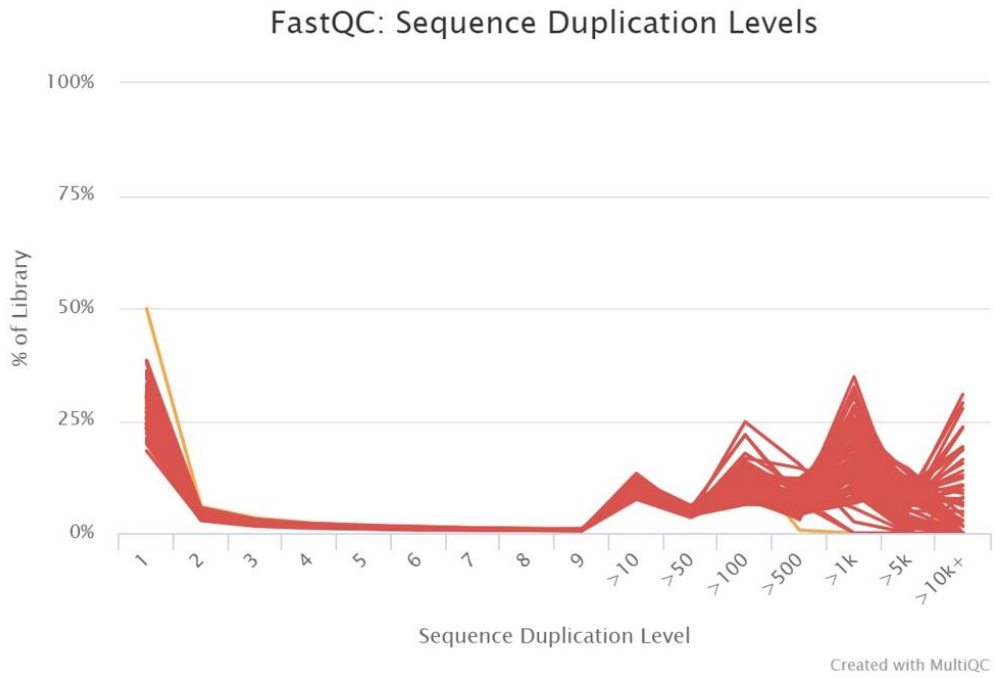


Figure 3-19. Sequence duplication levels for 86 sequences from British Columbia sequenced by Nanopore. The green lines represent the sequence that passed the module, the red lines show failure, and the orange line represents sequences with warnings. One out of 86 sequences showed warnings, and the rest failed the module (therefore there is no green line in this Figure).

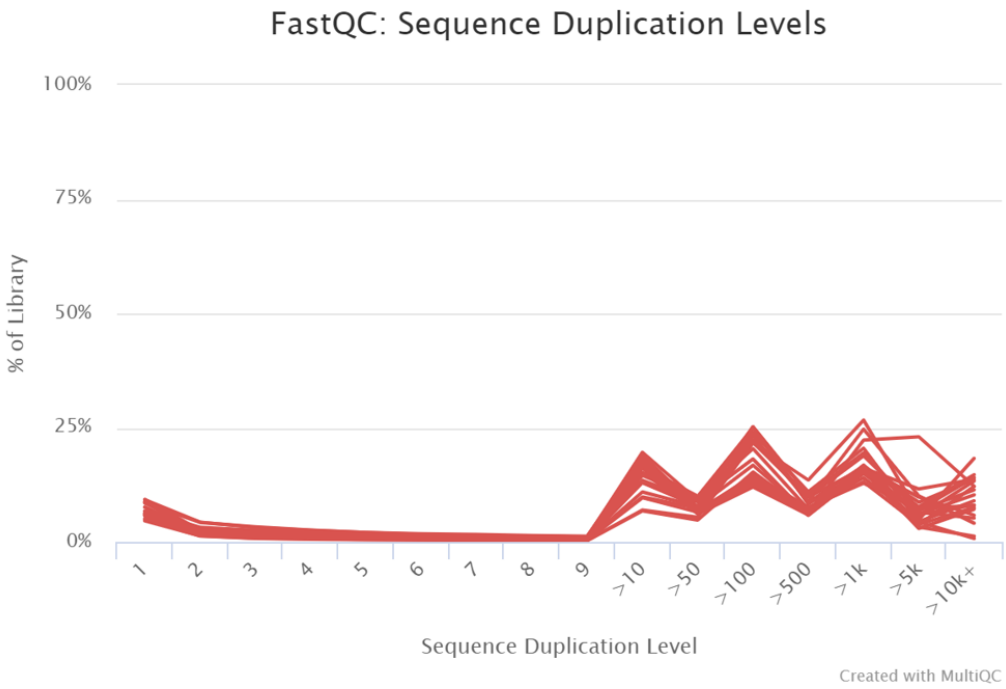


Figure 3-20. Sequence duplication levels for 20 sequences from British Columbia sequenced by Illumina. The green lines represent the sequence that passed the module, the red lines show failure, and the orange line represents sequences with warnings. All 20 sequences failed the module (Therefore there is no green line in this Figure).

Overrepresented Sequences

The overrepresented module shows sequences that appeared more than expected in the reads, which can signify that it is highly significant, or the library is contaminated. Moreover, if the per-sequence GC content module does not have a good result, overrepresented sequences aid in identifying the problem's source. An overrepresented sequence should appear in at least 0.1 percent of the reads [129, 143, 144, 145, 146, 147].

In the following figures, Figure 3-21 for Nanopore and Figure 3-22 for Illumina, as it appears, many overrepresented sequences exist because of amplicon methods. The results were also expected to be not good from the report of the sequence duplication levels modules for both sequencers.

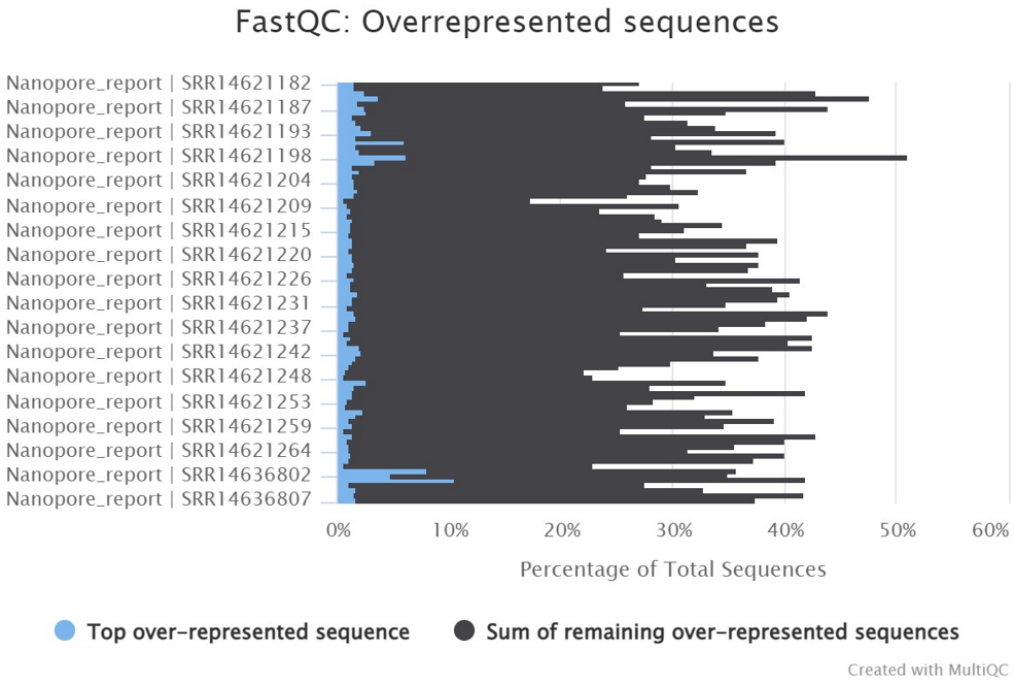


Figure 3-21. Overrepresented sequences for 86 sequences from British Columbia sequenced by Nanopore. The blue parts in the figure show the percentage of top-overrepresented sequences, and the black parts show the percentage of the sum of remaining over-represented sequences.

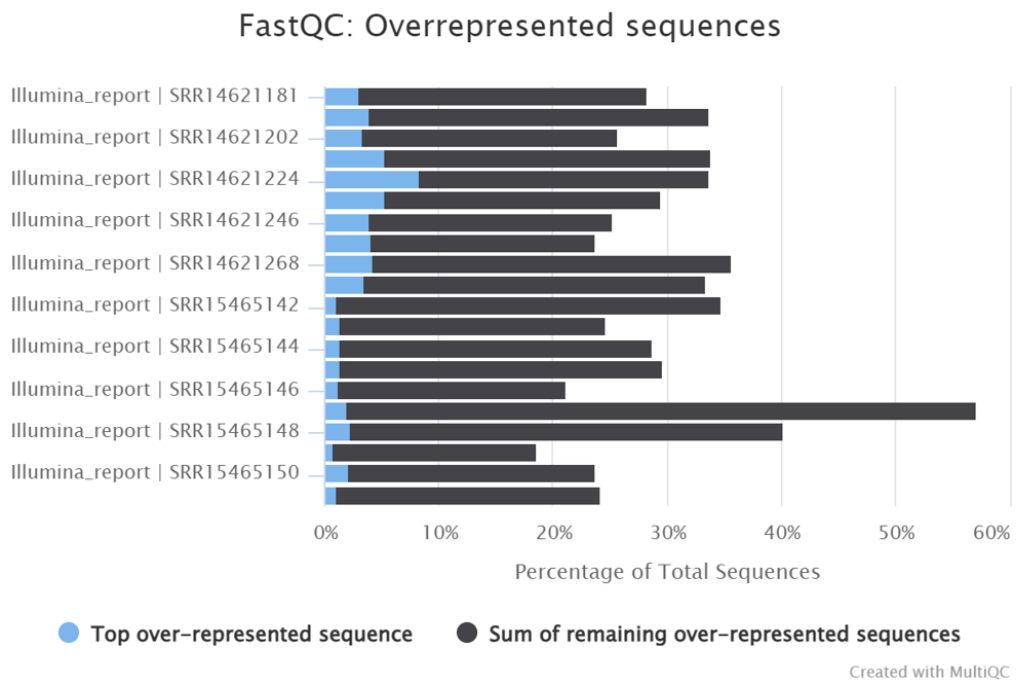


Figure 3-22. Overrepresented sequences for 20 sequences from British Columbia sequenced by Illumina. The blue parts in the figure show the percentage of top-overrepresented sequences, and the black parts show the percentage of the sum of remaining over-represented sequences.

3.2.6 Relationship between sequencing technologies, consensus nucleotides and ARTIC protocol[11]

Exploring the relationship between sequencing technologies, their protocol, and consensus nucleotides was the next goal of the pipeline. Therefore, the pipeline takes an MSA file and generates a CSV report for sequences with IUPAC codes. From the CSV report, some diagrams are generated to visualize a broad and thorough picture of this relationship.

[11] ARTIC protocol is a sequencing protocol used during the COVID-19 pandemic; the explanation of this protocol is available in section 1.4.

In this section, sequences from January 2020 to 18 February 2021 were used for all countries worldwide, provided in the GISAID database. Data from multiple sequence alignment was saved into a dictionary. Then in the following steps, the data were analyzed.

Table 3-6 and Figure 3-23 show the relationship between sequencing technology and consensus nucleotides; the number of consensus nucleotides in Illumina is distinctly more than in Nanopore but the proportion of consensus nucleotide per sequencing technology are almost the same. Forty-five thousand thirty-two sequences from 94059 sequences sequenced by Nanopore have these letters, and 148486 out of 308972 sequences sequenced by Illumina also have the letters. It was expected to have more results from Illumina because of the error rate; data from Illumina is considered more accurate in contrast to consensus nucleotides extracted from Nanopore, mostly considered error, and ignored.

Table 3-6. Relationship between distribution of consensus nucleotides and sequencing technologies.

Sequencing Technology	Number of sequences with consensus nucleotides	All sequences	Proportion of consensus nucleotide	Percentage of consensus sequences found with the sequencer in all consensus nucleotides
Nanopore	45,032	94,059	47.8763	23.2701
Illumina	148,486	308,972	48.0580	76.7298

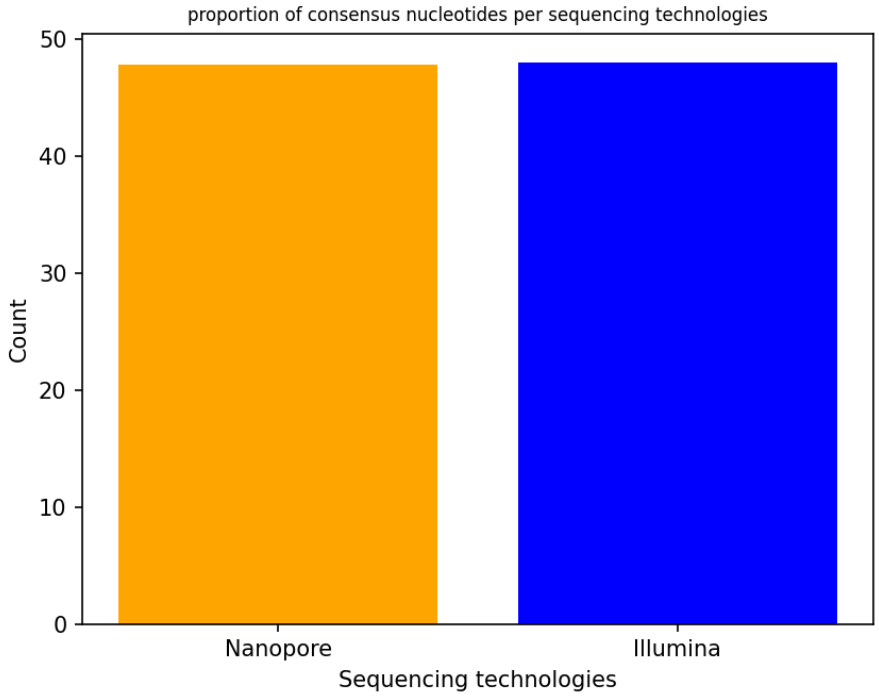


Figure 3-23. Relationship between the proportion of consensus nucleotide and sequencing technologies. The proportion of consensus nucleotides for both sequencing technologies was similar.

In general, there were 246475 consensus nucleotides found in the sequences, and among all found codes letter Y (C or T) with 100937 repetitions was the most common. Figure 3-24 and Table 3-7 show the distribution of consensus nucleotides in sequences.

Table 3-7 Distribution of IUPAC codes in different sequence in MSA file

Letter	Count
Y	100937
S	8832
W	16215
K	65478
R	35062
M	18643
H	505
D	393
B	201
V	209

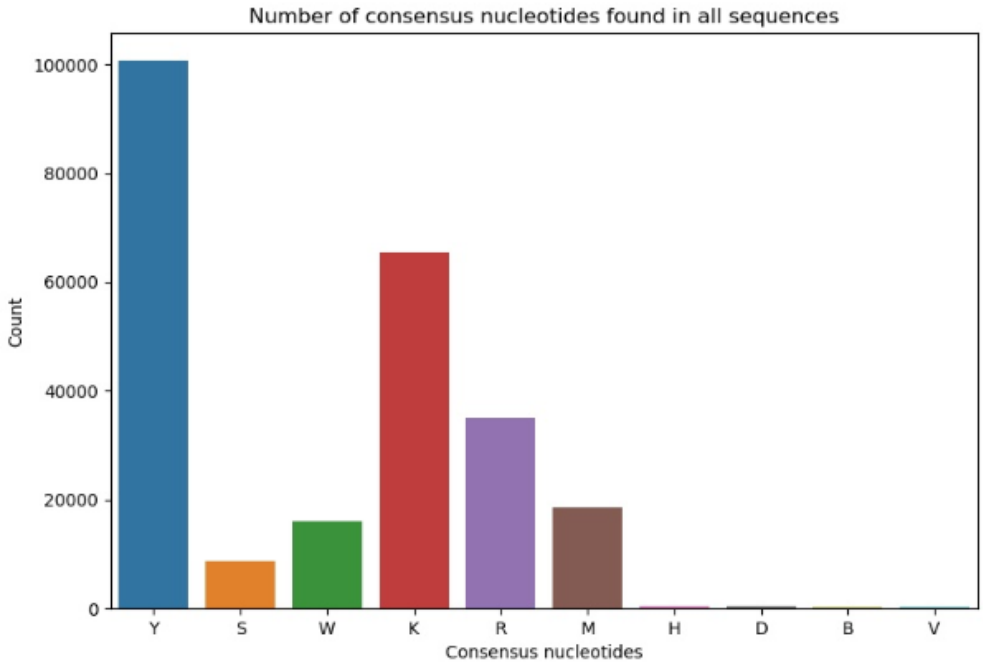


Figure 3-24. Distribution of consensus nucleotides in different sequences in MSA file. Letter Y has the highest amount, with 100,937 and letter B, with 201, has the lowest amount among all the letters.

As shown in Table 3-8 and Figure 3-24, the letter Y from IUPAC codes has the highest amount compared to other annotations because it is a point mutation called transition. *Transition* is a point mutation that changes between two purines and also between two pyrimidines [148].Moreover, it has a higher repetition among point mutations [149]. The second highest bar in Figure 3-24 belongs to the letter K. G to T mutation is one of the most frequent errors in Illumina [150], since Illumina recognizes 76.7298 percent of consensus nucleotides in this study, Figure 3-23 and Table 3-6, therefore, it was expected to see a high amount of this letter in the results, which is the IUPAC code for G or T. The third highest bar in Figure 3-24 is for the letter R (A-G); the other transition, similar to Y, was expected to have a high amount.

Knowing the distribution of consensus sequences in different continents to better understand the relationship between sequencing technologies and consensus sequences in different continents is crucial. The distribution of consensus nucleotide found worldwide is shown in Tables 3-8, and Figure 3-25 visualizes the information in Table 3-8. In February 2021, Europe, with 117495, had the highest number of consensus nucleotide. North America, with 83744, has the second rank, and the third rank belongs to Oceania, with 33562 letters.

Table 3-8 Distribution of consensus nucleotide worldwide.

Country	IUPAC codes
Europe	117495
North America	83744
Asia	9419
Oceania	33562
South America	671
Africa	1584

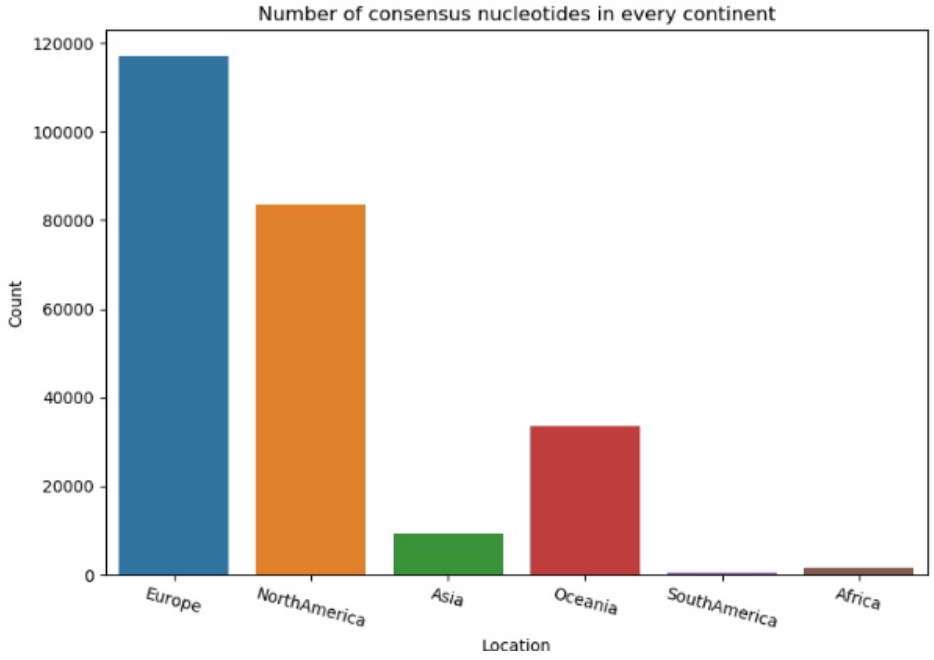


Figure 3-25. Distribution of consensus nucleotide in different continents. Europe, with 117,495, has the highest consensus nucleotide in the sequences, and South America, with 671, has the lowest amount.

Figure 3-26 shows the combination of data in Figures 3-24 and 3-25, which is the distribution of consensus nucleotides in different continents. The distribution of

consensus nucleotide in this Figure, Figure 3-26, also follows the rule that was explained in the previous figures, in which the letters Y, R and K have the most considerable amount among other nucleotides.

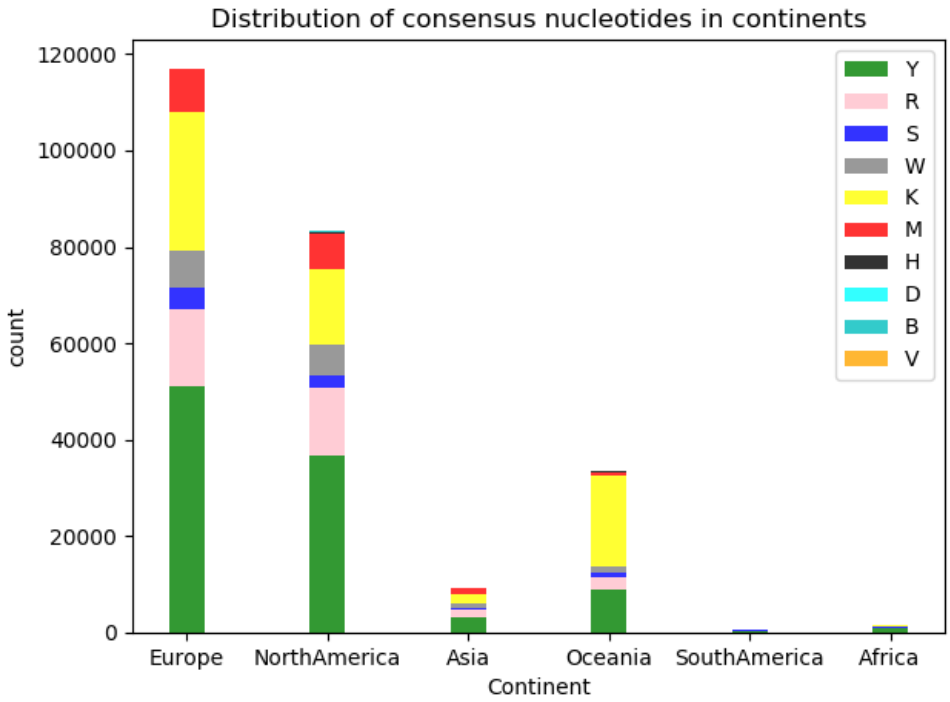


Figure 3-26. Relationship between various consensus nucleotides on different continents.

Also, to better understand and remove the effect of infection count in different regions, Figure 3-26 shows the ratio plot for consensus nucleotides in the different continents generated by the pipeline.

Table 3-9 Distribution of consensus nucleotides in different continents.

Consensus nucleotides / Region	M	R	W	S	Y	K	V	H	D	B
Europe	**8976**	**16248**	**7666**	**4475**	**51096**	**28893**	**20**	**33**	**58**	**29**
North America	**7380**	**14056**	**6181**	**2750**	**36777**	**15660**	**170**	**424**	**235**	**110**

Asia	**1297**	**1702**	**904**	**416**	**3130**	**1938**	**6**	**14**	**9**	**2**
Oceania	**878**	**2599**	**1306**	**1080**	**8872**	**18631**	**11**	**34**	**90**	**60**
South America	**32**	**212**	**61**	**63**	**206**	**95**	**1**	**0**	**0**	**0**
Africa	**80**	**245**	**96**	**47**	**852**	**261**	**1**	**0**	**1**	**0**

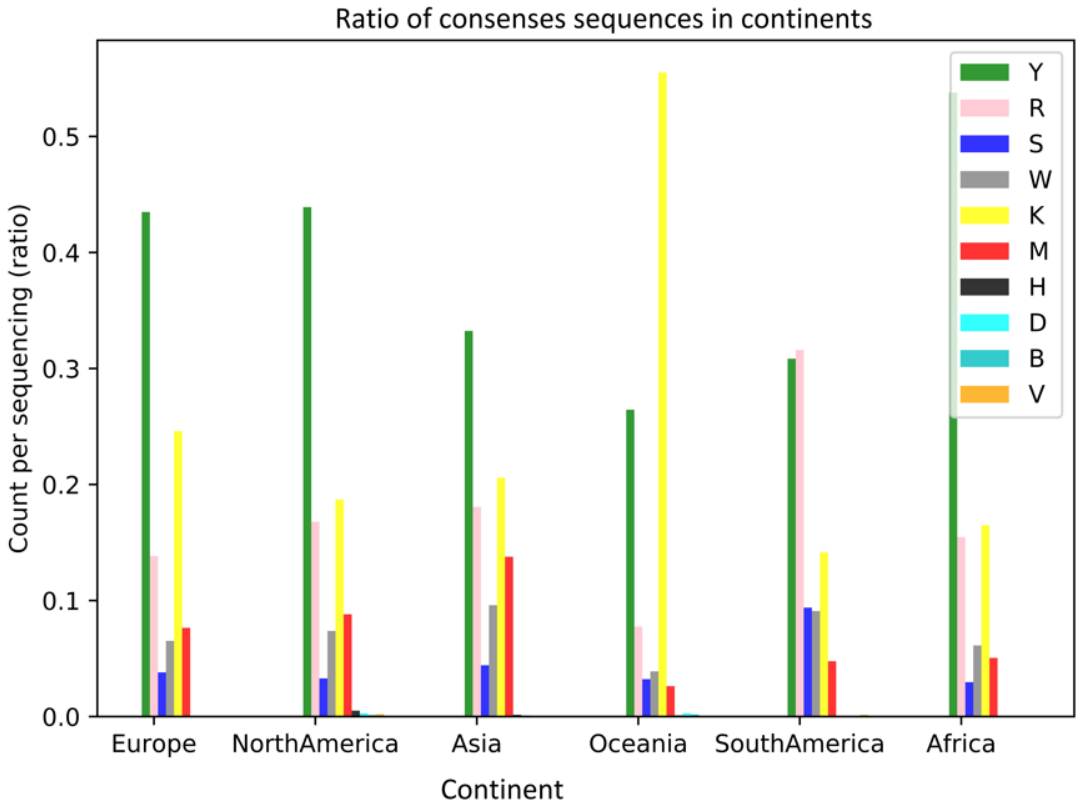


Figure 3-27. Percentage of distribution of consensus nucleotides in different continents.

After investigating the sequences, it was apparent that 26847 and 32442 with 7431 and 2929 repetitions are two indices in which the consensus nucleotides were more observed than other indices; the length of sequences is 35561. In sequences, the eight indices in which the repetition was more than 1000 are shown in Table 3-10.

Table 3-10 Eight indices with the repetition of consensus nucleotides more than 1000.

Indices	Repetition of consensus nucleotides	Most common consensus nucleotides on the location	Reference Genome
26847	7431	K (G or T)	T
32442	2929	M (A or C)	G
27244	1608	K (G or T)	C
11245	1174	Y (C or T)	G
19549	1117	R (A or G)	G
30187	1091	Y (C or T)	A
1499	1043	Y (C or T)	T
30547	1039	Y (C or T)	A

The result of analyzing data for repetition of more than 1000 (Table 3-10) shows that the letter K had more repetition in the two most remarkable indices, and after that, the letter Y had the most repetition; Figure 3-27 shows the distribution of consensus nucleotides in each of the eight indices.

Table 3-11 Eight indices and distribution of consensus nucleotide in each of them.

Indices	M	R	W	S	Y	K	V	H	D	B	-
26847	0	0	0	0	1	7428	0	0	0	1	0
27244	0	1	0	4	0	1602	0	0	0	0	0
30547	2	0	0	2	1034	0	0	0	0	0	0
32442	2875	0	2	10	12	0	7	19	0	3	0
1499	0	0	0	0	1042	0	0	0	0	0	0
11245	1	0	0	0	1172	0	0	0	0	0	0
19549	0	1116	0	0	0	0	0	0	0	0	0
30187	1	0	0	0	1089	0	0	0	0	0	0

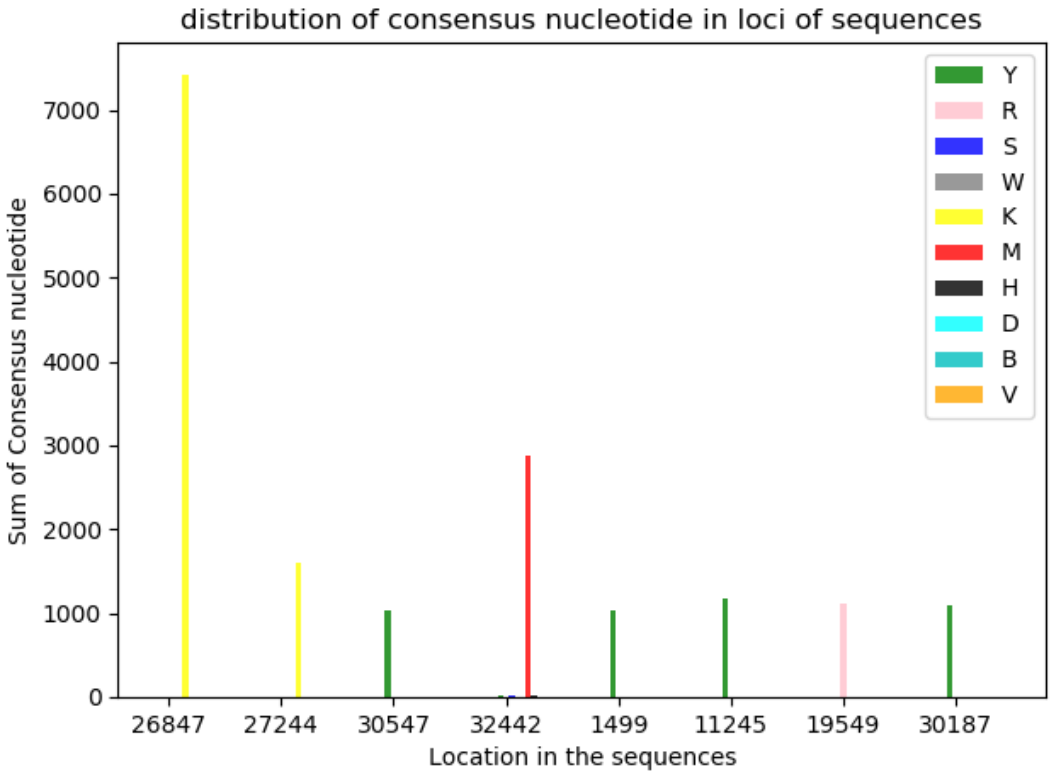


Figure 3-28. Distribution of consensus nucleotides in each of the eight indices with more than 1000 consensus nucleotides.

Afterwards, the sequences with most consensus nucleotides were studied among worldwide data, to find an importance in genomic regions, which, because the result was so close to each other, did not have a meaningful conclusion. Figure 3-28 shows the result for this part of the study.

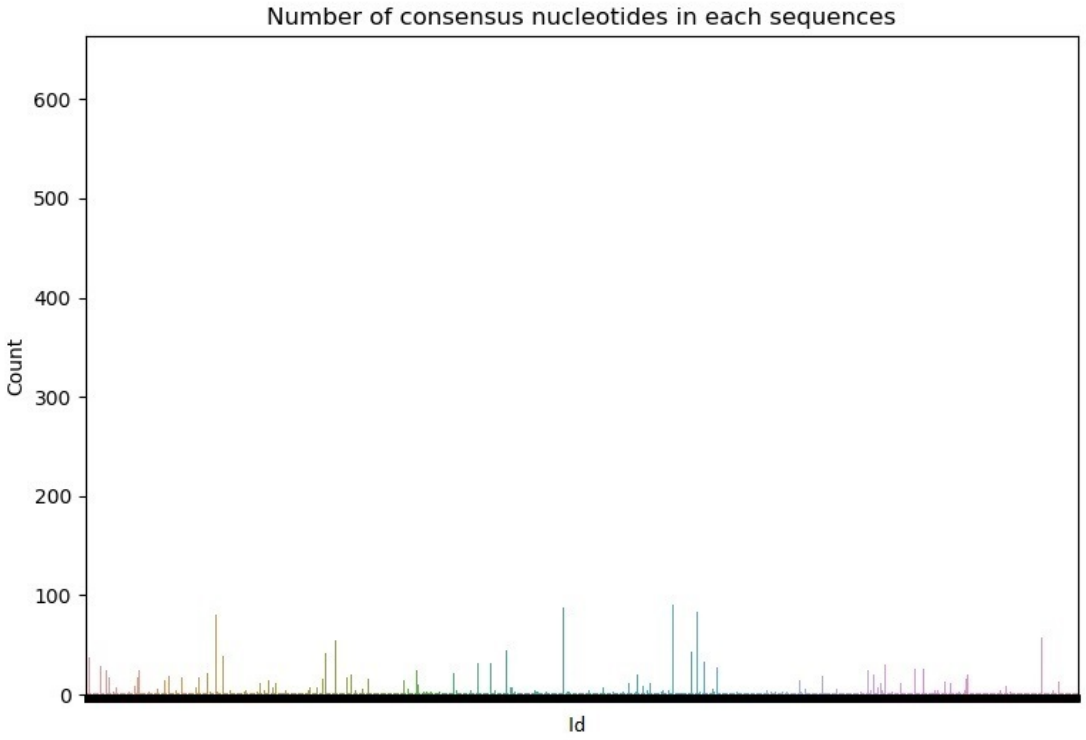


Figure 3-29. Distribution of consensus nucleotides in sequences in worldwide data. We could not draw a meaningful conclusion due to the large amount of data shown in this Figure.

Further, the relationship between sequencing technology and consensus nucleotides was studied. Table 3-12 shows that the letter Y (C or T) is more repetitive among consensus nucleotides. Also, the number of consensus nucleotides found in Illumina is higher than in Nanopore, Figure 3-30.

Table 3-12 Relationship between sequencing technology and consensus nucleotides. For example, the letter Y was 1366 times appeared in the sequences sequenced by Nanopore and 97527 times in sequenced that Illumina sequenced.

Letter	Nanopore	Illumina	Proportion letter for Nanopore	Proportion letter for Illumina
Y	1366	97527	1.3812	98.6187
S	278	8346	3.2235	96.7764
W	521	15100	3.3352	96.6647
K	998	63741	1.5415	98.4584
R	621	33628	1.8131	98.1868
M	928	17303	5.0902	94.9097
H	1	410	0.2433	99.7566
D	4	352	1.1235	98.8764
B	0	184	0	100
V	2	197	1.005	98.9949

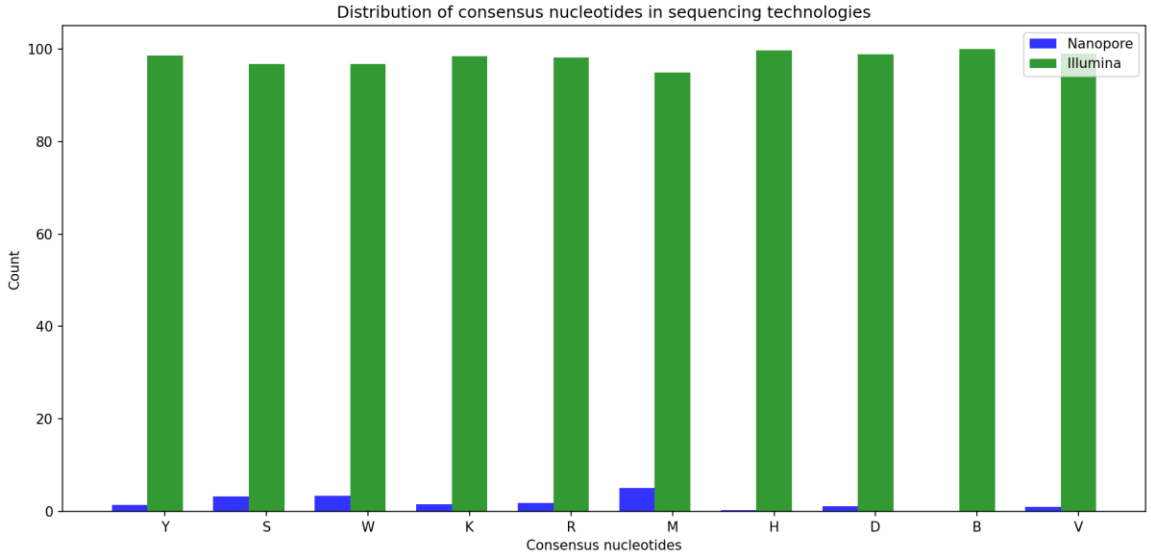


Figure 3-30. Relationship between sequencing technology and consensus nucleotides.

After generating data and plotting the above diagrams based on them, the last diagram is generated, a categorical plot to study the roles of the protocol in the availability of consensus nucleotides in sequences sequenced by different sequencing technologies in different continents.

Protocols have five different groups, ARTIC V1, ARTIC V2, ARTIC V3, ARTIC[12] protocol without mentioning the version in the given metadata file, and other protocols. The distribution of protocols is available in the Table 3-13 and visualized in Figure 3-31. Data for this part was from January 2020 to November 2021.

Table 3-13 Usage of protocols in different continents.

Protocol	Total in all continents
ARTIC	1131
ARTIC v1	199
ARTIC v2	8
ARTIC v3	8444
Other	58772

[12] The differences between these protocols are in alternative primers used in versions 2 and 3 and improved depth of coverage, which is explained in section 1.4.

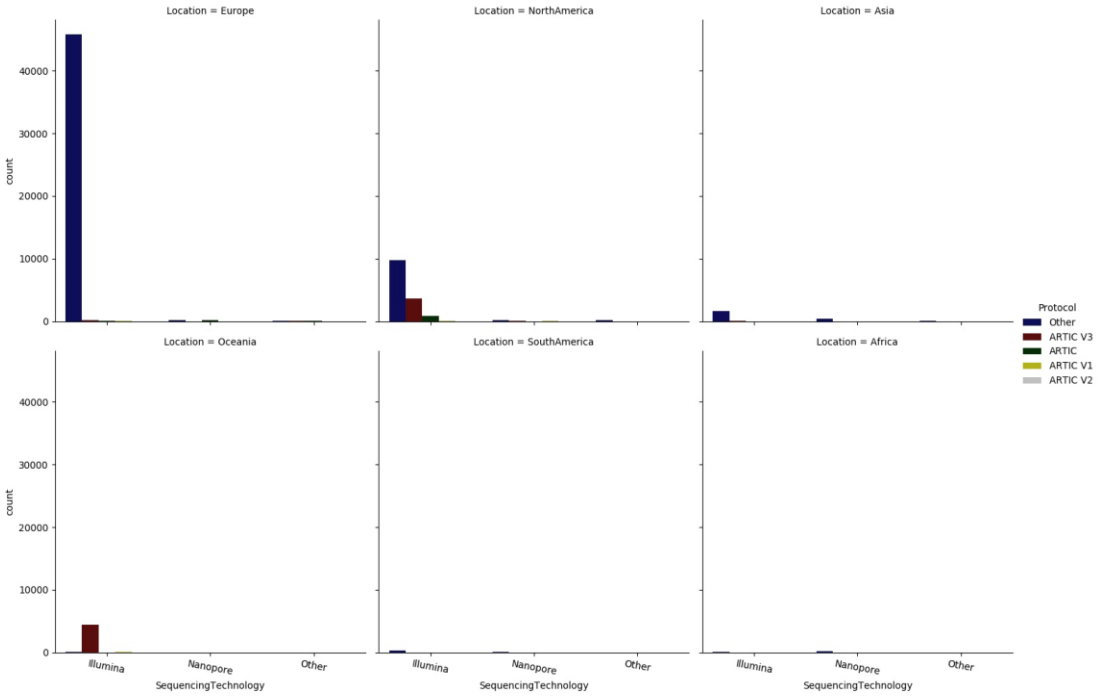


Figure 3-31. Usage of protocols in different continents. Europ has the most significant amount of usage for other protocols. Because of the large number of other protocols, the ARTIC is not apparent in this figure.

The second plot ignored the other protocol, and only the ARTIC protocol was shown in Figure 3-32. Oceania had the highest rank in using the ARTIC protocol and then North America; therefore, we could not observe any relationship between the protocol and IUPAC annotations and the sequencing technologies.

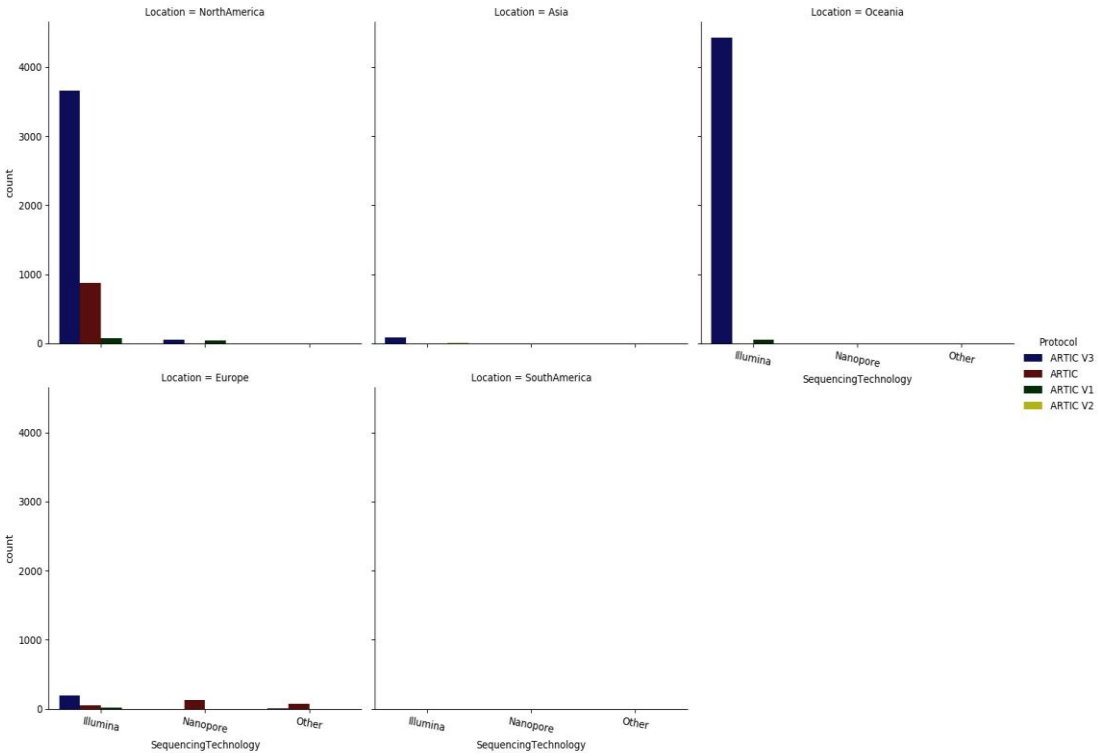


Figure 3-32. Usage of ARTIC protocols in different continents. This figure displays the ARTIC protocol exclusively because of its significance in being widely used worldwide and being the primary protocol utilized in the GISAID database.

3.2.7 COVID-19 Signal pipeline

The 20 sequences from the GISAID database with the highest rate of IUPAC codes from section 3.2.6 were chosen for employing the signal pipeline, and for positive control, 20 sequences without IUPAC code from the MSA file were randomly selected. Since the GISAID Id is unique for its website, a query was generated and specified for each id based on the metadata. As a result of the query, an SRA sequence id from the NCBI SRA database was found. Tables 3-14 and 3-15 show the GISAID Id, query, and SRA Id.

Table 3-14 Twenty sequences with highest number of IUPAC codes in the MSA file.

GISAID Id	Search Query	NCBI SRA database Id
EPI_ISL_732791	Spain coronavirus Illumina_MiSeq Logrono Male San Pedro 2020-11-01-	ERR5730727
EPI_ISL_466939	No items found.	-
EPI_ISL_591886	coronavirus Australia Victoria 2020-08-11 male Illumina NextSeq 500 VIDRL ARTIC v3 VIC11897/2020	SRR12894495
EPI_ISL_591945	coronavirus Australia Victoria 2020-08-10 Female Illumina NextSeq 500 VIDRL ARTIC v3 age 56	SRR12896086
EPI_ISL_592187	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 2020-08-06 betacoronavirus VIC13097	SRR12895904
EPI_ISL_591958	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 2020-08-11 betacoronavirus VIC12323	SRR12894517
EPI_ISL_562115	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 2020-07-23 betacoronavirus VIC8667	SRR12751148
EPI_ISL_561450	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus 2020-08-14 VIC10103	SRR12751432
EPI_ISL_565136	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 betacoronavirus 2020-07-23 VIC8668	SRR12751094
EPI_ISL_591949	coronavirus Illumina NextSeq 500 Australia Female ARTIC v3 betacoronavirus 2020-08-09 VIC12283	SRR12895222
EPI_ISL_591903	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 betacoronavirus 2020-08-11 VIC11955	SRR12894474
EPI_ISL_591877	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 betacoronavirus 2020-08-11 VIC11868	SRR12894441
EPI_ISL_592141	coronavirus Illumina NextSeq 500 Australia Male ARTIC v3 betacoronavirus 2020-08-04 VIC12994	SRR13130087
EPI_ISL_591944	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus VIC12275 2020-08-08	SRR12895975
EPI_ISL_592173	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus VIC13065 2020-08-05	SRR12894265
EPI_ISL_592158	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus VIC13034 2020-08-10	SRR12894865
EPI_ISL_591960	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus 2020-08-05 VIC12328	SRR12895784

EPI_ISL_591609	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus VIC13121 2020-08-09	SRR12896088
EPI_ISL_592513	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus VIC14480 2020-08-18	SRR12895420
EPI_ISL_592140	coronavirus Illumina NextSeq 500 Australia ARTIC v3 betacoronavirus VIC12993 2020-08-10	SRR12894926

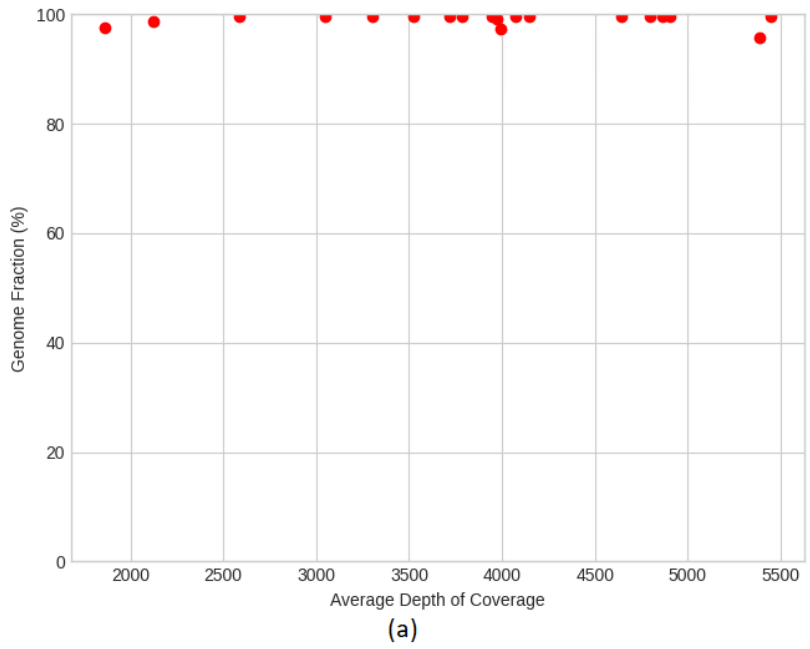
Also, the table shows the 20 random sequences without IUPAC codes extracted from the GISAID database.

*Table 3-15 Twenty sequences **without** IUPAC codes in the MSA file*

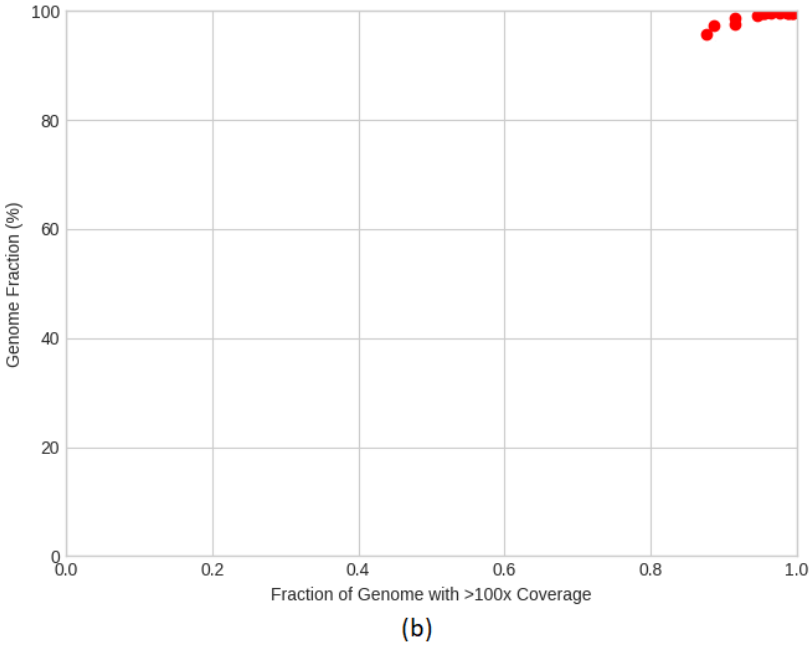
GISAID Id	Search Query	NCBI SRA database Id
EPI_ISL_565039	betacoronavirus 2020-07-27 Illumina VIC8472	SRR12751887
EPI_ISL_521425	betacoronavirus 2020-07-20 Illumina VIC6669	SRR12531797
EPI_ISL_518096	betacoronavirus 2020-07-26 Illumina VIC6058	SRR12529687
EPI_ISL_430567	betacoronavirus Illumina VIC1119 2020-04-08	SRR11622234
EPI_ISL_519836	betacoronavirus Illumina VIC5032 2020-07-17	SRR12536895
EPI_ISL_518148	betacoronavirus Illumina VIC6137 2020-07-23	SRR12531764
EPI_ISL_519894	betacoronavirus Illumina VIC5091 2020-07-19	SRR12537188
EPI_ISL_427053	betacoronavirus Illumina VIC778 2020-03-31	SRR11578274
EPI_ISL_521191	betacoronavirus Illumina VIC6399 2020-07-24	SRR12530055
EPI_ISL_563595	betacoronavirus Illumina VIC12796 2020-09-13	SRR12754829
EPI_ISL_626039	betacoronavirus Illumina VIC3559 2020-07-08	SRR12530735
EPI_ISL_427153	betacoronavirus Illumina VIC920 2020-03-27	SRR11578384
EPI_ISL_521558	betacoronavirus Illumina VIC2721 2020-07-04	SRR12530860
EPI_ISL_663930	betacoronavirus Illumina VIC17576 2020-07-28	SRR13179033
EPI_ISL_565807	betacoronavirus Illumina VIC9962 2020-07-23	SRR12753825

EPI_ISL_518696	betacoronavirus Illumina VIC5917 2020-07-21	SRR12529896
EPI_ISL_593086	betacoronavirus Illumina VIC14972 2020-08-19	SRR12894829
EPI_ISL_521856	betacoronavirus Illumina VIC3027 2020-07-01	SRR12530887
EPI_ISL_519375	betacoronavirus Illumina VIC3985 2020-07-15	SRR12528826
EPI_ISL_519948	betacoronavirus Illumina VIC5150 2020-07-12	SRR12531204

Then data from the SRA database were downloaded using a script and passed to the COVID-19-Signal pipeline [134]. Figure3-33 and 3-34 are the results of the COVID-19-Signal pipeline. As it is apparent from the figures, both sequences with IUPAC codes and without IUPAC codes have the same pattern; a relationship between the existence of the IUPAC code and average depth of coverage and genome fraction, which would help assess the reliability of data was not observed. Therefore, the existence of IUPAC code in a sequence is not a factor of bias.



(a)



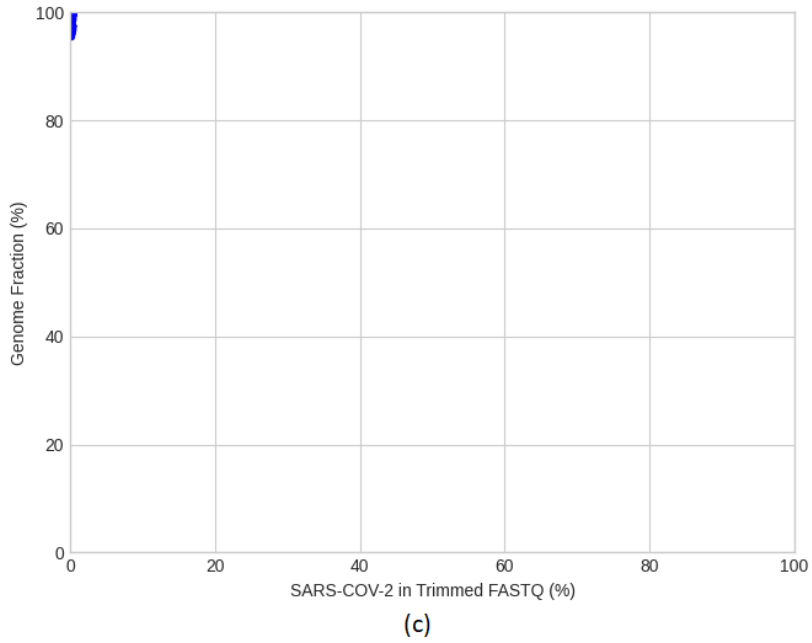
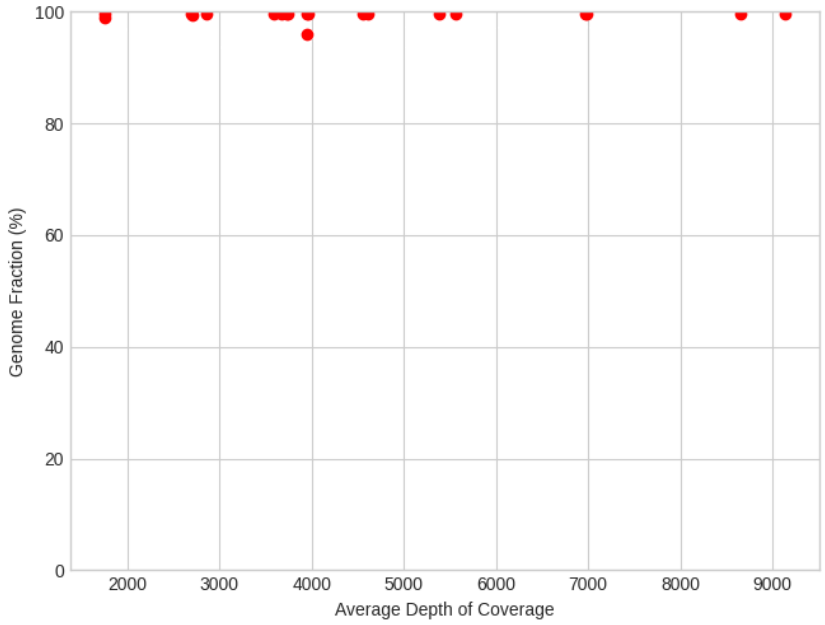
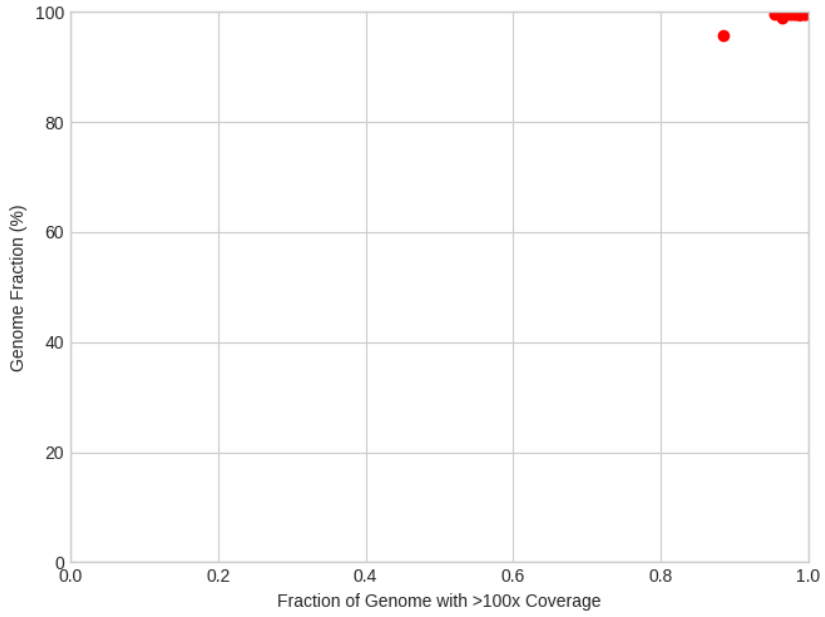


Figure 3-33. COVID-19-signal results for twenty sequences with highest amount of IUPAC codes in the MSA file (a), (b), (c). Each point represents one sample.


100
80
60
40
20
0
Genome Fraction (%)
2000
3000
4000
5000
6000
7000
8000
9000
Average Depth of Coverage

(a)


100
80
60
40
20
0
Genome Fraction (%)
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Genome with >100x Coverage

(b)

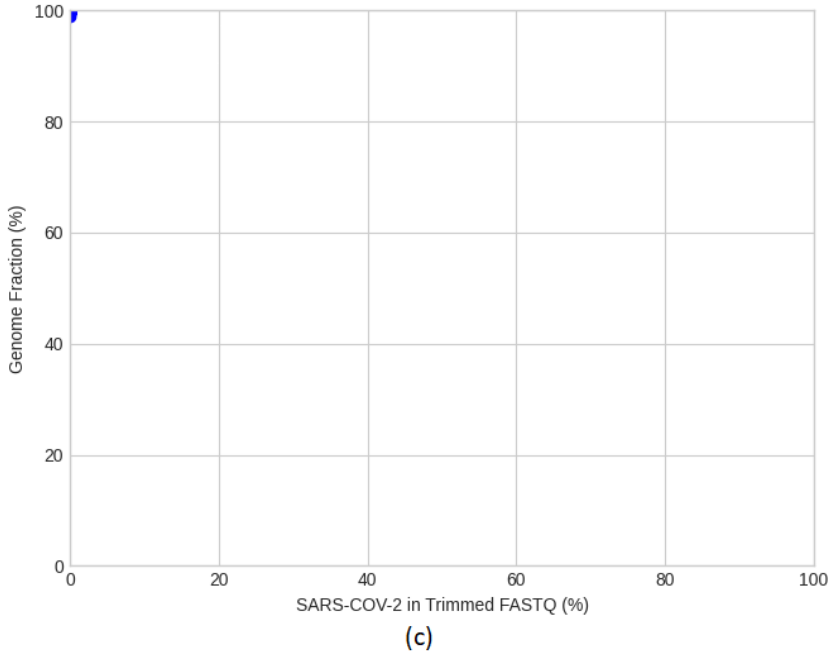


Figure 3-34. COVID-19-signal results for twenty sequences without IUPAC codes in the MSA file (a), (b), (c). Each point represents one sample.

3.3 Exploring the effect of mutations on depth of coverage by sequencing technology

Data for this section are extracted randomly. From the start of the pandemic until September 2022, 50 sequences per month with the available depth of coverage were selected from the GISAID website. The result of mutation's effect on the depth of coverage of Nanopore and Illumina are shown in Figures 3-35 and 3-36. It was expected to gradually decrease in depth of coverage from the start of the pandemic because the primers used were designed for the Wuhan strain. Therefore, it was predicted that the depth of coverage would have gaps or decreases during the time due to the introduction of numerous mutations in the virus' genome. However, the figures only show some change

of pattern and decrease in certain months, which was different from what was expected. Accordingly, no relationship was found between mutation's effects on the depth of coverage in different sequencing technology in this period.

The depth of coverage was expected to decrease over time because of many available variants. The depth of coverage is a mean, and when a small portion of Sars-Cov-2 is affected by mutation, it may not affect the mean, and some amplicons in the prepared kits may rescue the parts in the Sars-Cov-2 that were mutated in the process of amplification. Unfortunately, we cannot test amplicon by amplicon because this information is unavailable.

After examining the metadata of the data, it is regrettable that I could not provide an explanation for the increase in depth of coverage observed in November 2021 for Nanopore and in January 2021 for Illumina.

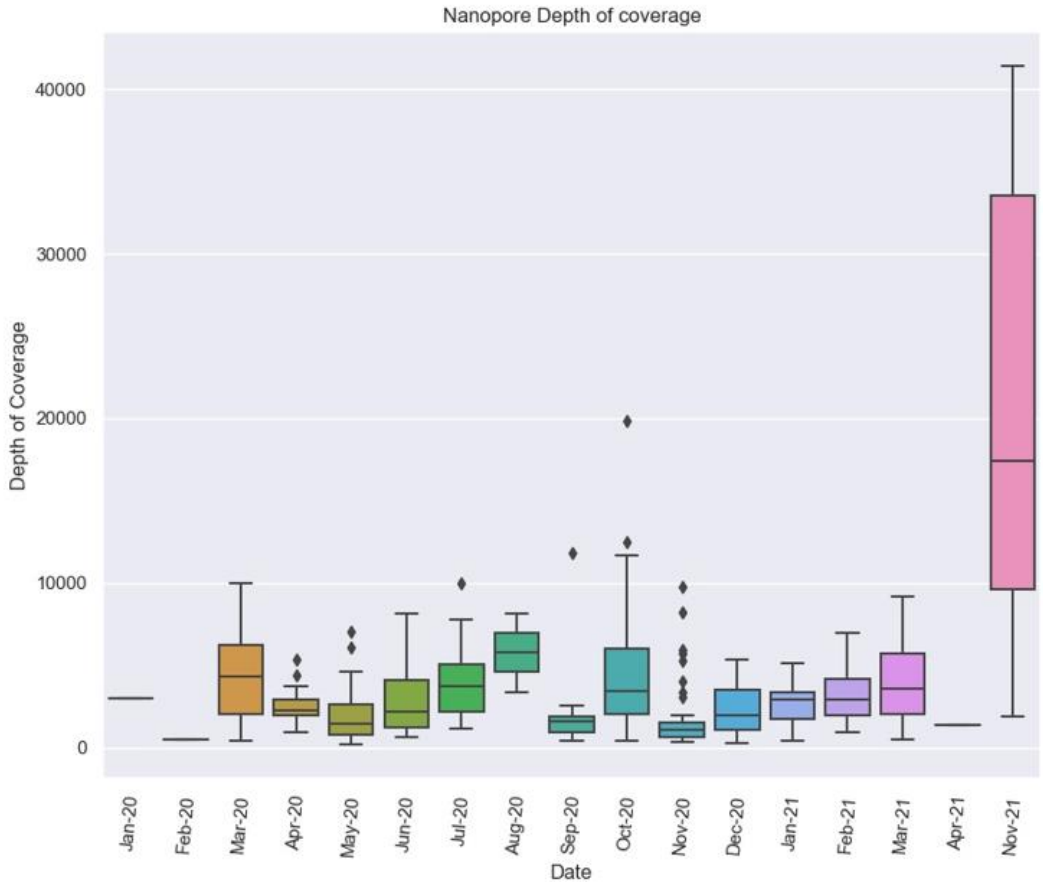


Figure 3-35 Depth of coverage of 50 random sequences for every month sequenced by Nanopore from the start of the pandemic until September 2022. Months without any data did not appear in the figure.

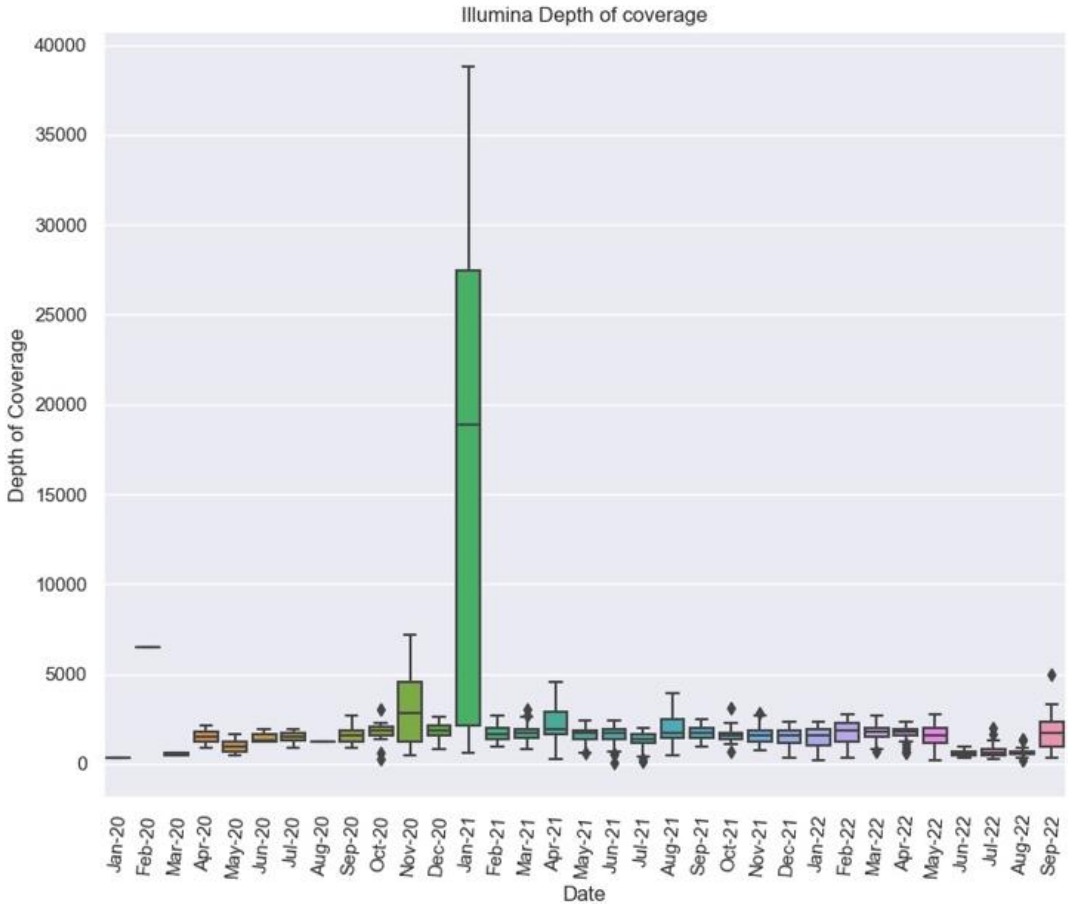


Figure 3-36. Depth of coverage of 50 random sequences for every month sequenced by Illumina from the start of the pandemic until September 2022.

In the next step finding the relationship at the time of appearing, variants were studied to see if there was a relationship. The Figure 3-37 and figure 3-38 show the time of starting each variant for Nanopore and Illumina, but the result did not show any relationship.

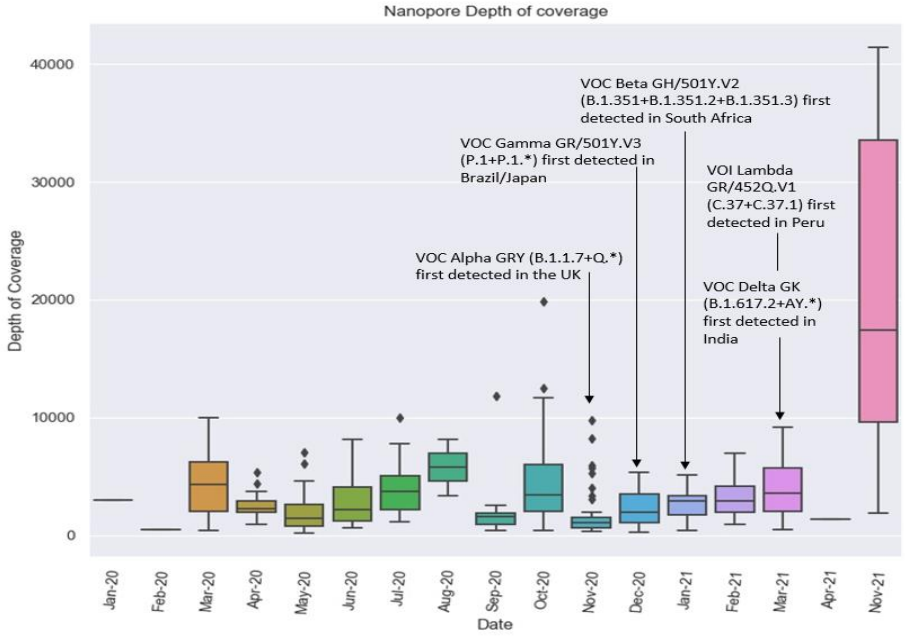


Figure 3-37. Starting point of variants in the depth of coverage of 50 random sequences for every month sequenced by Nanopore from the start of the pandemic until September 2022.

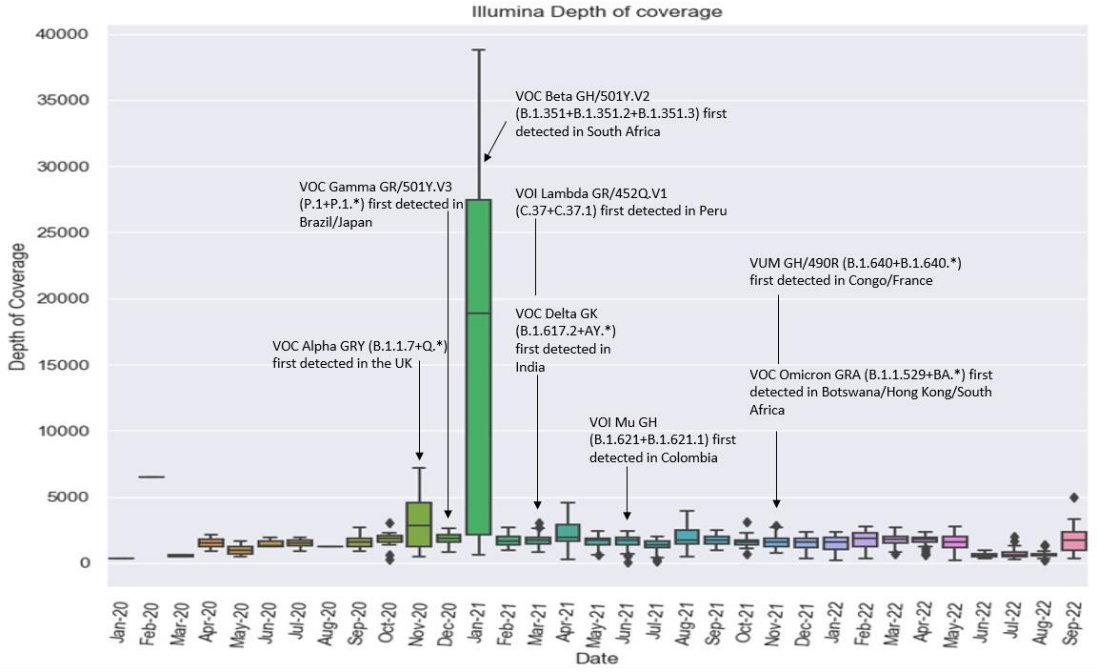


Figure 3-38. Starting point of variants in the depth of coverage of 50 random sequences for every month sequenced by Illumina from the start of the pandemic until September 2022.

4 Conclusions

Metagenomics approaches are culture-independent methods used to identify and classify microbial communities [151]. The approach to studying microbial communities depends on that community's abundance and the project [152].

Some microbes are low abundant and cannot be detected. Figure 4-1 shows the relationship between the level of abundance and detection in different microbiomes, every microbe above the threshold is easy to identify with sequencing methods. Below the threshold are two groups: the gray area, an uncertain area, and the ones below the gray zone, which is low abundant and cannot be detected by targeted sequencing, SARS-CoV-2 belongs to this group. Metagenomics' approaches can be divided into two main categories targeted sequencing (16SrRNA) and whole genome shotgun sequencing [153].

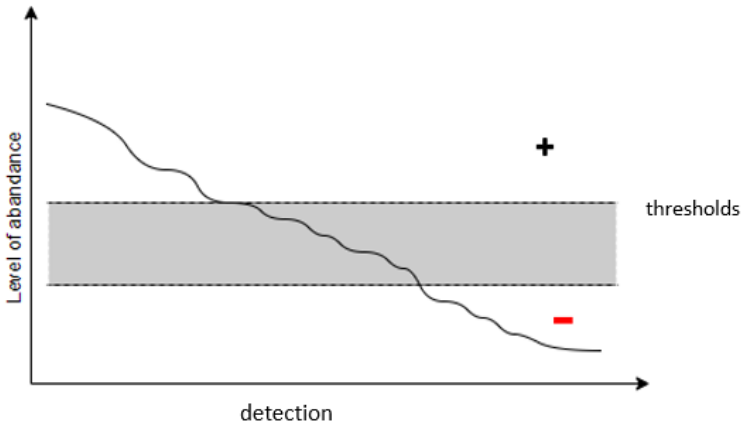


Figure 4-1. Relationship between the level of abundance and detection in different microbiomes.

This project aimed to design a pipeline that helps to analyze microbiomes with low abundance, and SARS-CoV-2 was chosen because of its importance and availability as an example to test the accuracy and results of different pipeline parts in two sequencing technologies, Nanopore and Illumina.

The pipeline is designed and developed to better understand microbiomes with low abundance with visualization tools and to research sequencing technologies' effect in providing bias and investigating the impact of mutations on the depth of coverage by sequencing technology. Data supplied from the pipeline allowed the following conclusion regarding SARS-CoV-2 data.

Data for SARS-CoV-2 from the GISAID website were employed to test the different sections of the project. Some data showed a relationship between sequencing technology and bias in the data, while time and location (BC vs Quebec) did not have an effect as a confounding factor. There was no apparent relationship between IUPAC codes and the usage of ARTIC protocol in different continents. There was not any evidence proving there is an effect from a mutation on the depth of coverage by sequencing technology.

In summary, this study helps enhance the accuracy of diagnostics, track viral variants, optimize vaccine design, and support informed public health decision-making. This study contributes to our understanding and management of COVID-19 by addressing biases in sequencing technology. The emphasis was on accurately identifying COVID-19 cases and facilitating prompt and effective public health responses. Additionally, the research has a role in monitoring the virus's genetic evolution and providing valuable insights for the interventions.

5 Perspectives

Studying some microbes because of their relevance to human disease is essential. However, their low abundance and hard-to-reach detection by meta-genomics approaches are the reason for using targeted sequencing to amplify the microbes. This project helps identify and characterize potential bias related to sequencing technologies using a publicly available dataset of SARS-CoV-2 swab specimens.

In future studies, there is potential for improving the bioinformatics pipeline used for analyzing specimen metagenomics. The computational resources available were limited, so only a small dataset was used for certain parts of the project. Thousands of sequencing datasets are available that could be used for further research. The raw data was not accessible for this project, and I had to do metadata extraction for some of the analysis. FastQC was used for comprehensive analysis, but there are newer technologies in other areas, such as graph genome, that could be explored by others. Further work on improving the graph genome and creating a comprehensive database of human pathogenic and non-pathogenic microbes at the strain level could lead to identifying new virulent variants and aiding vaccine development.

5.1 Graph genome [154]

A *graph genome* is a graph that represents genetic variation [155]. The graph shows alignment mismatches from the reference genome and includes the reference genome and all variants from multiple sequence alignment [154]. Moreover, it helps simplify

understanding and discovering patterns in the genome and helps analyze genome evolution [154, 156].

An interval was defined for this graph, and a similar section is shown with one line in the graph, and the sections where all nucleotides are not the same are shown with different lines. The process of searching for sequences and comparing them involves the utilization of a suffix array [157] for each mismatch. A *suffix array* is a sorted data structure of a string's suffixes [157]. An edge is separated from the main graph based on the reference genome for each mismatch. To reduce memory usage, every time a subsequence is available in previous sequences, the subsequence is replaced with a specific character in the new sequence, so it is not considered twice in the calculations. In addition, to increase the speed and memory usage, instead of using nucleotides after comparing them, each nucleotide's location on the y-axis is calculated; afterwards, each sequence is plotted on the graph. The pseudocode 5-1 shows the algorithm's steps.

```
ProcessSequences method :
Generate y-axis
Generate empty x List with reference genome length
Find sequencing technology for the sequence

Perform for loop in all sequences:
        Perform for loop in all altered sequences:
                Perform for loop in all nucleotides:
                        Check if the nucleotide in altered sequences list are equal to
                        the nucleotide in the altered sequences list in i location,
                           If they are equal and
                                If it's not the last nucleotide of the sequence add the
                        nucleotide to
                                        The segment list , which contains the similar
                        nucleotides
```

```
                If it is the last nucleotide add that to the string of y-
axis(newLine)
   else
                If they're not equal add the segment to the newLine
and clear the segment and
    increase the repeatList for the nucleotides in the segment

add the sequence to altered sequences list
 draw y axis to graph genome plot
```

Algorithm 5-1. Pseudocode that shows the high-level logic of the graph genome algorithm and explains different algorithm steps.

A file containing multiple sequence alignments should be given to the pipeline as input; the file that was used for this purpose is downloaded from the GISAID website to generate a graph genome. Separate lines were drawn if they were different in a specific interval. Otherwise, if they were similar in that interval, they were drawn by one line. In other words, the contigs in specific intervals in sequences were drawn, and other regions of sequences were ignored. A sample for the graph genome with an example sequence part is explained in Figure 6-1.

There were around 68554 sequences available at the time -April 13th, 2021- to increase the readability of the graph; because of the size of the data, it was impossible to draw the graph with available servers; therefore, spike protein data from Newfoundland and Labrador was used. At first, the spike protein sequence from all Newfoundland and Labrador sequences was extracted from the multiple sequence alignments file; then, the pipeline produced the graph genome for spike sequences.

In summary, a tool for comparing all sequences to each other and reference genome was developed to ensure the DNA of spike proteins extracted from the multiple sequence

alignment file. For COVID-19, spikes for each sequence were extracted and given to the pipeline as input. Also, the SARS-CoV-2 reference genome spike was extracted from the UCSC genome browser [136].

The result shows that all the sequences in 38 sites differ from extracted DNA from UCSC. Since this region's DNA is 3822 base pairs long, 38 sites are reasonable, and the MSA spike's DNA is reliable.

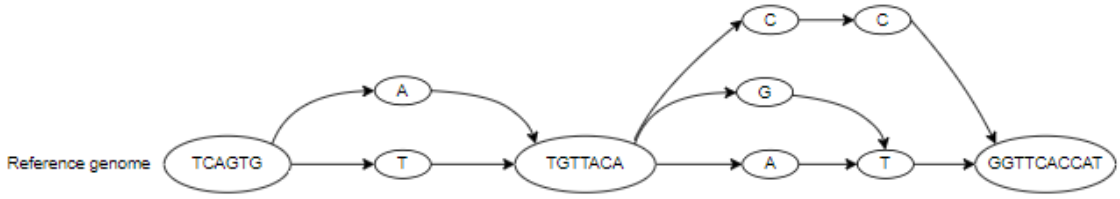


Figure 5-1. Graph genome. An example of a graph genome with a built sequence is the nodes are nucleotides, and edges are between nucleotides in the sequence. The figure's first six nucleotides (TCAGTG) are the same between sequences. Therefore, there is one circle for all six nucleotides. For the next nucleotide, some sequences have 'A,' and some sequences have 'T' since the reference genome is 'T,' the straight line is 'T,' and 'A' is shown as a branch of it. The following seven nucleotides are the same among sequences. Therefore, there is one sequence for all of them. The following two nucleotides in the reference genome are 'A' and 'T,' but in some sequences, it is 'G,' 'T,' and 'C,' and 'C'; therefore, some branches were made for those sequences in the graph genome. In this graph genome, the threshold was considered 5. If the number of similar sequences exceeds the threshold, they will break into different branches.

After uploading all data and projects on the Graham server on Compute Canada [158], the result comes as the following graph. The data for this part of the project was for multiple sequence alignment of Canadian data gathered from the GISAID database in January 2022. The yellow line in the figure 6-2 represents the reference genome, the blue lines are sequences that utilized Nanopore sequencing technology, and the green lines used Illumina sequencing technology.

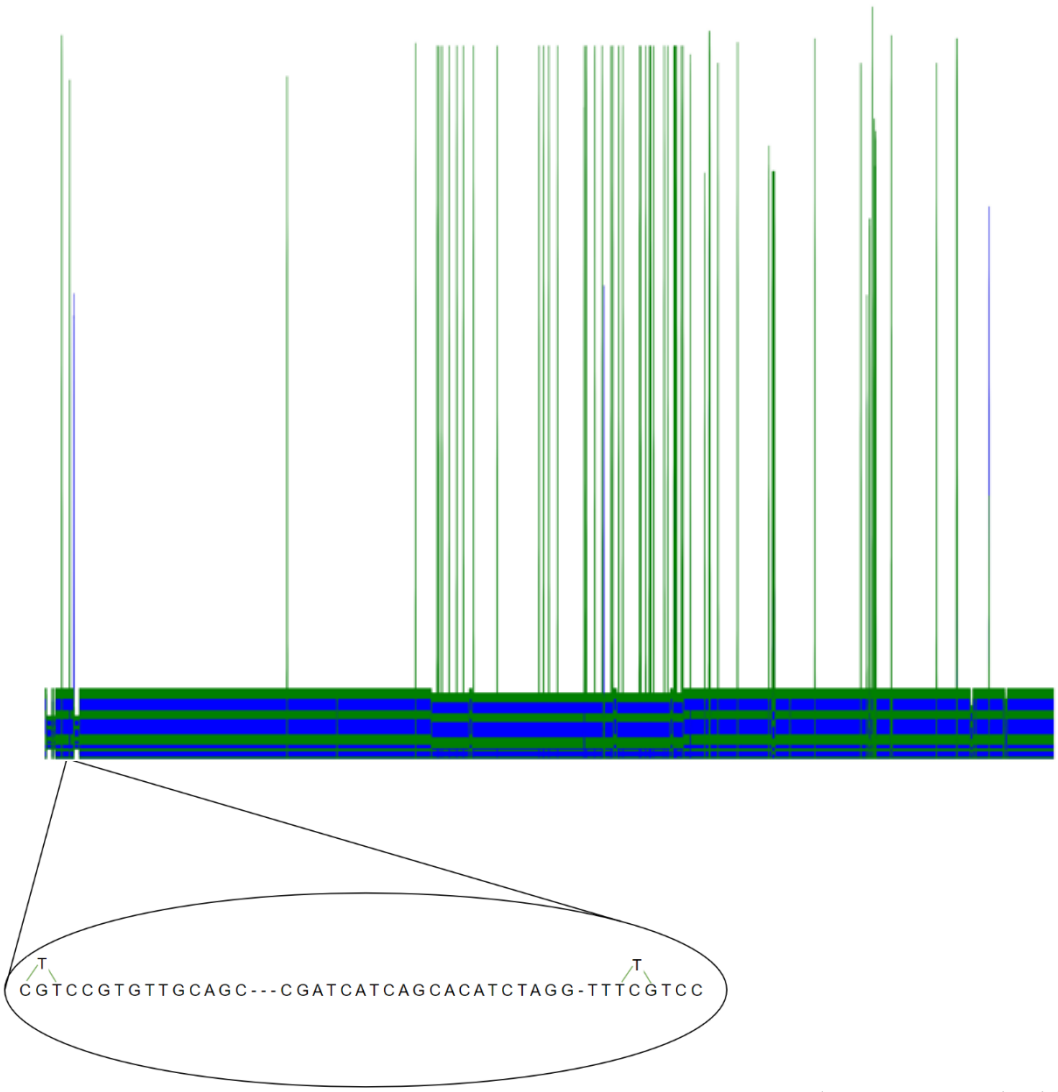


Figure 5-2. Graph genome for Canadian data for Illumina and Nanopore. The blue lines are sequences that are sequenced by Nanopore sequencing technology, and the green lines are for Illumina. In the part of the graph that is magnified, part of the sequences is cut, nucleotide 415 to 447 in 10 first sequences in the MSA file, the nucleotide C matches between selected sequences, and after that, nucleotide T is available in some sequences, so in the graph, there is a branch to nucleotide T, and because Illumina sequences this nucleotide in these sequences, the line is green. Then "TCCGTGTTGCAGC---CGATCATCAGCACATCTAGG-TTT" are the same in the selected sequences. Then there is a nucleotide C in the reference genome and in some of the selected sequences, but in some sequences, there is nucleotide T in this position; therefore, there is an edge from the previous nucleotide, nucleotide T, to this nucleotide and since Illumina sequences this sequences the edge is green. And then, "GTCC" is similar in the selected sequences.

During the last phase of my thesis, I explored a few possible ways to improve my work, and graph genome was one that still needs improvement.

Bibliography

[1] S. Khare *et al.*, “GISAID’s Role in Pandemic Response,” *China CDC Wkly*, vol. 3, no. 49, pp. 1049–1051, 2021, doi: 10.46234/ccdcw2021.255.

[2] R. Sender, S. Fuchs, and R. Milo, “Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans.,” *Cell*, vol. 164, no. 3, pp. 337–40, Jan. 2016, doi: 10.1016/j.cell.2016.01.013.

[3] Integrative HMP (iHMP) Research Network Consortium, “The Integrative Human Microbiome Project.,” *Nature*, vol. 569, no. 7758, pp. 641–648, 2019, doi: 10.1038/s41586-019-1238-8.

[4] D. Zheng, T. Liwinski, and E. Elinav, “Interaction between microbiota and immunity in health and disease.,” *Cell Res*, vol. 30, no. 6, pp. 492–506, 2020, doi: 10.1038/s41422-020-0332-7.

[5] Human Microbiome Project Consortium, “Structure, function and diversity of the healthy human microbiome.,” *Nature*, vol. 486, no. 7402, pp. 207–14, Jun. 2012, doi: 10.1038/nature11234.

[6] M. J. Bull and N. T. Plummer, “Part 1: The Human Gut Microbiome in Health and Disease.,” *Integr Med (Encinitas)*, vol. 13, no. 6, pp. 17–22, Dec. 2014.

[7] K. A. Lee, M. K. Luong, H. Shaw, P. Nathan, V. Bataille, and T. D. Spector, “The gut microbiome: what the oncologist ought to know.,” *Br J Cancer*, vol. 125, no. 9, pp. 1197–1209, 2021, doi: 10.1038/s41416-021-01467-x.

[8] M. F. Fernández, I. Reina-Pérez, J. M. Astorga, A. Rodríguez-Carrillo, J. Plaza-Díaz, and L. Fontana, "Breast Cancer and Its Relationship with the Microbiota.," *Int J Environ Res Public Health*, vol. 15, no. 8, 2018, doi: 10.3390/ijerph15081747.

[9] M. J. Bull and N. T. Plummer, "Part 1: The Human Gut Microbiome in Health and Disease.," *Integr Med (Encinitas)*, vol. 13, no. 6, pp. 17–22, Dec. 2014.

[10] D. Rothschild *et al.*, "Environment dominates over host genetics in shaping human gut microbiota.," *Nature*, vol. 555, no. 7695, pp. 210–215, 2018, doi: 10.1038/nature25973.

[11] D. Rothschild *et al.*, "Environment dominates over host genetics in shaping human gut microbiota.," *Nature*, vol. 555, no. 7695, pp. 210–215, 2018, doi: 10.1038/nature25973.

[12] Y. Belkaid and T. W. Hand, "Role of the microbiota in immunity and inflammation.," *Cell*, vol. 157, no. 1, pp. 121–41, Mar. 2014, doi: 10.1016/j.cell.2014.03.011.

[13] H. Wang *et al.*, "Potential Associations Between Microbiome and COVID-19," *Front Med (Lausanne)*, vol. 8, Dec. 2021, doi: 10.3389/fmed.2021.785496.

[14] A. Sharma, S. Tiwari, M. K. Deb, and J. L. Marty, "Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies.," *Int J Antimicrob Agents*, vol. 56, no. 2, p. 106054, Aug. 2020, doi: 10.1016/j.ijantimicag.2020.106054.

[15] A. R. Sahin, "2019 Novel Coronavirus (COVID-19) Outbreak: A Review of the Current Literature," *Eurasian J Med Oncol*, 2020, doi: 10.14744/ejmo.2020.12220.

[16] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, "Author Correction: Characteristics of SARS-CoV-2 and COVID-19," *Nat Rev Microbiol*, vol. 20, no. 5, pp. 315–315, May 2022, doi: 10.1038/s41579-022-00711-2.

[17] M. Kumar and S. Al Khodor, “Pathophysiology and treatment strategies for COVID-19.,” *J Transl Med*, vol. 18, no. 1, p. 353, 2020, doi: 10.1186/s12967-020-02520-8.

[18] M. Giovanetti *et al.*, “Evolution patterns of SARS-CoV-2: Snapshot on its genome variants,” *Biochem Biophys Res Commun*, vol. 538, pp. 88–91, Jan. 2021, doi: 10.1016/j.bbrc.2020.10.102.

[19] M. Bartas *et al.*, “In-Depth Bioinformatic Analyses of Nidovirales Including Human SARS-CoV-2, SARS-CoV, MERS-CoV Viruses Suggest Important Roles of Non-canonical Nucleic Acid Structures in Their Lifecycles,” *Front Microbiol*, vol. 11, Jul. 2020, doi: 10.3389/fmicb.2020.01583.

[20] Y.-C. Wu, C.-S. Chen, and Y.-J. Chan, “The outbreak of COVID-19: An overview,” *Journal of the Chinese Medical Association*, vol. 83, no. 3, pp. 217–220, Mar. 2020, doi: 10.1097/JCMA.0000000000000270.

[21] M. By Patrick R. Murray, PhD, Ken S. Rosenthal, PhD and Michael A. Pfaller, *MEDICAL MICROBIOLOGY*. Elsevier, 2020.

[22] C. I. Paules, H. D. Marston, and A. S. Fauci, “Coronavirus Infections—More Than Just the Common Cold,” *JAMA*, vol. 323, no. 8, p. 707, Feb. 2020, doi: 10.1001/jama.2020.0757.

[23] A. M. Al-Qaaneh, T. Alshammari, R. Aldahhan, H. Aldossary, Z. A. Alkhalifah, and J. F. Borgio, “Genome composition and genetic characterization of SARS-CoV-2,” *Saudi J Biol Sci*, vol. 28, no. 3, pp. 1978–1989, Mar. 2021, doi: 10.1016/j.sjbs.2020.12.053.

[24] C. Ogimi, Y. J. Kim, E. T. Martin, H. J. Huh, C.-H. Chiu, and J. A. Englund, “What’s New With the Old Coronaviruses?,” *J Pediatric Infect Dis Soc*, vol. 9, no. 2, pp. 210–217, Apr. 2020, doi: 10.1093/jpids/piaa037.

[25] F. Lin, X. Wang, and M. Zhou, “How trade affects pandemics? Evidence from severe acute respiratory syndromes in 2003,” *World Econ*, vol. 45, no. 7, pp. 2270–2283, Jul. 2022, doi: 10.1111/twec.13127.

[26] World Health Organization, “WHO Coronavirus (COVID-19) Dashboard,” 2022. [Online]. Available: https://covid19.who.int/

[27] Government of Canada, “COVID-19: Outbreak update,” 2022. [Online]. Available: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html

[28] Centers for Disease Control and Prevention, “SARS Basics Fact Sheet,” 2017. doi: https://www.cdc.gov/sars/about/fs-sars.html.

[29] World Health Organization, “Middle East respiratory syndrome,” 2022. [Online]. Available: https://www.emro.who.int/health-topics/mers-cov/mers-outbreaks.html

[30] Centers for Disease Control and Prevention, “Middle East Respiratory Syndrome (MERS),Symptoms & Complications,” 2019. [Online]. Available: https://www.cdc.gov/coronavirus/mers/about/symptoms.html

[31] Centers for Disease Control and Prevention, “Middle East Respiratory Syndrome (MERS),Transmission,” 2019. [Online]. Available: https://www.cdc.gov/coronavirus/mers/about/transmission.html

[32] G. Chowell *et al.*, “Transmission characteristics of MERS and SARS in the healthcare setting: a comparative study,” *BMC Med*, vol. 13, no. 1, p. 210, Dec. 2015, doi: 10.1186/s12916-015-0450-0.

[33] Cable News Network, “Covid-19 Pandemic Timeline Fast Facts,” 2022. doi: https://www.cnn.com/2021/08/09/health/covid-19-pandemic-timeline-fast-facts/index.html.

[34] P. Pagliano, C. Sellitto, V. Conti, T. Ascione, and S. Esposito, “Characteristics of viral pneumonia in the COVID-19 era: an update,” *Infection*, vol. 49, no. 4, pp. 607–616, Aug. 2021, doi: 10.1007/s15010-021-01603-y.

[35] N. Zhu *et al.*, “A Novel Coronavirus from Patients with Pneumonia in China, 2019,” *New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/NEJMoa2001017.

[36] Y.-C. Wu, C.-S. Chen, and Y.-J. Chan, “The outbreak of COVID-19: An overview,” *Journal of the Chinese Medical Association*, vol. 83, no. 3, pp. 217–220, Mar. 2020, doi: 10.1097/JCMA.0000000000000270.

[37] S. Jiang, L. Du, and Z. Shi, “An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies,” *Emerg Microbes Infect*, vol. 9, no. 1, pp. 275–277, Jan. 2020, doi: 10.1080/22221751.2020.1723441.

[38] Y.-C. Wu, C.-S. Chen, and Y.-J. Chan, “The outbreak of COVID-19: An overview,” *Journal of the Chinese Medical Association*, vol. 83, no. 3, pp. 217–220, Mar. 2020, doi: 10.1097/JCMA.0000000000000270.

[39] B. Hu, H. Guo, P. Zhou, and Z.-L. Shi, “Characteristics of SARS-CoV-2 and COVID-19,” *Nat Rev Microbiol*, vol. 19, no. 3, pp. 141–154, Mar. 2021, doi: 10.1038/s41579-020-00459-7.

[40] World Health Organization (WHO), “Origins of the SARS-CoV-2 virus,” 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/origins-of-the-virus

[41] W. Li *et al.*, “Bats Are Natural Reservoirs of SARS-Like Coronaviruses,” *Science (1979)*, vol. 310, no. 5748, pp. 676–679, Oct. 2005, doi: 10.1126/science.1118391.

[42] World Health Organization (WHO), “Coronavirus disease (COVID-19): How is it transmitted?,” 2021. [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted

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

[44] A. Bal, R. Agrawal, P. Vaideeswar, S. Arava, and A. Jain, “COVID-19: An up-to-date review – from morphology to pathogenesis,” *Indian J Pathol Microbiol*, vol. 63, no. 3, p. 358, 2020, doi: 10.4103/IJPM.IJPM_779_20.

[45] L. Mousavizadeh and S. Ghasemi, “Genotype and phenotype of COVID-19: Their roles in pathogenesis,” *Journal of Microbiology, Immunology and Infection*, vol. 54, no. 2, pp. 159–163, Apr. 2021, doi: 10.1016/j.jmii.2020.03.022.

[46] L. Mousavizadeh and S. Ghasemi, “Genotype and phenotype of COVID-19: Their roles in pathogenesis,” *Journal of Microbiology, Immunology and Infection*, vol. 54, no. 2, pp. 159–163, Apr. 2021, doi: 10.1016/j.jmii.2020.03.022.

[47] M. Giovanetti *et al.*, "Evolution patterns of SARS-CoV-2: Snapshot on its genome variants," *Biochem Biophys Res Commun*, vol. 538, pp. 88–91, Jan. 2021, doi: 10.1016/j.bbrc.2020.10.102.

[48] P. K. Singh, U. Kulsum, S. B. Rufai, S. R. Mudliar, and S. Singh, "Mutations in SARS-CoV-2 Leading to Antigenic Variations in Spike Protein: A Challenge in Vaccine Development," *J Lab Physicians*, vol. 12, no. 02, pp. 154–160, Aug. 2020, doi: 10.1055/s-0040-1715790.

[49] A. A. T. Naqvi *et al.*, "Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach.," *Biochim Biophys Acta Mol Basis Dis*, vol. 1866, no. 10, p. 165878, 2020, doi: 10.1016/j.bbadis.2020.165878.

[50] R. A. Khailany, M. Safdar, and M. Ozaslan, "Genomic characterization of a novel SARS-CoV-2," *Gene Rep*, vol. 19, p. 100682, Jun. 2020, doi: 10.1016/j.genrep.2020.100682.

[51] D. Yesudhas, A. Srivastava, and M. M. Gromiha, "COVID-19 outbreak: history, mechanism, transmission, structural studies and therapeutics," *Infection*, vol. 49, no. 2, pp. 199–213, Apr. 2021, doi: 10.1007/s15010-020-01516-2.

[52] S. R. Weiss and S. Navas-Martin, "Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus," *Microbiology and Molecular Biology Reviews*, vol. 69, no. 4, pp. 635–664, Dec. 2005, doi: 10.1128/MMBR.69.4.635-664.2005.

[53] T. Behl *et al.*, "CD147-spike protein interaction in COVID-19: Get the ball rolling with a novel receptor and therapeutic target," *Science of The Total Environment*, vol. 808, p. 152072, Feb. 2022, doi: 10.1016/j.scitotenv.2021.152072.

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

[55] D. Yesudhas, A. Srivastava, and M. M. Gromiha, “COVID-19 outbreak: history, mechanism, transmission, structural studies and therapeutics,” *Infection*, vol. 49, no. 2, pp. 199–213, Apr. 2021, doi: 10.1007/s15010-020-01516-2.

[56] A. Bal, R. Agrawal, P. Vaideeswar, S. Arava, and A. Jain, “COVID-19: An up-to-date review – from morphology to pathogenesis,” *Indian J Pathol Microbiol*, vol. 63, no. 3, p. 358, 2020, doi: 10.4103/IJPM.IJPM_779_20.

[57] S. J. R. da Silva *et al.*, “Two Years into the COVID-19 Pandemic: Lessons Learned,” *ACS Infect Dis*, vol. 8, no. 9, pp. 1758–1814, Sep. 2022, doi: 10.1021/acsinfecdis.2c00204.

[58] T. Behl *et al.*, “CD147-spike protein interaction in COVID-19: Get the ball rolling with a novel receptor and therapeutic target,” *Science of The Total Environment*, vol. 808, p. 152072, Feb. 2022, doi: 10.1016/j.scitotenv.2021.152072.

[59] A. A. T. Naqvi *et al.*, “Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach.,” *Biochim Biophys Acta Mol Basis Dis*, vol. 1866, no. 10, p. 165878, 2020, doi: 10.1016/j.bbadis.2020.165878.

[60] R. A. Khailany, M. Safdar, and M. Ozaslan, “Genomic characterization of a novel SARS-CoV-2,” *Gene Rep*, vol. 19, p. 100682, Jun. 2020, doi: 10.1016/j.genrep.2020.100682.

[61] M. Giovanetti *et al.*, “Evolution patterns of SARS-CoV-2: Snapshot on its genome variants,” *Biochem Biophys Res Commun*, vol. 538, pp. 88–91, Jan. 2021, doi: 10.1016/j.bbrc.2020.10.102.

[62] K. E. Kistler, J. Huddleston, and T. Bedford, “Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2,” *Cell Host Microbe*, vol. 30, no. 4, pp. 545-555.e4, Apr. 2022, doi: 10.1016/j.chom.2022.03.018.

[63] Government of Canada, “COVID-19: Symptoms, treatment, what to do if you feel sick,” 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.html

[64] A. V. Raveendran, R. Jayadevan, and S. Sashidharan, “Long COVID: An overview,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 15, no. 3, pp. 869–875, May 2021, doi: 10.1016/j.dsx.2021.04.007.

[65] Centers for Disease Control and Prevention, “Symptoms of COVID-19,” 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[66] X. Chen *et al.*, “A systematic review of neurological symptoms and complications of COVID-19,” *J Neurol*, vol. 268, no. 2, pp. 392–402, Feb. 2021, doi: 10.1007/s00415-020-10067-3.

[67] H. C. Koc, J. Xiao, W. Liu, Y. Li, and G. Chen, “Long COVID and its Management,” *Int J Biol Sci*, vol. 18, no. 12, pp. 4768–4780, 2022, doi: 10.7150/ijbs.75056.

[68] S. Lopez-Leon *et al.*, “More than 50 long-term effects of COVID-19: a systematic review and meta-analysis,” *Sci Rep*, vol. 11, no. 1, p. 16144, Dec. 2021, doi: 10.1038/s41598-021-95565-8.

[69] Centers for Disease Control and Prevention, “Post-COVID Conditions: Information for Healthcare Providers,” 2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/post-covid-conditions.html

[70] G. Van Vo, E. Bagyinszky, and S. S. A. An, “COVID-19 Genetic Variants and Their Potential Impact in Vaccine Development,” *Microorganisms*, vol. 10, no. 3, p. 598, Mar. 2022, doi: 10.3390/microorganisms10030598.

[71] D. Tshiabuila *et al.*, “Comparison of SARS-CoV-2 sequencing using the ONT GridION and the Illumina MiSeq.,” *BMC Genomics*, vol. 23, no. 1, p. 319, Apr. 2022, doi: 10.1186/s12864-022-08541-5.

[72] P. Yang and X. Wang, “COVID-19: a new challenge for human beings,” *Cell Mol Immunol*, vol. 17, no. 5, pp. 555–557, May 2020, doi: 10.1038/s41423-020-0407-x.

[73] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, “Overview of Next-Generation Sequencing Technologies.,” *Curr Protoc Mol Biol*, vol. 122, no. 1, p. e59, 2018, doi: 10.1002/cpmb.59.

[74] H. P. J. Buermans and J. T. den Dunnen, “Next generation sequencing technology: Advances and applications,” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014, doi: 10.1016/j.bbadis.2014.06.015.

[75] I. Illumina, “Explore Illumina sequencing technology.” [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

[76] R. K. Ravi, K. Walton, and M. Khosroheidari, “MiSeq: A Next Generation Sequencing Platform for Genomic Analysis,” 2018, pp. 223–232. doi: 10.1007/978-1-4939-7471-9_12.

[77] W. Gu, S. Miller, and C. Y. Chiu, “Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection.,” *Annu Rev Pathol*, vol. 14, pp. 319–338, 2019, doi: 10.1146/annurev-pathmechdis-012418-012751.

[78] I. Illumina, “Understanding the NGS workflow.” [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html

[79] H. P. J. Buermans and J. T. den Dunnen, “Next generation sequencing technology: Advances and applications,” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, Oct. 2014, doi: 10.1016/j.bbadis.2014.06.015.

[80] W. Gu, S. Miller, and C. Y. Chiu, “Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection.,” *Annu Rev Pathol*, vol. 14, pp. 319–338, 2019, doi: 10.1146/annurev-pathmechdis-012418-012751.

[81] M. Kircher, P. Heyn, and J. Kelso, “Addressing challenges in the production and analysis of illumina sequencing data,” *BMC Genomics*, vol. 12, no. 1, p. 382, Dec. 2011, doi: 10.1186/1471-2164-12-382.

[82] W. Gu, S. Miller, and C. Y. Chiu, “Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection.,” *Annu Rev Pathol*, vol. 14, pp. 319–338, 2019, doi: 10.1146/annurev-pathmechdis-012418-012751.

[83] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, “Overview of Next-Generation Sequencing Technologies.,” *Curr Protoc Mol Biol*, vol. 122, no. 1, p. e59, 2018, doi: 10.1002/cpmb.59.

[84] W. Gu, S. Miller, and C. Y. Chiu, “Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection.,” *Annu Rev Pathol*, vol. 14, pp. 319–338, 2019, doi: 10.1146/annurev-pathmechdis-012418-012751.

[85] R. K. Ravi, K. Walton, and M. Khosroheidari, “MiSeq: A Next Generation Sequencing Platform for Genomic Analysis,” 2018, pp. 223–232. doi: 10.1007/978-1-4939-7471-9_12.

[86] A. Żmieńko and A. Satyr, “Nanopore Sequencing and its Application in Biology,” *Postepy Biochem*, vol. 66, no. 3, pp. 193–204, 2020, doi: 10.18388/pb.2020_328.

[87] Y. Zhou *et al.*, “Application of Nanopore Sequencing in the Detection of Foodborne Microorganisms,” *Nanomaterials*, vol. 12, no. 9, p. 1534, May 2022, doi: 10.3390/nano12091534.

[88] H. Lu, F. Giordano, and Z. Ning, “Oxford Nanopore MinION Sequencing and Genome Assembly.,” *Genomics Proteomics Bioinformatics*, vol. 14, no. 5, pp. 265–279, Oct. 2016, doi: 10.1016/j.gpb.2016.05.004.

[89] N. Kono and K. Arakawa, “Nanopore sequencing: Review of potential applications in functional genomics,” *Dev Growth Differ*, vol. 61, no. 5, pp. 316–326, Jun. 2019, doi: 10.1111/dgd.12608.

[90] Oxford Nanopore Technologies, “How does nanopore DNA sequencing work?” [Online]. Available: https://nanoporetech.com/applications/dna-nanopore-sequencing

[91] L. L. Zhang, C. Zhang, and J. P. Peng, “Application of Nanopore Sequencing Technology in the Clinical Diagnosis of Infectious Diseases.,” *Biomed Environ Sci*, vol. 35, no. 5, pp. 381–392, May 2022, doi: 10.3967/bes2022.054.

[92] H. Lu, F. Giordano, and Z. Ning, “Oxford Nanopore MinION Sequencing and Genome Assembly,” *Genomics Proteomics Bioinformatics*, vol. 14, no. 5, pp. 265–279, Oct. 2016, doi: 10.1016/j.gpb.2016.05.004.

[93] A. P. Heikema *et al.*, "Comparison of Illumina versus Nanopore 16S rRNA Gene Sequencing of the Human Nasal Microbiota," *Genes (Basel)*, vol. 11, no. 9, p. 1105, Sep. 2020, doi: 10.3390/genes11091105.

[94] L. M. Petersen, I. W. Martin, W. E. Moschetti, C. M. Kershaw, and G. J. Tsongalis, "Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing," *J Clin Microbiol*, vol. 58, no. 1, Dec. 2019, doi: 10.1128/JCM.01315-19.

[95] B. Egeter *et al.*, "Speeding up the detection of invasive bivalve species using environmental DNA: A Nanopore and Illumina sequencing comparison," *Mol Ecol Resour*, vol. 22, no. 6, pp. 2232–2247, Aug. 2022, doi: 10.1111/1755-0998.13610.

[96] A. Zhang, "Are there any new sequencing technologies coming out? How do they work?" [Online]. Available: https://www.thetech.org/ask-a-geneticist/3rd-generation-sequencing

[97] J. Lang, "NanoCoV19: An analytical pipeline for rapid detection of severe acute respiratory syndrome coronavirus 2," *Front Genet*, vol. 13, Sep. 2022, doi: 10.3389/fgene.2022.1008792.

[98] Artic Network, "SARS-CoV-2." [Online]. Available: https://artic.network/ncov-2019

[99] Artic Network, "hCoV-2019/nCoV-2019 Version 3 Amplicon Set," 2020.

[100] H. R. Benson, "An introduction to benchmarking in healthcare.," *Radiol Manage*, vol. 16, no. 4, pp. 35–9, 1994.

[101] T. Liu *et al.*, "A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples.," *iScience*, vol. 24, no. 8, p. 102892, Aug. 2021, doi: 10.1016/j.isci.2021.102892.

[102] Wikipedia Contributors, “Phylogenetic tree.” 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Phylogenetic_tree

[103] P. Kapli, Z. Yang, and M. J. Telford, “Phylogenetic tree building in the genomic age.,” *Nat Rev Genet*, vol. 21, no. 7, pp. 428–444, 2020, doi: 10.1038/s41576-020-0233-0.

[104] Wikipedia Contributors, “Newick format,” *Wikipedia*. Wikipedia, The Free Encyclopedia., 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Newick_format

[105] D. Slutsky, “The Effective Use of Graphs,” *J Wrist Surg*, vol. 03, no. 02, pp. 067–068, May 2014, doi: 10.1055/s-0034-1375704.

[106] Sauro Jeff and Lewis Jim, “cumulative-graphs.” [Online]. Available: https://measuringu.com/cumulative-graphs/

[107] S. Crase and S. N. Thennadil, “An analysis framework for clustering algorithm selection with applications to spectroscopy,” *PLoS One*, vol. 17, no. 3, p. e0266369, Mar. 2022, doi: 10.1371/journal.pone.0266369.

[108] D. Xu, Rui and Wunsch, “Cluster Analysis,” in *Clustering*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2009, pp. 1–13. doi: 10.1002/9780470382776.ch1.

[109] Cormen Thomas H., Leiserson Charles E., Rivest Ronald L., and Stein Clifford, *Introduction to Algorithms*, Second. MIT Press, 2001.

[110] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput.,” *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–7, 2004, doi: 10.1093/nar/gkh340.

[111] “Biopython.” [Online]. Available: https://biopython.org/docs/1.76/api/Bio.Align.Applications.html

[112] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004, doi: 10.1093/nar/gkh340.

[113] M. T. Pervez *et al.*, “Evaluating the accuracy and efficiency of multiple sequence alignment methods.,” *Evol Bioinform Online*, vol. 10, pp. 205–17, 2014, doi: 10.4137/EBO.S19199.

[114] C. Kemena and C. Notredame, “Upcoming challenges for multiple sequence alignment methods in the high-throughput era.,” *Bioinformatics*, vol. 25, no. 19, pp. 2455–65, Oct. 2009, doi: 10.1093/bioinformatics/btp452.

[115] “sequence alignement viewer.” [Online]. Available: https://dmnfarrell.github.io/bioinformatics/bokeh-sequence-aligner

[116] F. S. D. D. and A. W. des Higgins, “Clustal W / Clustal X.” Accessed: Jan. 31, 2023. [Online]. Available: http://www.clustal.org/clustal2/

[117] A. D. Johnson, “An extended IUPAC nomenclature code for polymorphic nucleic acids.,” *Bioinformatics*, vol. 26, no. 10, pp. 1386–9, May 2010, doi: 10.1093/bioinformatics/btq098.

[118] “IUPAC codes.” [Online]. Available: https://genome.ucsc.edu/goldenPath/help/iupac.html

[119] Wikipedia Contributors, “UPGMA,” *Wikipedia*. Wikipedia, The Free Encyclopedia., 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/UPGMA

[120] Sabine Landau, Dr Sabine Landau, Morven Leese, Daniel Stahl, Brian S. Everitt, and Dr Morven Leese, *Cluster Analysis*, 5th ed. John Wiley & Sons, Incorporated, 2011.

[121] Basel Abu-Jamous, Rui Fa, and Asoke K. Nandi, *Integrative Cluster Analysis in Bioinformatics*. John Wiley & Sons, Incorporated, 2015.

[122] M. Weiß and M. Göker, "Molecular Phylogenetic Reconstruction," in *The Yeasts*, Elsevier, 2011, pp. 159–174. doi: 10.1016/B978-0-444-52149-1.00012-4.

[123] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi, "How to control confounding effects by statistical analysis.," *Gastroenterol Hepatol Bed Bench*, vol. 5, no. 2, pp. 79–83, 2012.

[124] L. Thomas, "Confounding Variables | Definition, Examples & Controls." [Online]. Available: https://www.scribbr.com/methodology/confounding-variables/

[125] Wikipedia Contributors, "Confounding." Wikipedia, The Free Encyclopedia. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Confounding

[126] Wikipedia Contributors, "Fisher's exact test," *Wikipedia*. Wikipedia, 2023. Accessed: Feb. 20, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Fisher%27s_exact_test#

[127] J. H. 2014. McDonald, "Fisher's exact test of independence," in *Handbook of Biological Statistics* , 3rd ed., M. Baltimore, Ed.

[128] P. Bhandari and P. Bhandari, "Independent vs. Dependent Variables | Definition & Examples." [Online]. Available: https://www.scribbr.com/methodology/independent-and-dependent-variables/

[129] Babraham Institute., "Babraham Bioinformatics fastqc." [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[130] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report.," *Bioinformatics*, vol. 32, no. 19, pp. 3047–8, 2016, doi: 10.1093/bioinformatics/btw354.

[131] Wikipedia Contributors, "DNA and RNA codon tables." 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/DNA_and_RNA_codon_tables

[132] A. D. Johnson, "An extended IUPAC nomenclature code for polymorphic nucleic acids.," *Bioinformatics*, vol. 26, no. 10, pp. 1386–9, May 2010, doi: 10.1093/bioinformatics/btq098.

[133] Wikipedia Contributors, "Consensus sequence." 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Consensus_sequence

[134] J. A. Nasir *et al.*, "A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture," *Viruses*, vol. 12, no. 8, p. 895, Aug. 2020, doi: 10.3390/v12080895.

[135] Derek Caetano-Anolles, "Phred-scaled quality scores." Accessed: Feb. 13, 2023. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360035531872-Phred-scaled-quality-scores

[136] "IUPAC codes", [Online]. Available: https://www.bioinformatics.org/sms/iupac.html

[137] Wikipedia Contributors, "Coverage (genetics)." Wikipedia, The Free Encyclopedia., 2023. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Coverage_(genetics)

[138] Wikipedia Contributors, “DNA replication.” Wikipedia, The Free Encyclopedia., 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=DNA_replication&oldid=1113947229

[139] Illumina Inc, “Understanding Illumina Quality Scores”, Accessed: Apr. 11, 2023. [Online]. Available: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

[140] Illumina inc, “Quality Scores for Next-Generation Sequencing ,” 2011, Accessed: Apr. 12, 2023. [Online]. Available: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

[141] C. Delahaye and J. Nicolas, “Sequencing DNA with nanopores: Troubles and biases,” *PLoS One*, vol. 16, no. 10, p. e0257521, Oct. 2021, doi: 10.1371/journal.pone.0257521.

[142] Illumina Inc, “Measuring sequencing accuracy,” 2023, Accessed: Apr. 12, 2023. [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html#:~:text=Illumina%20Sequencing%20Quality%20Scores,sequencing%20applications%2C%20including%20clinical%20research

[143] “Fastqc.” [Online]. Available: https://scienceparkstudygroup.github.io/rna-seq-lesson/03-qc-of-sequencing-results/index.html

[144] “FastQC Tutorial & FAQ.” [Online]. Available: https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

[145] “about_fastqc_aggregate_report.” [Online]. Available: https://gau.ccr.cancer.gov/tools/about_fastqc_aggregate_report/

[146] "FastQC_Manual." [Online]. Available: https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

[147] Babraham Bioinformatics group, *Using fastQC to check the quality of high throughput sequence*. [Online Video]. Available: https://www.youtube-nocookie.com/embed/bz93ReOv87Y?rel=0

[148] Wikipedia contributors, "Transition (genetics)." Wikipedia, The Free Encyclopedia., 2022.

[149] F. Vogel and M. Kopun, "Higher frequencies of transitions among point mutations," *J Mol Evol*, vol. 9, no. 2, pp. 159–180, Jun. 1977, doi: 10.1007/BF01732746.

[150] F. Pfeiffer *et al.*, "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Sci Rep*, vol. 8, no. 1, p. 10950, Dec. 2018, doi: 10.1038/s41598-018-29325-6.

[151] A. E. Pérez-Cobas, L. Gomez-Valero, and C. Buchrieser, "Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses.," *Microb Genom*, vol. 6, no. 8, 2020, doi: 10.1099/mgen.0.000409.

[152] N. I. of H. (NIH), *The New Science of Metagenomics*. Washington, D.C.: National Academies Press, 2007. doi: 10.17226/11902.

[153] N. Drou *et al.*, "Metagenomics." [Online]. Available: https://learn.gencore.bio.nyu.edu/metgenomics/

[154] A. Ameur, "Goodbye reference, hello genome graphs," *Nat Biotechnol*, vol. 37, no. 8, pp. 866–868, Aug. 2019, doi: 10.1038/s41587-019-0199-7.

[155] R. M. Colquhoun *et al.*, “Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs,” *Genome Biol*, vol. 22, no. 1, p. 267, Dec. 2021, doi: 10.1186/s13059-021-02473-1.

[156] S. Nusrat, T. Harbig, and N. Gehlenborg, “Tasks, Techniques, and Tools for Genomic Data Visualization,” *Computer Graphics Forum*, vol. 38, no. 3, pp. 781–805, Jun. 2019, doi: 10.1111/cgf.13727.

[157] Wikipedia Contributors, “Suffix array.” Wikipedia, The Free Encyclopedia., 2022. Accessed: Mar. 22, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Suffix_array

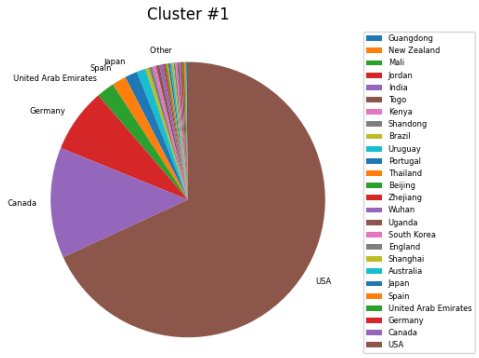
[158] Graham, “Compute Canada,”https://ccdb.computecanada.ca/security/login. Accessed: Jan. 24, 2023. [Online]. Available: https://ccdb.computecanada.ca/security/login

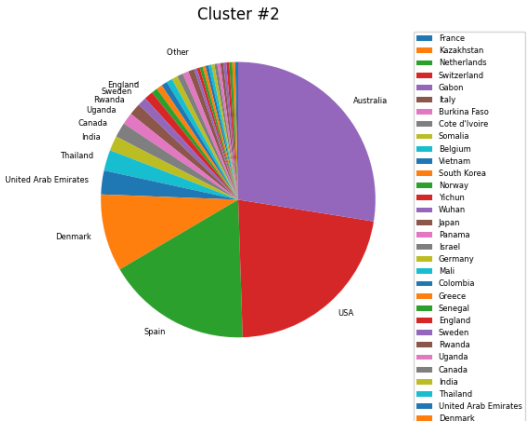
[159] S. J. Balin and M. Cascalho, “The rate of mutation of a single gene,” *Nucleic Acids Res*, vol. 38, no. 5, pp. 1575–1582, Mar. 2010, doi: 10.1093/nar/gkp1119.

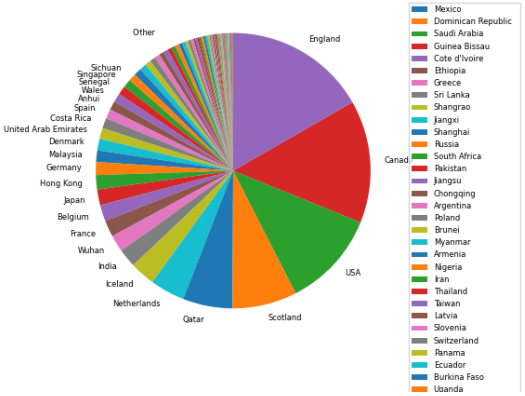
Appendix A: Sequence alignment viewer modified code

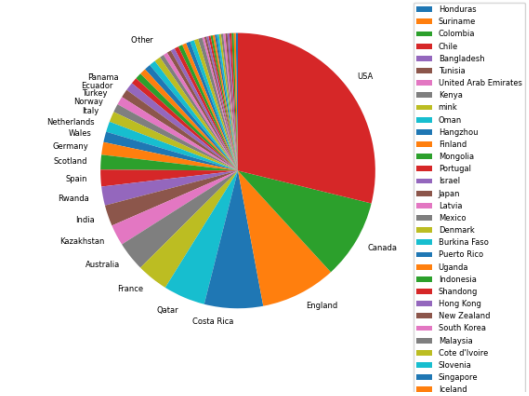
```
def make_seq(length=40):
    return
".join([random.choice(['A','C','T','G','Y','S','W','K','R','M','H','D','B','V','N']) for i in
range(length)])


def mutate_seq(seq):
    """mutate a sequence randomly"""
    seq = list(seq)
    pos = np.random.randint(1,len(seq),17)
    for i in pos:
        seq[i] = random.choice(['A','C','T','G','Y','S','W','K','R','M','H','D','B','V','N'])
    return ".join(seq)


def get_colors(seqs):
    """make colors for bases in sequence"""
    text = [i for s in list(seqs) for i in s]
    clrs =  {'A':'red','T':'green','G':'orange','C':'blue','-
':'white','N':'white','Y':'peach','S':'gold','W':'black','K':'eggplant','R':'slategrey','M':'pe
ru','H':'silver','D':'palegreen','B':'greenyellow','V':'aqua'}
    colors = [clrs[i] for i in text]
    return colors
```

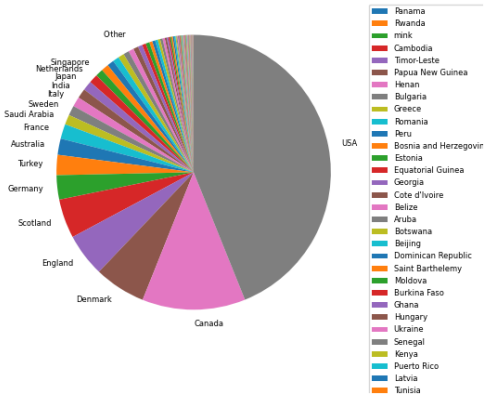
Appendix B: Distribution of COVID-19 variants in different cities/countries in different clusters.

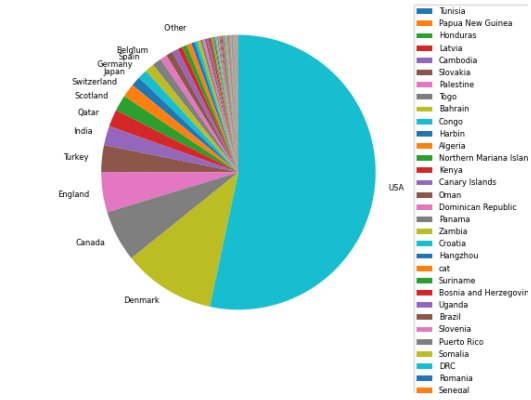

Cluster #1
Other
Japan
Spain
United Arab Emirates
Germany
Canada
USA
Guangdong
New Zealand
Mali
Jordan
India
Togo
Kenya
Shandong
Brazil
Uruguay
Portugal
Thailand
Beijing
Zhejiang
Wuhan
Uganda
South Korea
England
Shanghai
Australia
Japan
Spain
United Arab Emirates
Germany
Canada
USA


Cluster #2
Other
England
Sweden
Rwanda
Uganda
Canada
India
Thailand
United Arab Emirates
Denmark
Spain
USA
Australia
France
Kazakhstan
Netherlands
Switzerland
Gabon
Italy
Burkina Faso
Cote d'Ivoire
Somalia
Belgium
Vietnam
South Korea
Norway
Yichun
Wuhan
Japan
Panama
Israel
Germany
Mali
Colombia
Greece
Senegal
England
Sweden
Rwanda
Uganda
Canada
India
Thailand
United Arab Emirates
Denmark

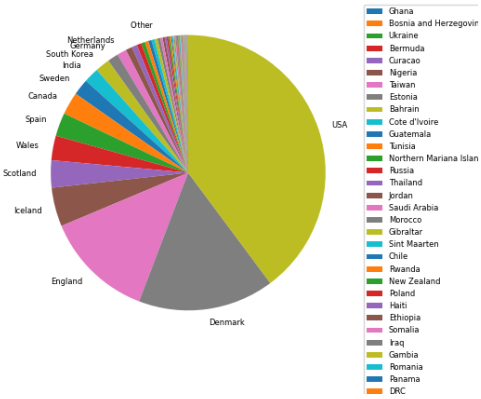

Cluster #3
Other
England
Sichuan
Singapore
Senegal
Wales
Anhui
Spain
Costa Rica
United Arab Emirates
Denmark
Malaysia
Germany
Hong Kong
Japan
Belgium
France
Wuhan
India
Iceland
Netherlands
Qatar
Scotland
USA
Canad
Mexico
Dominican Republic
Saudi Arabia
Guinea Bissau
Cote d'Ivoire
Ethiopia
Greece
Sri Lanka
Shangrao
Jiangxi
Shanghai
Russia
South Africa
Pakistan
Jiangsu
Chongqing
Argentina
Poland
Brunei
Myanmar
Armenia
Nigeria
Iran
Thailand
Taiwan
Latvia
Slovenia
Switzerland
Panama
Ecuador
Burkina Faso
Uganda


Cluster #4
Other
Panama
Ecuador
Turkey
Norway
Italy
Netherlands
Wales
Germany
Scotland
Spain
Rwanda
India
Kazakhstan
Australia
France
Qatar
Costa Rica
England
Canada
USA
Honduras
Suriname
Colombia
Chile
Bangladesh
Tunisia
United Arab Emirates
Kenya
mink
Oman
Hangzhou
Finland
Mongolia
Portugal
Israel
Japan
Latvia
Mexico
Denmark
Burkina Faso
Puerto Rico
Uganda
Indonesia
Shandong
Hong Kong
New Zealand
South Korea
Malaysia
Cote d'Ivoire
Slovenia
Singapore
Iceland

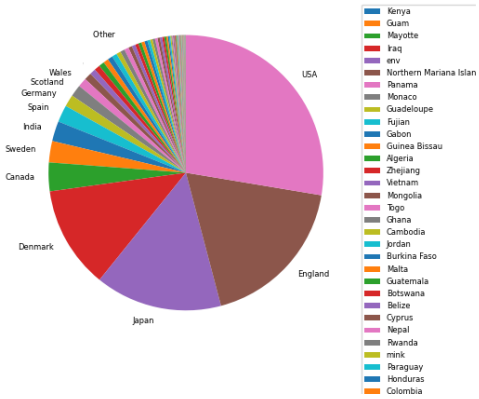

Cluster #5
Other
Singapore
Netherlands
Japan
India
Italy
Sweden
Saudi Arabia
France
Australia
Turkey
Germany
Scotland
England
Denmark
Canada
USA
Panama
Rwanda
mink
Cambodia
Timor-Leste
Papua New Guinea
Henan
Bulgaria
Greece
Romania
Peru
Bosnia and Herzegovir
Estonia
Equatorial Guinea
Georgia
Cote d'Ivoire
Belize
Aruba
Botswana
Beijing
Dominican Republic
Saint Barthelemy
Moldova
Burkina Faso
Ghana
Hungary
Ukraine
Senegal
Kenya
Puerto Rico
Latvia
Tunisia


Cluster #6
Other
Belgium
Spain
Germany
Japan
Switzerland
Scotland
Qatar
India
Turkey
England
Canada
Denmark
USA
Tunisia
Papua New Guinea
Honduras
Latvia
Cambodia
Slovakia
Palestine
Togo
Bahrain
Congo
Harbin
Algeria
Northern Mariana Islan
Kenya
Canary Islands
Oman
Dominican Republic
Panama
Zambia
Croatia
Hangzhou
cat
Suriname
Bosnia and Herzegovir
Uganda
Brazil
Slovenia
Puerto Rico
Somalia
DRC
Romania
Senegal

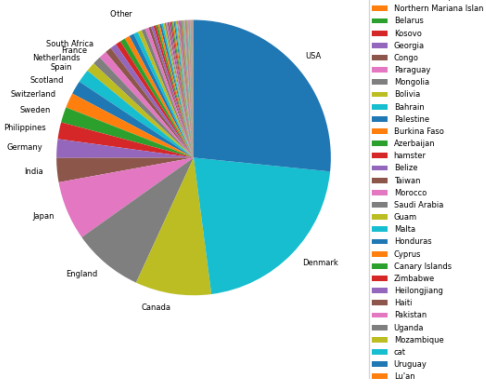
Cluster #7



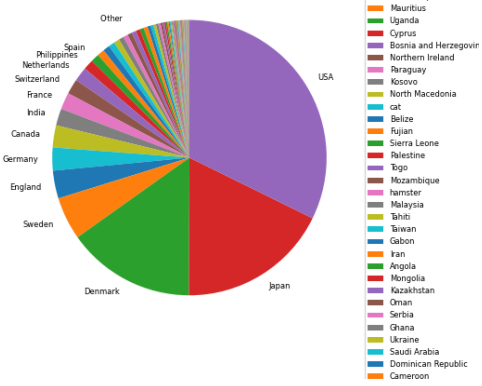
Cluster #8



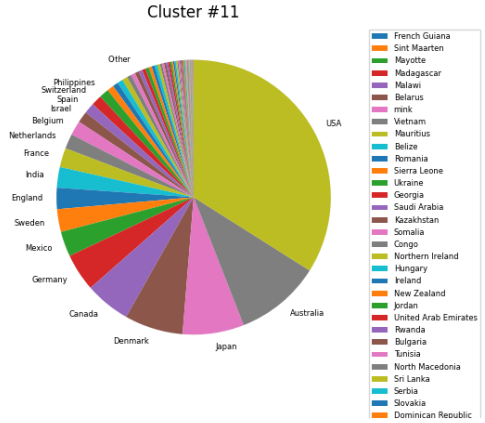
Cluster #9



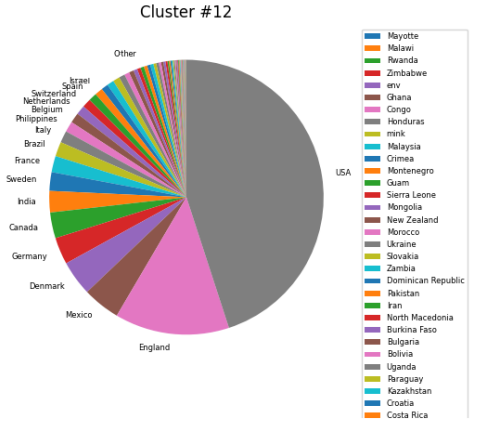
Cluster #10

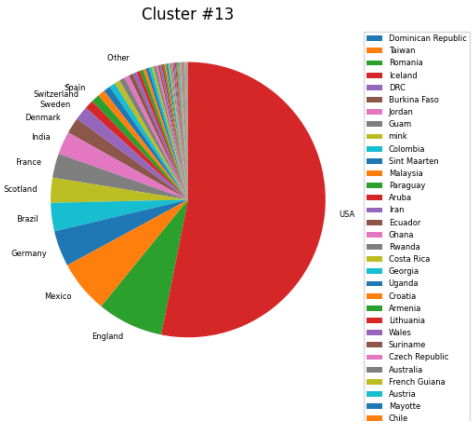


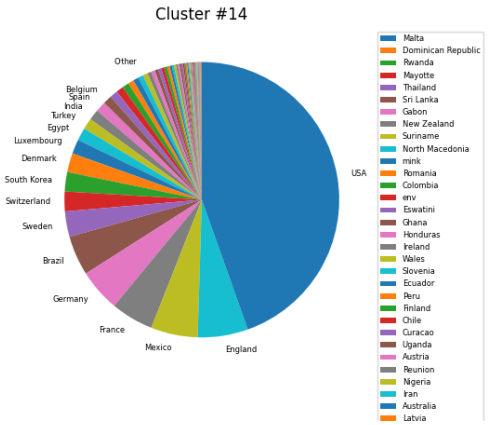
Cluster #11

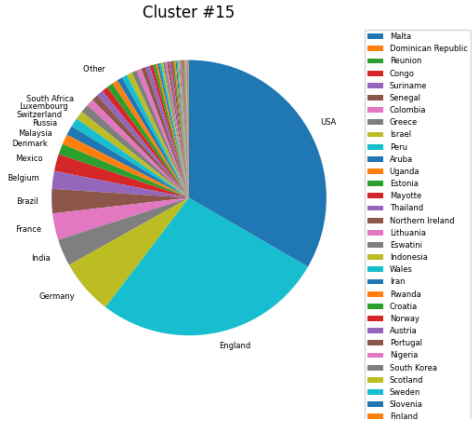


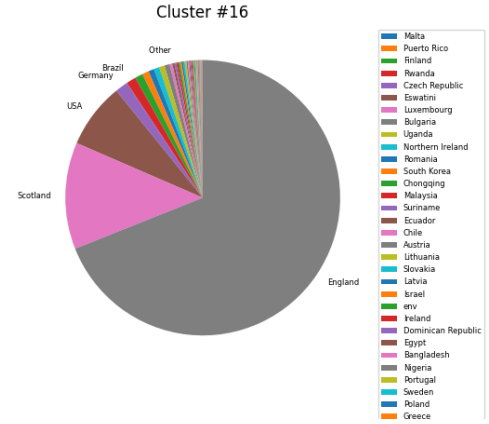
Cluster #12

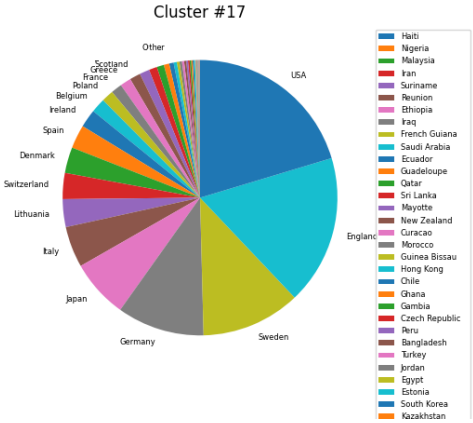


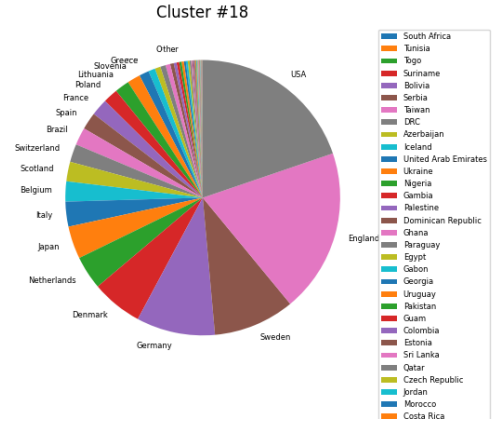

Cluster #13
Other
Spain
Switzerland
Sweden
Denmark
India
France
Scotland
Brazil
Germany
Mexico
England
USA
Dominican Republic
Taiwan
Romania
Iceland
DRC
Burkina Faso
Jordan
Guam
mink
Colombia
Sint Maarten
Malaysia
Paraguay
Aruba
Iran
Ecuador
Ghana
Rwanda
Costa Rica
Georgia
Uganda
Croatia
Armenia
Lithuania
Wales
Suriname
Czech Republic
Australia
French Guiana
Austria
Mayotte
Chile

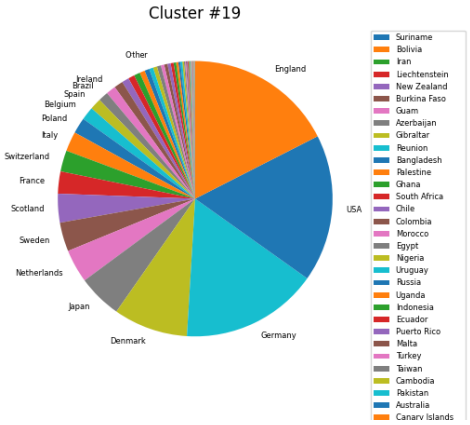

Cluster #14
Other
Belgium
Spain
India
Turkey
Egypt
Luxembourg
Denmark
South Korea
Switzerland
Sweden
Brazil
Germany
France
Mexico
England
USA
Malta
Dominican Republic
Rwanda
Mayotte
Thailand
Sri Lanka
Gabon
New Zealand
Suriname
North Macedonia
mink
Romania
Colombia
env
Eswatini
Ghana
Honduras
Ireland
Wales
Slovenia
Ecuador
Peru
Finland
Chile
Curacao
Uganda
Austria
Reunion
Nigeria
Iran
Australia
Latvia

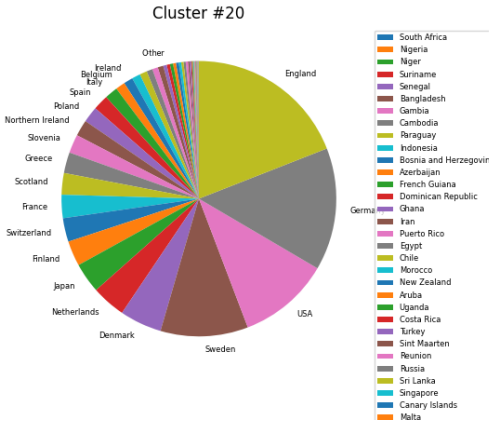

Cluster #15
Other
South Africa
Luxembourg
Switzerland
Russia
Malaysia
Denmark
Mexico
Belgium
Brazil
France
India
Germany
England
USA
Malta
Dominican Republic
Reunion
Congo
Suriname
Senegal
Colombia
Greece
Israel
Peru
Aruba
Uganda
Estonia
Mayotte
Thailand
Northern Ireland
Lithuania
Eswatini
Indonesia
Wales
Iran
Rwanda
Croatia
Norway
Austria
Portugal
Nigeria
South Korea
Scotland
Sweden
Slovenia
Finland

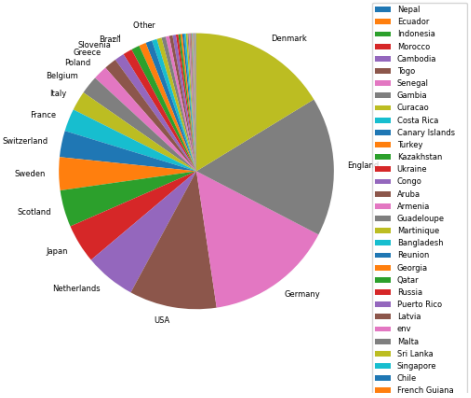

Cluster #16
Other
Brazil
Germany
USA
Scotland
England
Malta
Puerto Rico
Finland
Rwanda
Czech Republic
Eswatini
Luxembourg
Bulgaria
Uganda
Northern Ireland
Romania
South Korea
Chongqing
Malaysia
Suriname
Ecuador
Chile
Austria
Lithuania
Slovakia
Latvia
Israel
env
Ireland
Dominican Republic
Egypt
Bangladesh
Nigeria
Portugal
Sweden
Poland
Greece

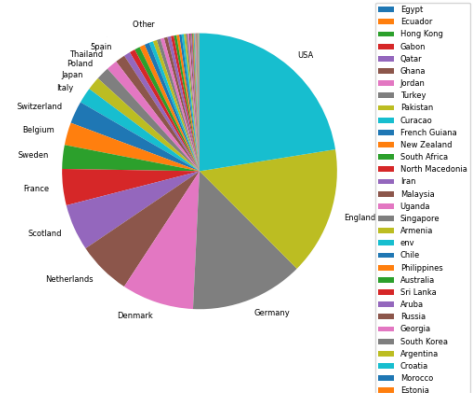

Cluster #17
Other
Scotland
Greece
France
Poland
Belgium
Ireland
Spain
Denmark
Switzerland
Lithuania
Italy
Japan
Germany
Sweden
USA
England
Haiti
Nigeria
Malaysia
Iran
Suriname
Reunion
Ethiopia
Iraq
French Guiana
Saudi Arabia
Ecuador
Guadeloupe
Qatar
Sri Lanka
Mayotte
New Zealand
Curacao
Morocco
Guinea Bissau
Hong Kong
Chile
Ghana
Gambia
Czech Republic
Peru
Bangladesh
Turkey
Jordan
Egypt
Estonia
South Korea
Kazakhstan

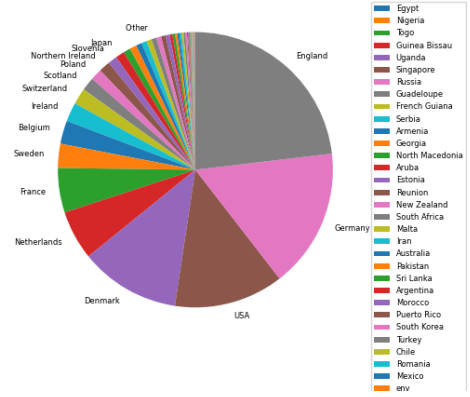

Cluster #18
Other
Greece
Slovenia
Lithuania
Poland
France
Spain
Brazil
Switzerland
Scotland
Belgium
Italy
Japan
Netherlands
Denmark
Germany
Sweden
USA
England
South Africa
Tunisia
Togo
Suriname
Bolivia
Serbia
Taiwan
DRC
Azerbaijan
Iceland
United Arab Emirates
Ukraine
Nigeria
Gambia
Palestine
Dominican Republic
Ghana
Paraguay
Egypt
Gabon
Georgia
Uruguay
Pakistan
Guam
Colombia
Estonia
Sri Lanka
Qatar
Czech Republic
Jordan
Morocco
Costa Rica

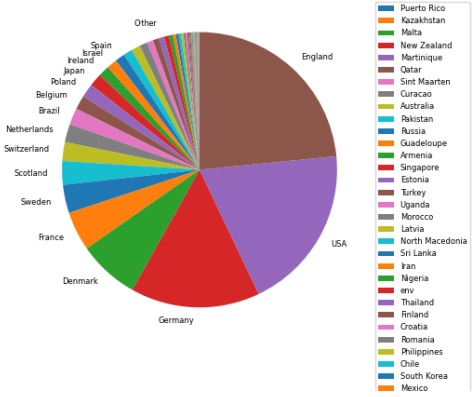

Cluster #19
Other
England
Ireland
Brazil
Spain
Belgium
Poland
Italy
Switzerland
France
Scotland
Sweden
Netherlands
Japan
Denmark
Germany
USA
Suriname
Bolivia
Iran
Liechtenstein
New Zealand
Burkina Faso
Guam
Azerbaijan
Gibraltar
Reunion
Bangladesh
Palestine
Ghana
South Africa
Chile
Colombia
Morocco
Egypt
Nigeria
Uruguay
Russia
Uganda
Indonesia
Ecuador
Puerto Rico
Malta
Turkey
Taiwan
Cambodia
Pakistan
Australia
Canary Islands

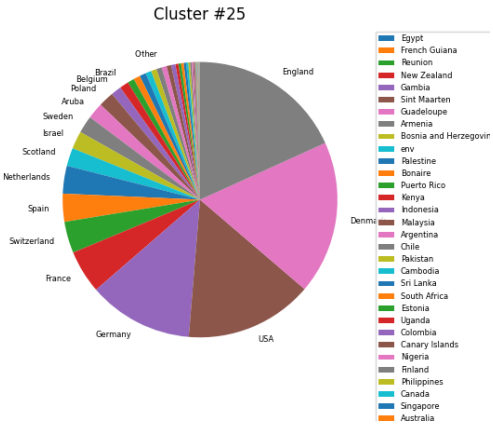

Cluster #20
Other
England
Ireland
Belgium
Italy
Spain
Poland
Northern Ireland
Slovenia
Greece
Scotland
France
Switzerland
Finland
Japan
Netherlands
Denmark
Sweden
USA
South Africa
Nigeria
Niger
Suriname
Senegal
Bangladesh
Gambia
Cambodia
Paraguay
Indonesia
Bosnia and Herzegovir
Azerbaijan
French Guiana
Dominican Republic
Ghana
Iran
Puerto Rico
Egypt
Chile
Morocco
New Zealand
Aruba
Uganda
Costa Rica
Turkey
Sint Maarten
Reunion
Russia
Sri Lanka
Singapore
Canary Islands
Malta

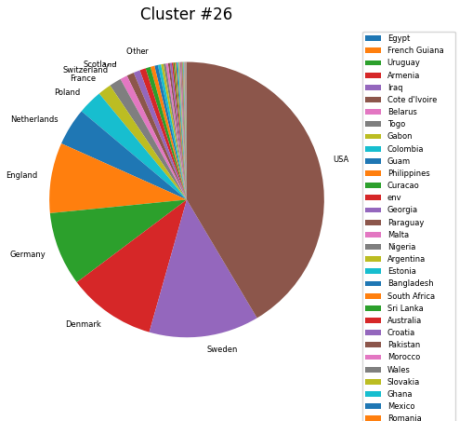

Cluster #21
Other
Brazil
Slovenia
Greece
Poland
Belgium
Italy
France
Switzerland
Sweden
Scotland
Japan
Netherlands
USA
Germany
Denmark
Nepal
Ecuador
Indonesia
Morocco
Cambodia
Togo
Senegal
Gambia
Curacao
Costa Rica
Canary Islands
Turkey
Kazakhstan
Ukraine
Congo
Aruba
Armenia
Guadeloupe
Martinique
Bangladesh
Reunion
Georgia
Qatar
Russia
Puerto Rico
Latvia
env
Malta
Sri Lanka
Singapore
Chile
French Guiana

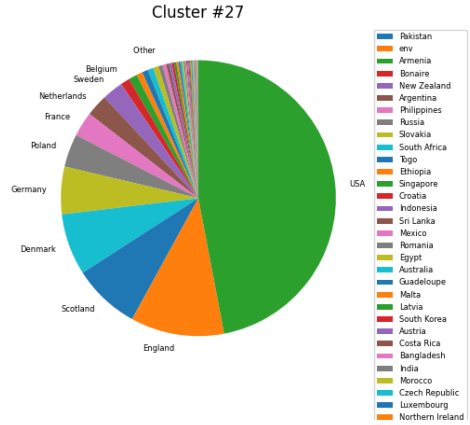

Cluster #22
Other
Spain
Thailand
Poland
Japan
Italy
Switzerland
Belgium
Sweden
France
Scotland
Netherlands
Denmark
Germany
USA
Egypt
Ecuador
Hong Kong
Gabon
Qatar
Ghana
Jordan
Turkey
Pakistan
Curacao
French Guiana
New Zealand
South Africa
North Macedonia
Iran
Malaysia
Uganda
Singapore
Armenia
env
Chile
Philippines
Australia
Sri Lanka
Aruba
Russia
Georgia
South Korea
Argentina
Croatia
Morocco
Estonia

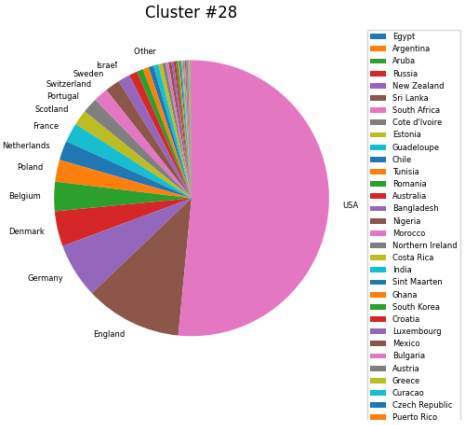

Cluster #23
Other
Japan
Slovenia
Northern Ireland
Poland
Scotland
Switzerland
Ireland
Belgium
Sweden
France
Netherlands
Denmark
USA
Germany
England
Egypt
Nigeria
Togo
Guinea Bissau
Uganda
Singapore
Russia
Guadeloupe
French Guiana
Serbia
Armenia
Georgia
North Macedonia
Aruba
Estonia
Reunion
New Zealand
South Africa
Morocco
Malta
Iran
Australia
Pakistan
Sri Lanka
Argentina
Morocco
Puerto Rico
South Korea
Turkey
Chile
Romania
Mexico
env

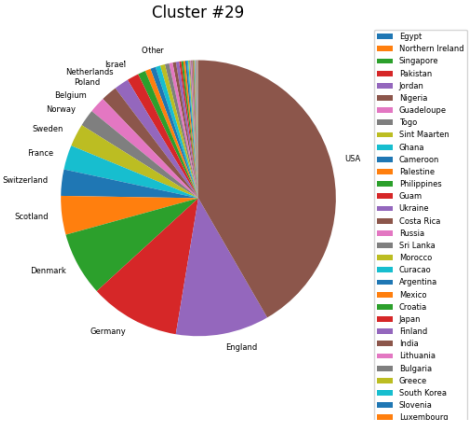

Cluster #24
Other
Spain
Israel
Ireland
Japan
Poland
Belgium
Brazil
Netherlands
Switzerland
Scotland
Sweden
France
Denmark
Germany
USA
England
Puerto Rico
Kazakhstan
Malta
New Zealand
Martinique
Qatar
Sint Maarten
Curacao
Australia
Pakistan
Russia
Guadeloupe
Armenia
Singapore
Estonia
Turkey
Uganda
Morocco
Latvia
North Macedonia
Sri Lanka
Iran
Nigeria
env
Thailand
Finland
Croatia
Romania
Philippines
Chile
South Korea
Mexico

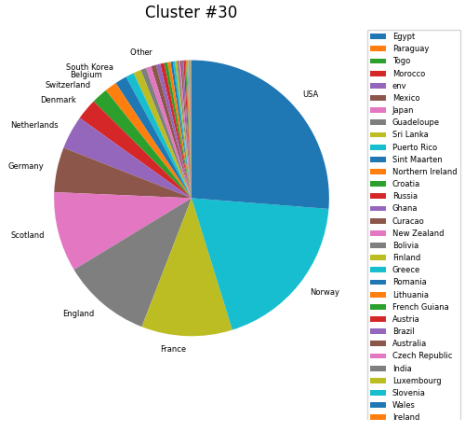

Cluster #25
Other
England
Brazil
Belgium
Poland
Aruba
Sweden
Israel
Scotland
Netherlands
Spain
Switzerland
France
Germany
USA
Egypt
French Guiana
Reunion
New Zealand
Gambia
Sint Maarten
Guadeloupe
Armenia
Bosnia and Herzegovin
env
Palestine
Bonaire
Puerto Rico
Kenya
Indonesia
Malaysia
Argentina
Chile
Pakistan
Cambodia
Sri Lanka
South Africa
Estonia
Uganda
Colombia
Canary Islands
Nigeria
Finland
Philippines
Canada
Singapore
Australia

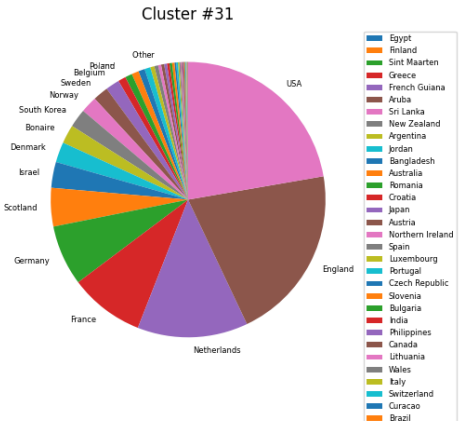

Cluster #26
Other
Scotland
Switzerland
France
Poland
Netherlands
England
Germany
Denmark
Sweden
USA
Egypt
French Guiana
Uruguay
Armenia
Iraq
Cote d'Ivoire
Belarus
Togo
Gabon
Colombia
Guam
Philippines
Curacao
env
Georgia
Paraguay
Malta
Nigeria
Argentina
Estonia
Bangladesh
South Africa
Sri Lanka
Australia
Croatia
Pakistan
Morocco
Wales
Slovakia
Ghana
Mexico
Romania

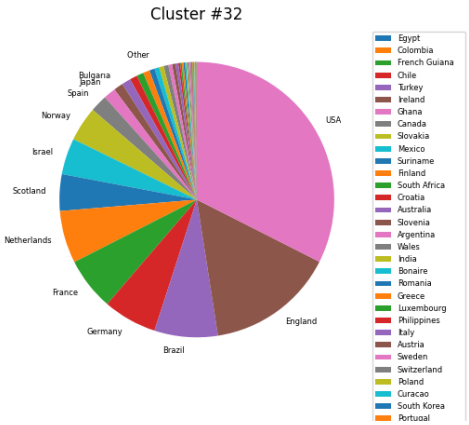

Cluster #27
Other
Belgium
Sweden
Netherlands
France
Poland
Germany
Denmark
Scotland
England
USA
Pakistan
env
Armenia
Bonaire
New Zealand
Argentina
Philippines
Russia
Slovakia
South Africa
Togo
Ethiopia
Singapore
Croatia
Indonesia
Sri Lanka
Mexico
Romania
Egypt
Australia
Guadeloupe
Malta
Latvia
South Korea
Austria
Costa Rica
Bangladesh
India
Morocco
Czech Republic
Luxembourg
Northern Ireland

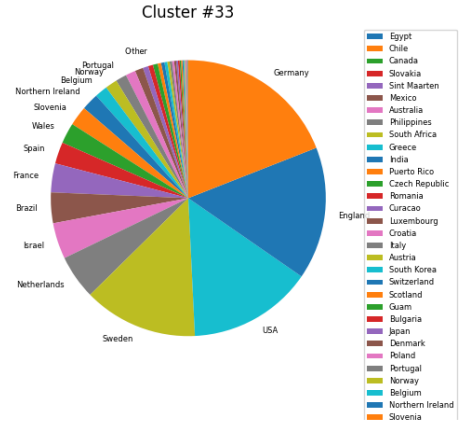

Cluster #28
Other
Israel
Sweden
Switzerland
Portugal
Scotland
France
Netherlands
Poland
Belgium
Denmark
Germany
England
USA
Egypt
Argentina
Aruba
Russia
New Zealand
Sri Lanka
South Africa
Cote d'Ivoire
Estonia
Guadeloupe
Chile
Tunisia
Romania
Australia
Bangladesh
Nigeria
Morocco
Northern Ireland
Costa Rica
India
Sint Maarten
Ghana
South Korea
Croatia
Luxembourg
Mexico
Bulgaria
Austria
Greece
Curacao
Czech Republic
Puerto Rico

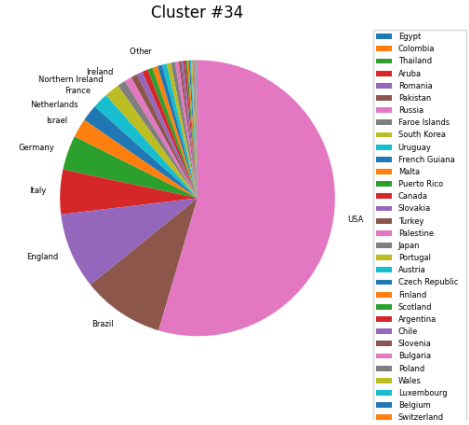

Cluster #29
Other
Israel
Netherlands
Poland
Belgium
Norway
Sweden
France
Switzerland
Scotland
Denmark
Germany
England
USA
Egypt
Northern Ireland
Singapore
Pakistan
Jordan
Nigeria
Guadeloupe
Togo
Sint Maarten
Ghana
Cameroon
Palestine
Philippines
Guam
Ukraine
Costa Rica
Russia
Sri Lanka
Morocco
Curacao
Argentina
Mexico
Croatia
Japan
Finland
India
Lithuania
Bulgaria
Greece
South Korea
Slovenia
Luxembourg

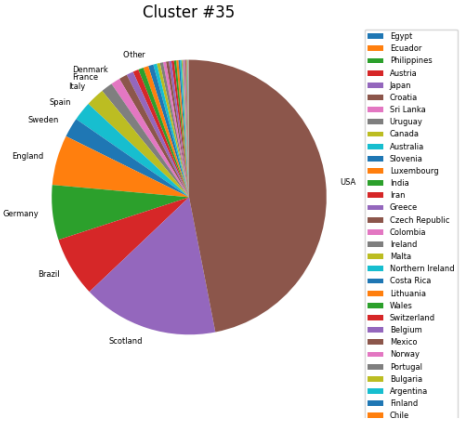

Cluster #30
Other
South Korea
Belgium
Switzerland
Denmark
Netherlands
Germany
Scotland
England
France
Norway
USA
Egypt
Paraguay
Togo
Morocco
env
Mexico
Japan
Guadeloupe
Sri Lanka
Puerto Rico
Sint Maarten
Northern Ireland
Croatia
Russia
Ghana
Curacao
New Zealand
Bolivia
Finland
Greece
Romania
Lithuania
Austria
Brazil
Australia
Czech Republic
India
Luxembourg
Slovenia
Wales
Ireland

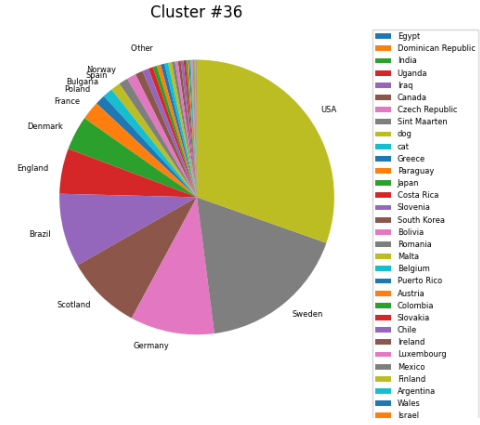

Cluster #31
Other
Poland
Belgium
Sweden
Norway
South Korea
Bonaire
Denmark
Israel
Scotland
Germany
France
Netherlands
England
USA
Egypt
Finland
Sint Maarten
Greece
French Guiana
Aruba
Sri Lanka
New Zealand
Argentina
Jordan
Bangladesh
Australia
Romania
Croatia
Japan
Austria
Northern Ireland
Spain
Luxembourg
Portugal
Czech Republic
Slovenia
Bulgaria
India
Philippines
Canada
Lithuania
Wales
Italy
Switzerland
Curacao
Brazil


Cluster #32
Other
Bulgaria
Japan
Spain
Norway
Israel
Scotland
Netherlands
France
Germany
Brazil
England
USA
Egypt
Colombia
French Guiana
Chile
Turkey
Ireland
Ghana
Canada
Slovakia
Mexico
Suriname
Finland
South Africa
Croatia
Australia
Slovenia
Argentina
Wales
India
Bonaire
Romania
Greece
Luxembourg
Philippines
Italy
Austria
Sweden
Switzerland
Poland
Curacao
South Korea
Portugal

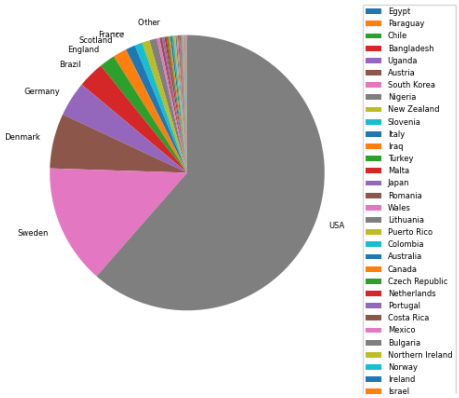

Cluster #33
Other
Portugal
Norway
Belgium
Northern Ireland
Slovenia
Wales
Spain
France
Brazil
Israel
Netherlands
Sweden
USA
Germany
Egypt
Chile
Canada
Slovakia
Sint Maarten
Mexico
Australia
Philippines
South Africa
Greece
India
Puerto Rico
Czech Republic
Romania
Curacao
Luxembourg
Croatia
Italy
Austria
South Korea
Switzerland
Scotland
Guam
Bulgaria
Japan
Denmark
Poland
Portugal
Norway
Belgium
Northern Ireland
Slovenia


Cluster #34
Other
Ireland
Northern Ireland
France
Netherlands
Israel
Germany
Italy
England
Brazil
USA
Egypt
Colombia
Thailand
Aruba
Romania
Pakistan
Russia
Faroe Islands
South Korea
Uruguay
French Guiana
Malta
Puerto Rico
Canada
Slovakia
Turkey
Palestine
Japan
Portugal
Austria
Czech Republic
Finland
Scotland
Argentina
Chile
Slovenia
Bulgaria
Poland
Wales
Luxembourg
Belgium
Switzerland

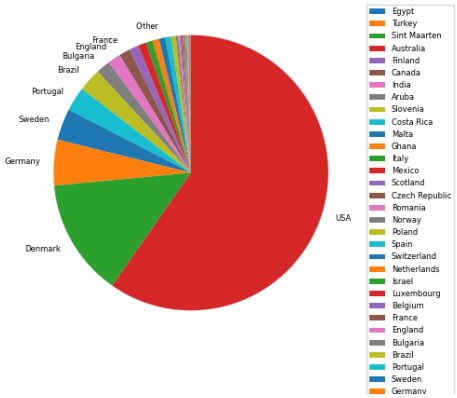

Cluster #35
Other
Denmark
France
Italy
Spain
Sweden
England
Germany
Brazil
Scotland
USA
Egypt
Ecuador
Philippines
Austria
Japan
Croatia
Sri Lanka
Uruguay
Canada
Australia
Slovenia
Luxembourg
India
Iran
Greece
Czech Republic
Colombia
Ireland
Malta
Northern Ireland
Costa Rica
Lithuania
Wales
Switzerland
Belgium
Mexico
Norway
Portugal
Bulgaria
Argentina
Finland
Chile


Cluster #36
Other
Norway
Spain
Bulgaria
Poland
France
Denmark
England
Brazil
Scotland
Germany
Sweden
USA
Egypt
Dominican Republic
India
Uganda
Iraq
Canada
Czech Republic
Sint Maarten
dog
cat
Greece
Paraguay
Japan
Costa Rica
Slovenia
South Korea
Bolivia
Romania
Malta
Belgium
Puerto Rico
Austria
Colombia
Slovakia
Chile
Ireland
Luxembourg
Mexico
Finland
Argentina
Wales
Israel

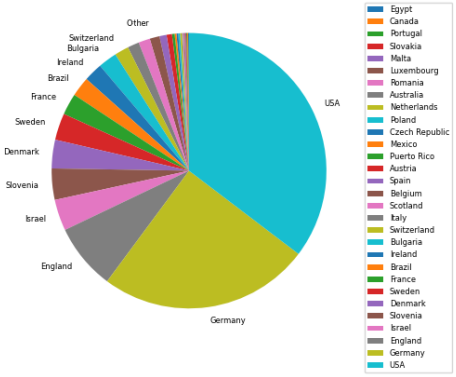
Cluster #37



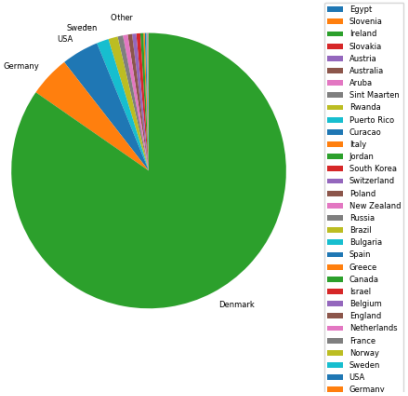
Cluster #38



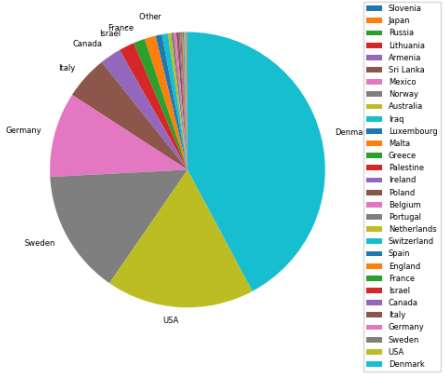
Cluster #39



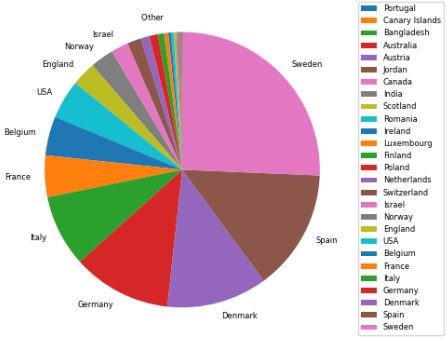
Cluster #40

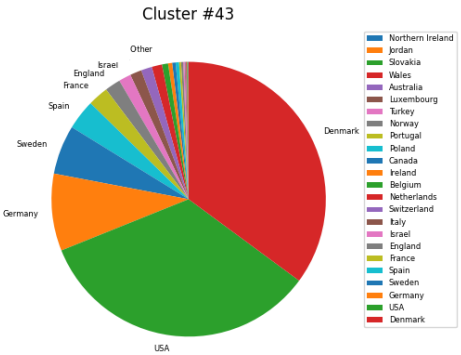


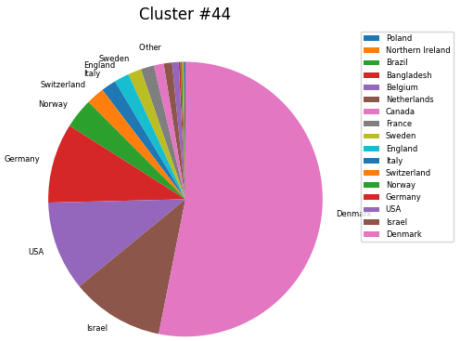
Cluster #41

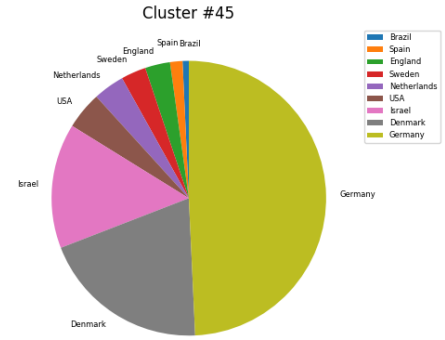


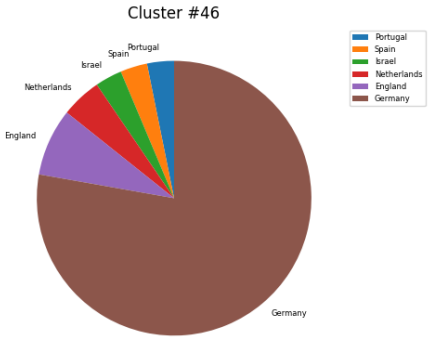
Cluster #42

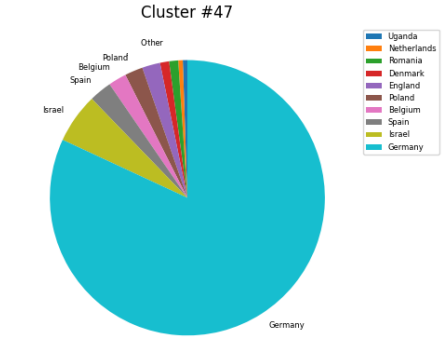


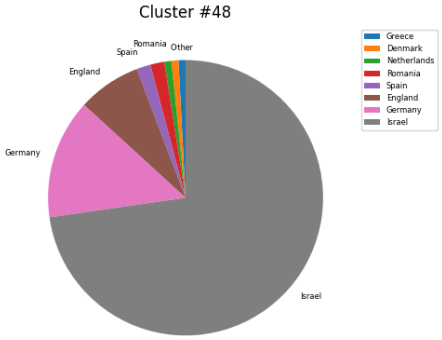

Cluster #43
Other
Israel
England
France
Spain
Sweden
Germany
USA
Denmark
Northern Ireland
Jordan
Slovakia
Wales
Australia
Luxembourg
Turkey
Norway
Portugal
Poland
Canada
Ireland
Belgium
Netherlands
Switzerland
Italy
Israel
England
France
Spain
Sweden
Germany
USA
Denmark


Cluster #44
Other
Sweden
England
Italy
Switzerland
Norway
Germany
USA
Israel
Poland
Northern Ireland
Brazil
Bangladesh
Belgium
Netherlands
Canada
France
Sweden
England
Italy
Switzerland
Norway
Germany
USA
Israel
Denmark

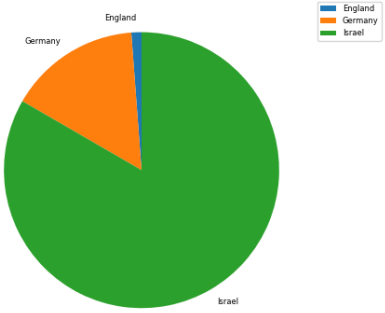

Cluster #45
Spain
Brazil
England
Sweden
Netherlands
USA
Israel
Germany
Denmark
Brazil
Spain
England
Sweden
Netherlands
USA
Israel
Denmark
Germany


Cluster #46
Portugal
Spain
Israel
Netherlands
England
Germany
Portugal
Spain
Israel
Netherlands
England
Germany

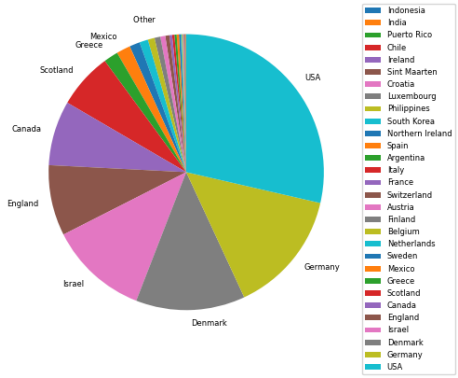

Cluster #47
Other
Poland
Belgium
Spain
Israel
Germany
Uganda
Netherlands
Romania
Denmark
England
Poland
Belgium
Spain
Israel
Germany


Cluster #48
Romania
Spain
Other
England
Germany
Israel
Greece
Denmark
Netherlands
Romania
Spain
England
Germany
Israel

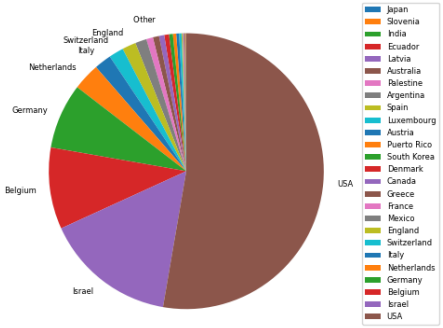
Cluster #49



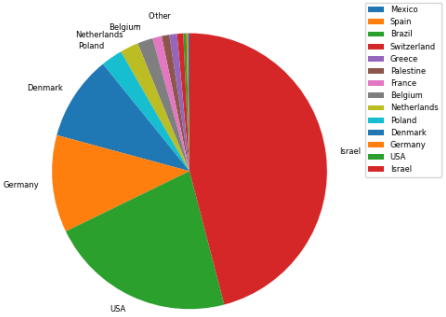
Cluster #50



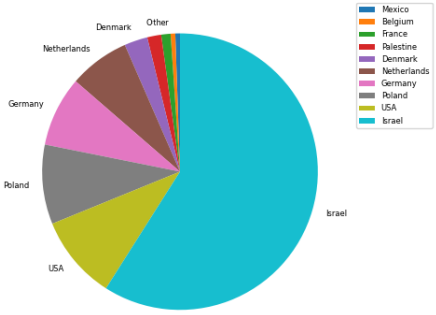
Cluster #51



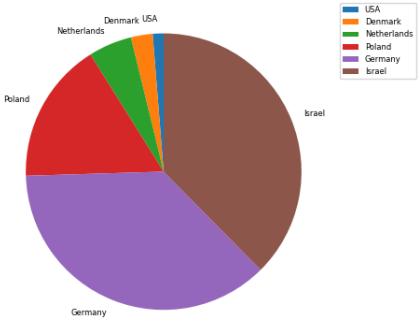
Cluster #52

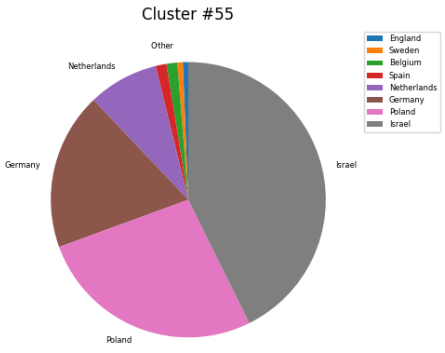


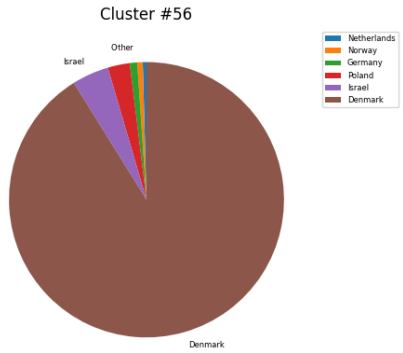
Cluster #53

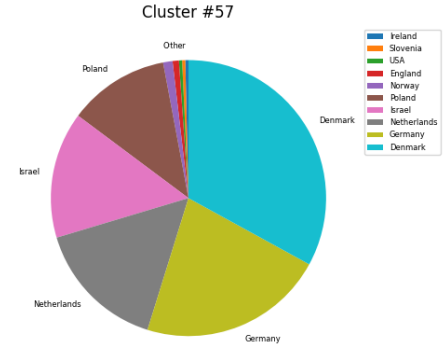


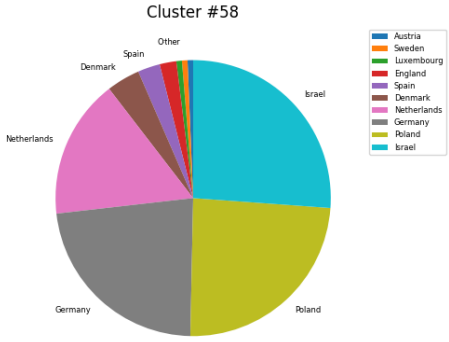
Cluster #54

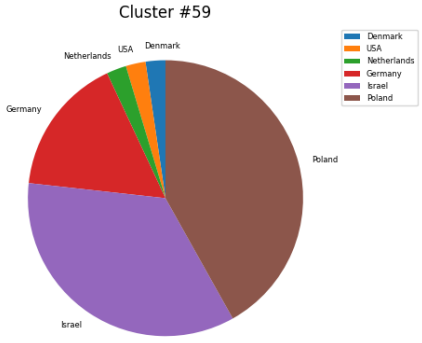


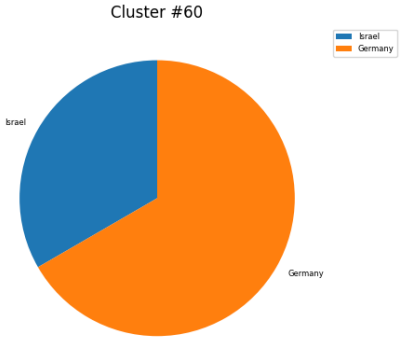

Cluster #55
Other
Netherlands
Germany
Poland
Israel
England
Sweden
Belgium
Spain
Netherlands
Germany
Poland
Israel

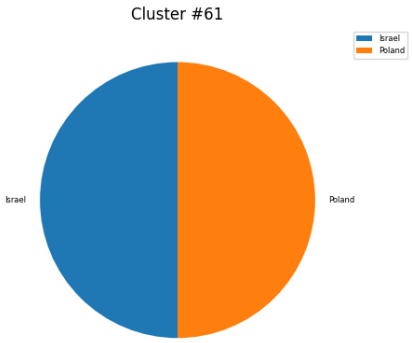

Cluster #56
Other
Israel
Denmark
Netherlands
Norway
Germany
Poland
Israel
Denmark

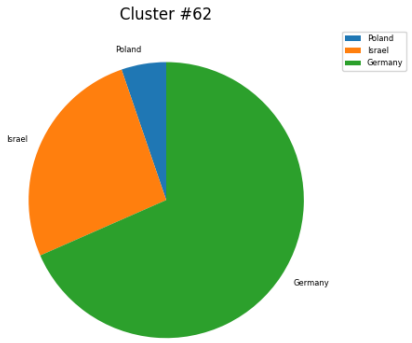

Cluster #57
Other
Poland
Denmark
Israel
Netherlands
Germany
Ireland
Slovenia
USA
England
Norway
Poland
Israel
Netherlands
Germany
Denmark

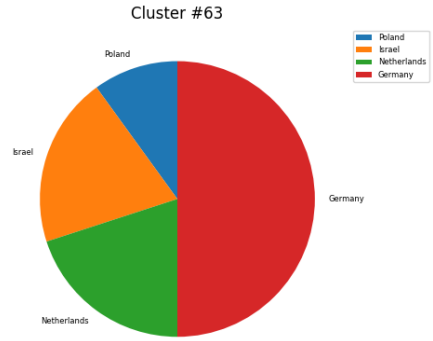

Cluster #58
Other
Spain
Denmark
Israel
Netherlands
Germany
Poland
Austria
Sweden
Luxembourg
England
Spain
Denmark
Netherlands
Germany
Poland
Israel


Cluster #59
Netherlands
USA
Denmark
Germany
Poland
Israel
Denmark
USA
Netherlands
Germany
Israel
Poland

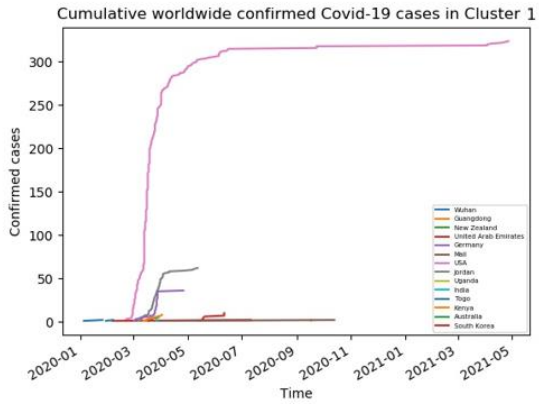

Cluster #60
Israel
Germany
Israel
Germany

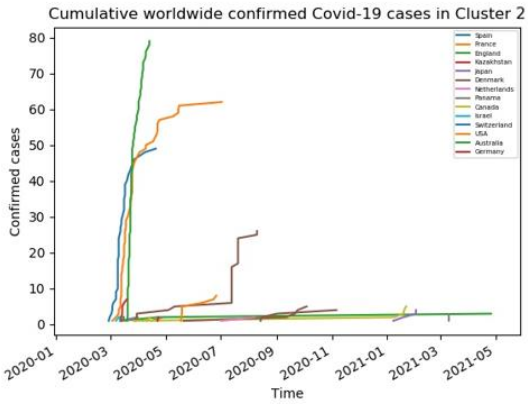

Cluster #61
Israel
Poland
Israel
Poland

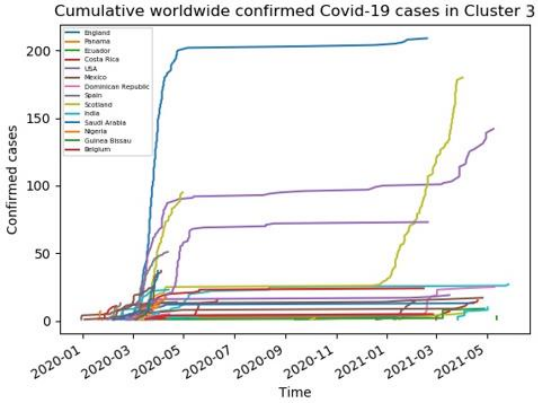

Cluster #62
Poland
Israel
Germany
Poland
Israel
Germany

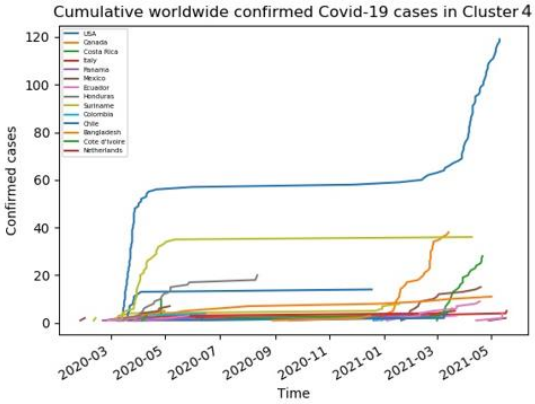

Cluster #63
Poland
Israel
Netherlands
Germany
Poland
Israel
Germany
Netherlands

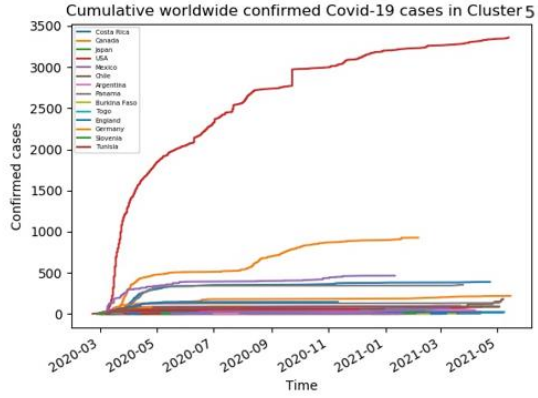
Appendix C: Cumulative diagram for clusters.

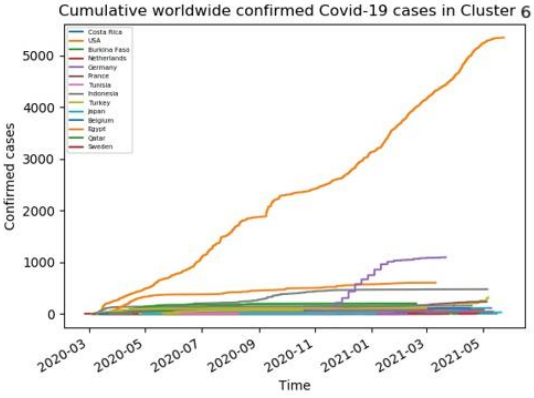

Cumulative worldwide confirmed Covid-19 cases in Cluster 1
Confirmed cases
Time
Wuhan
Guangdong
New Zealand
United Arab Emirates
Germany
Mali
USA
Jordan
Uganda
India
Togo
Kenya
Australia
South Korea

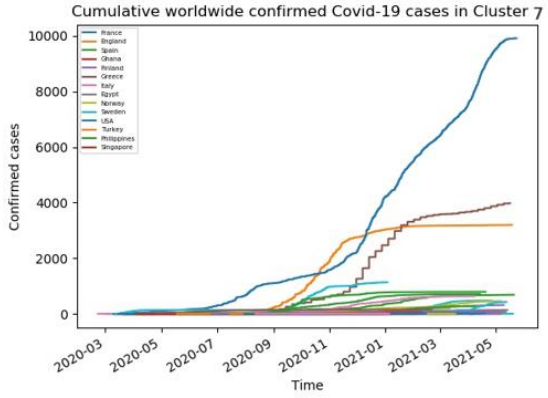

Cumulative worldwide confirmed Covid-19 cases in Cluster 2
Confirmed cases
Time
Spain
France
England
Kazakhstan
Japan
Denmark
Netherlands
Panama
Canada
Israel
Switzerland
USA
Australia
Germany

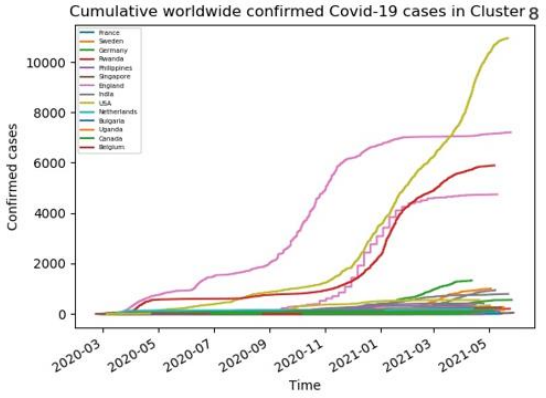

Cumulative worldwide confirmed Covid-19 cases in Cluster 3
Confirmed cases
Time
England
Panama
Ecuador
Costa Rica
USA
Mexico
Dominican Republic
Spain
Scotland
India
Saudi Arabia
Nigeria
Guinea Bissau
Belgium

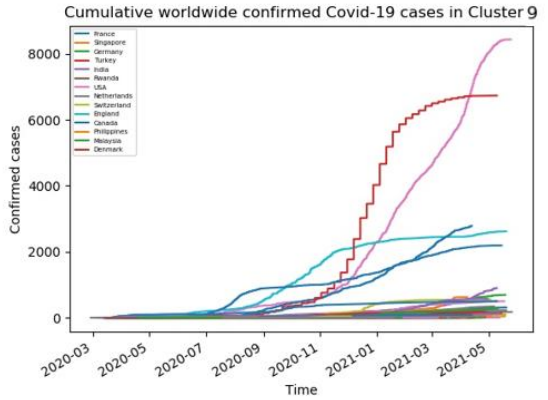

Cumulative worldwide confirmed Covid-19 cases in Cluster 4
Confirmed cases
Time
USA
Canada
Costa Rica
Italy
Panama
Mexico
Ecuador
Honduras
Suriname
Colombia
Chile
Bangladesh
Cote d'Ivoire
Netherlands

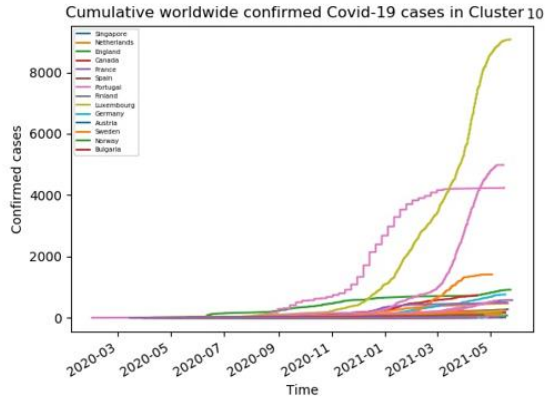

Cumulative worldwide confirmed Covid-19 cases in Cluster 5
Confirmed cases
Time
Costa Rica
Canada
Japan
USA
Mexico
Chile
Argentina
Panama
Burkina Faso
Togo
England
Germany
Slovenia
Tunisia

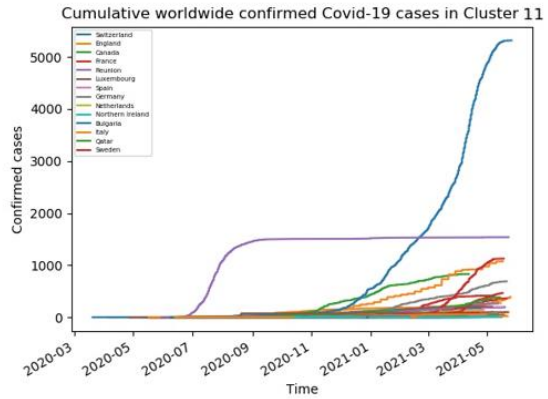

Cumulative worldwide confirmed Covid-19 cases in Cluster 6
Confirmed cases
Time
Costa Rica
USA
Burkina Faso
Netherlands
Germany
France
Tunisia
Indonesia
Turkey
Japan
Belgium
Egypt
Qatar
Sweden

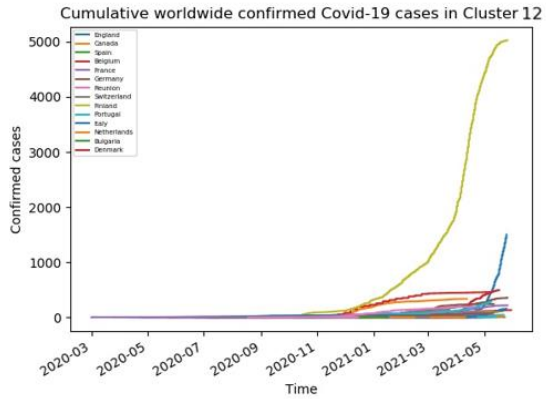

Cumulative worldwide confirmed Covid-19 cases in Cluster 7
Confirmed cases
Time
France
England
Spain
Ghana
Finland
Greece
Italy
Egypt
Norway
Sweden
USA
Turkey
Philippines
Singapore

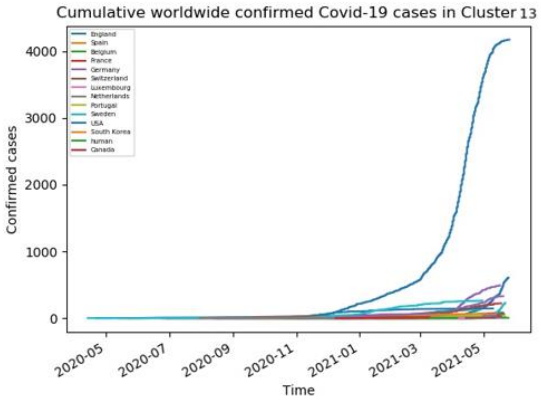

Cumulative worldwide confirmed Covid-19 cases in Cluster 8
Confirmed cases
Time
France
Sweden
Germany
Rwanda
Philippines
Singapore
England
India
USA
Netherlands
Bulgaria
Uganda
Canada
Belgium

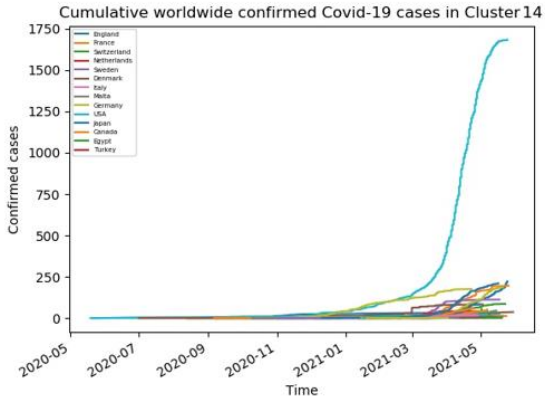

Cumulative worldwide confirmed Covid-19 cases in Cluster 9
Confirmed cases
Time
France
Singapore
Germany
Turkey
India
Rwanda
USA
Netherlands
Switzerland
England
Canada
Philippines
Malaysia
Denmark

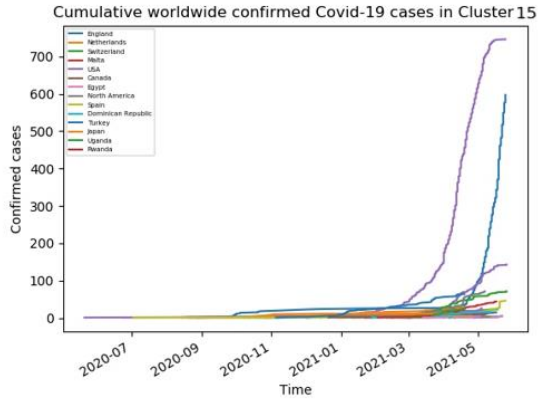

Cumulative worldwide confirmed Covid-19 cases in Cluster 10
Confirmed cases
Time
Singapore
Netherlands
England
Canada
France
Spain
Portugal
Finland
Luxembourg
Germany
Austria
Sweden
Norway
Bulgaria

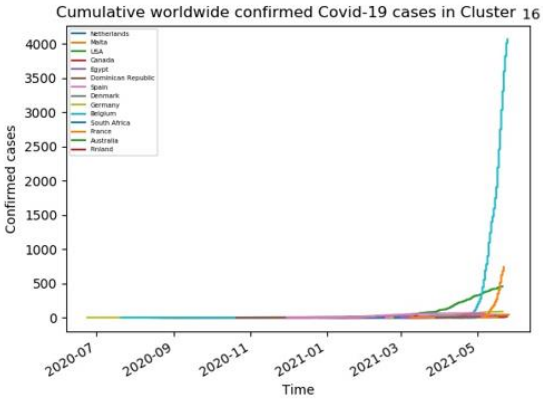

Cumulative worldwide confirmed Covid-19 cases in Cluster 11
Confirmed cases
Time
Switzerland
England
Canada
France
Reunion
Luxembourg
Spain
Germany
Netherlands
Northern Ireland
Bulgaria
Italy
Qatar
Sweden

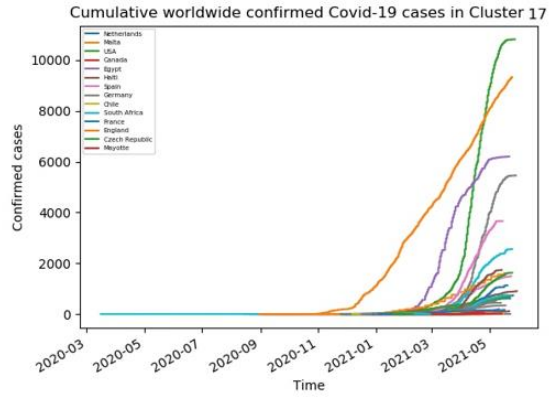

Cumulative worldwide confirmed Covid-19 cases in Cluster 12
Confirmed cases
Time
England
Canada
Spain
Belgium
France
Germany
Reunion
Switzerland
Finland
Portugal
Italy
Netherlands
Bulgaria
Denmark

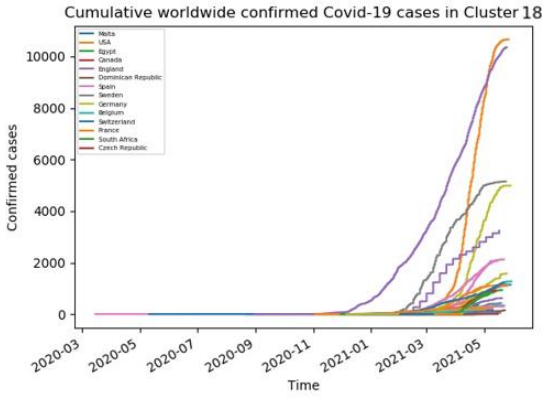

Cumulative worldwide confirmed Covid-19 cases in Cluster 13
Confirmed cases
Time
England
Spain
Belgium
France
Germany
Switzerland
Luxembourg
Netherlands
Portugal
Sweden
USA
South Korea
human
Canada

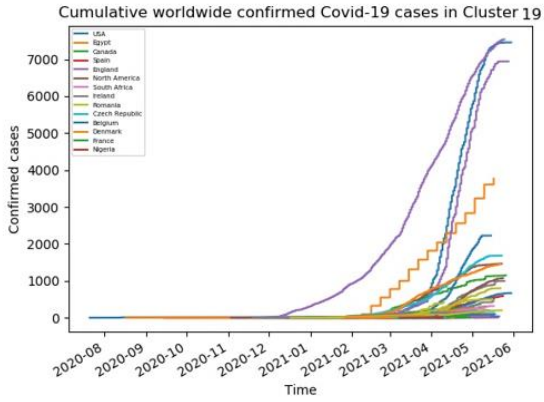

Cumulative worldwide confirmed Covid-19 cases in Cluster 14
Confirmed cases
Time
England
France
Switzerland
Netherlands
Sweden
Denmark
Italy
Malta
Germany
USA
Japan
Canada
Egypt
Turkey

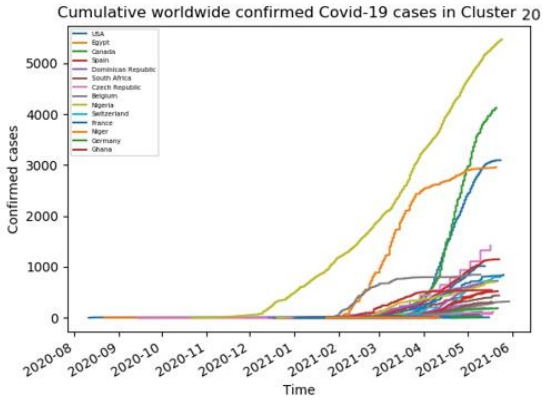

Cumulative worldwide confirmed Covid-19 cases in Cluster 15
Confirmed cases
Time
England
Netherlands
Switzerland
Malta
USA
Canada
Egypt
North America
Spain
Dominican Republic
Turkey
Japan
Uganda
Rwanda

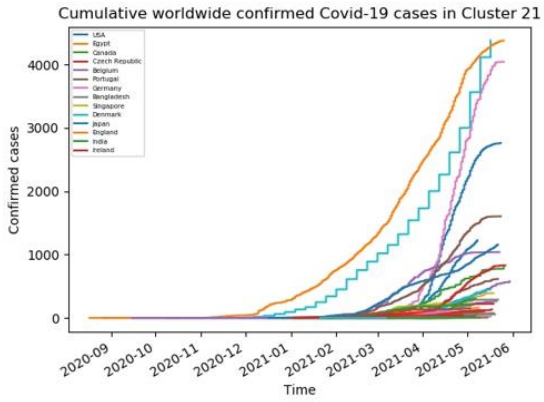

Cumulative worldwide confirmed Covid-19 cases in Cluster 16
Confirmed cases
Time
Netherlands
Malta
USA
Canada
Egypt
Dominican Republic
Spain
Denmark
Germany
Belgium
South Africa
France
Australia
Finland

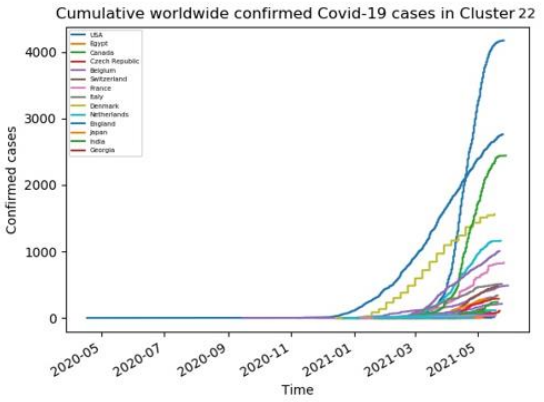

Cumulative worldwide confirmed Covid-19 cases in Cluster 17
Confirmed cases
Time
Netherlands
Malta
USA
Canada
Egypt
Haiti
Spain
Germany
Chile
South Africa
France
England
Czech Republic
Mayotte

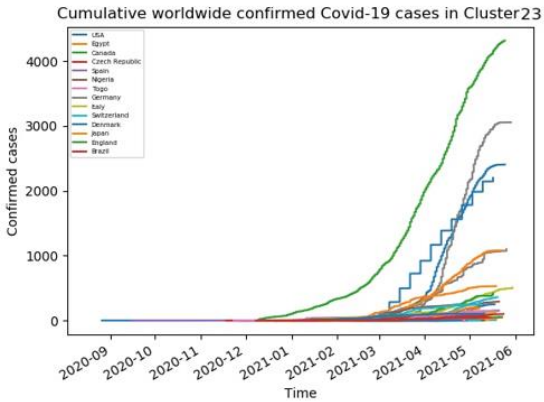

Cumulative worldwide confirmed Covid-19 cases in Cluster 18
Confirmed cases
Time
Malta
USA
Egypt
Canada
England
Dominican Republic
Spain
Sweden
Germany
Belgium
Switzerland
France
South Africa
Czech Republic

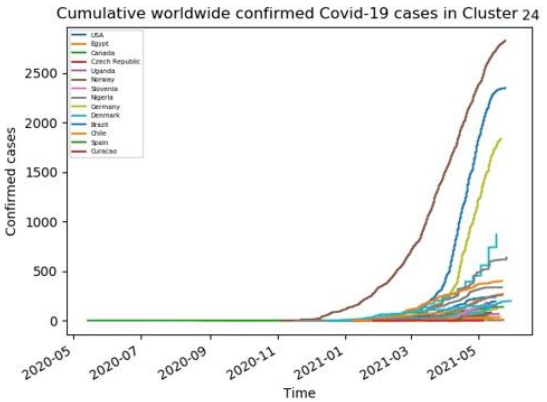

Cumulative worldwide confirmed Covid-19 cases in Cluster 19
Confirmed cases
Time
USA
Egypt
Canada
Spain
England
North America
South Africa
Ireland
Romania
Czech Republic
Belgium
Denmark
France
Nigeria

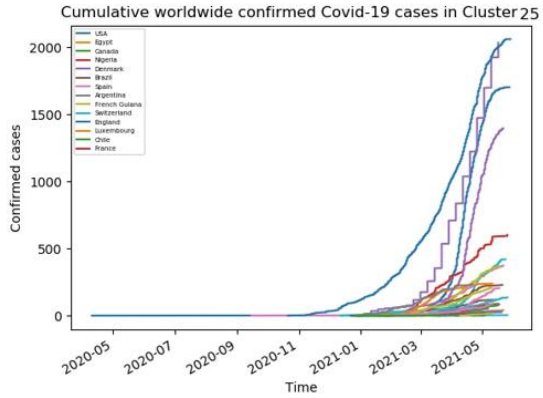

Cumulative worldwide confirmed Covid-19 cases in Cluster 20
Confirmed cases
Time
USA
Egypt
Canada
Spain
Dominican Republic
South Africa
Czech Republic
Belgium
Nigeria
Switzerland
France
Niger
Germany
Ghana

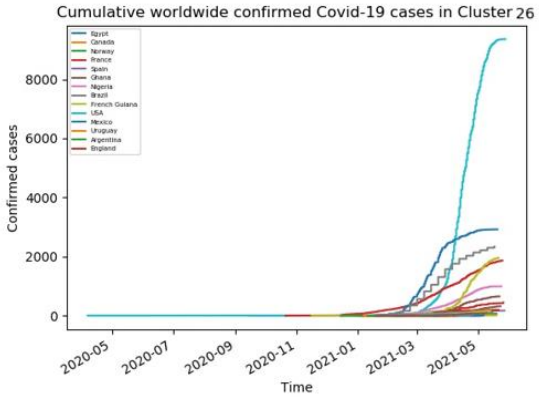

Cumulative worldwide confirmed Covid-19 cases in Cluster 21
Confirmed cases
Time
USA
Egypt
Canada
Czech Republic
Belgium
Portugal
Germany
Bangladesh
Singapore
Denmark
Japan
England
India
Ireland

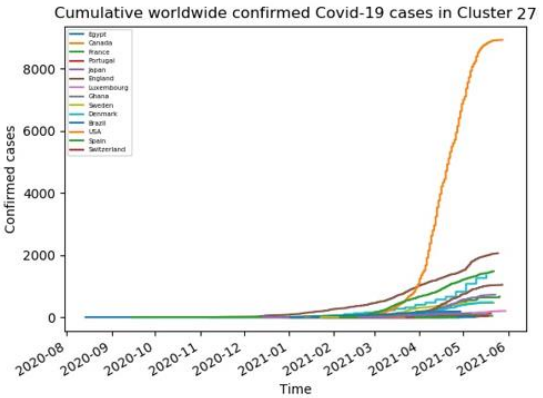

Cumulative worldwide confirmed Covid-19 cases in Cluster 22
Confirmed cases
Time
USA
Egypt
Canada
Czech Republic
Belgium
Switzerland
France
Italy
Denmark
Netherlands
England
Japan
India
Georgia

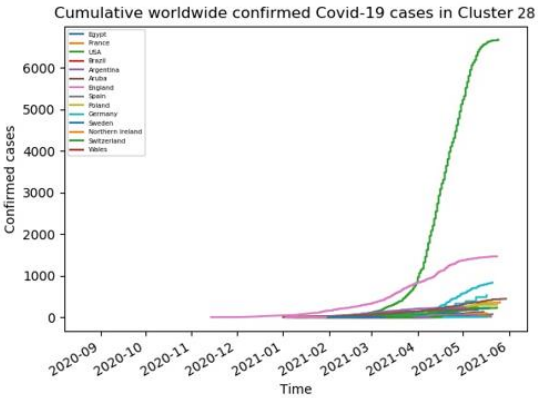

Cumulative worldwide confirmed Covid-19 cases in Cluster 23
Confirmed cases
Time
USA
Egypt
Canada
Czech Republic
Spain
Nigeria
Togo
Germany
Italy
Switzerland
Denmark
Japan
England
Brazil

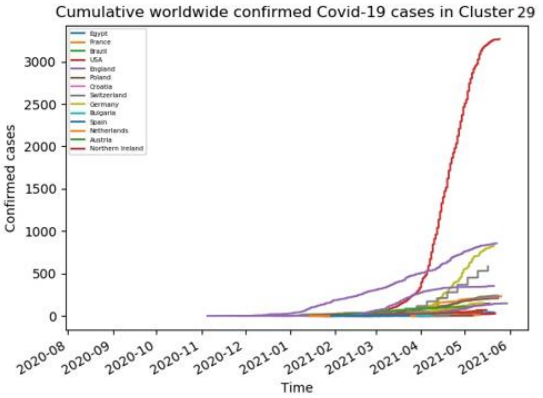

Cumulative worldwide confirmed Covid-19 cases in Cluster 24
Confirmed cases
Time
USA
Egypt
Canada
Czech Republic
Uganda
Norway
Slovenia
Nigeria
Germany
Denmark
Brazil
Chile
Spain
Curacao

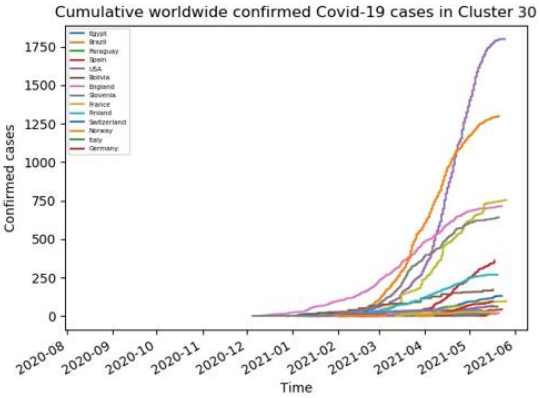

Cumulative worldwide confirmed Covid-19 cases in Cluster 25
Confirmed cases
Time
USA
Egypt
Canada
Nigeria
Denmark
Brazil
Spain
Argentina
French Guiana
Switzerland
England
Luxembourg
Chile
France

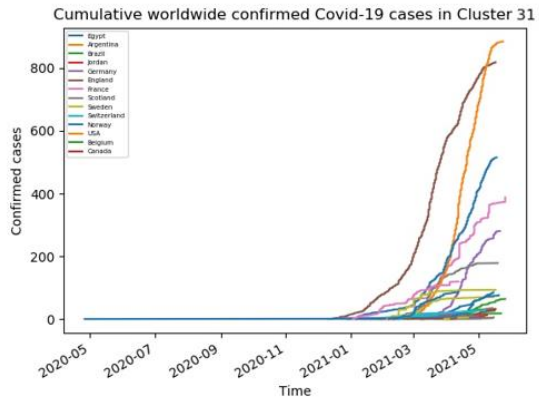

Cumulative worldwide confirmed Covid-19 cases in Cluster 26
Confirmed cases
Time
Egypt
Canada
Norway
France
Spain
Ghana
Nigeria
Brazil
French Guiana
USA
Mexico
Uruguay
Argentina
England

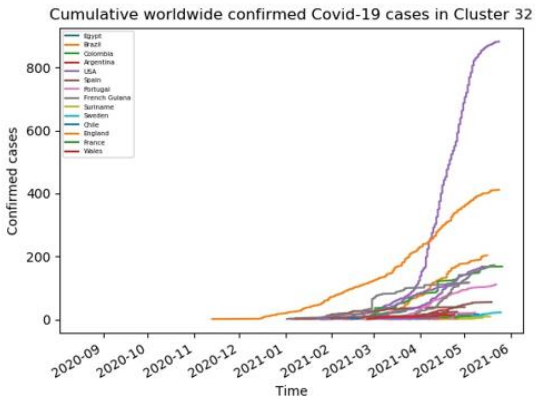

Cumulative worldwide confirmed Covid-19 cases in Cluster 27
Confirmed cases
Time
Egypt
Canada
France
Portugal
Japan
England
Luxembourg
Ghana
Sweden
Denmark
Brazil
USA
Spain
Switzerland

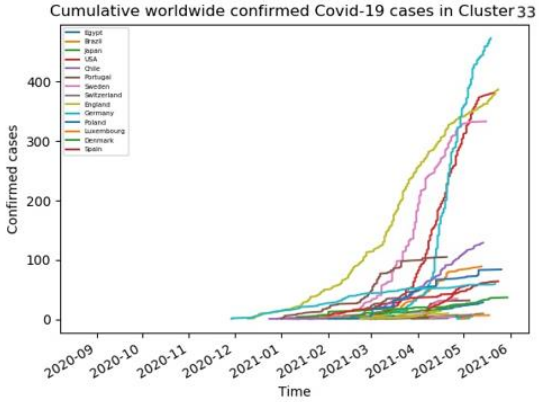

Cumulative worldwide confirmed Covid-19 cases in Cluster 28
Confirmed cases
Time
Egypt
France
USA
Brazil
Argentina
Aruba
England
Spain
Poland
Germany
Sweden
Northern Ireland
Switzerland
Wales

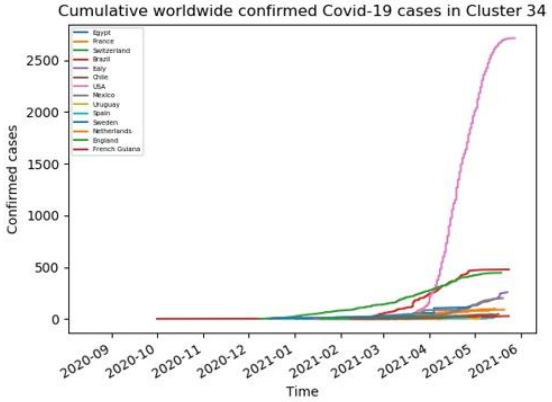

Cumulative worldwide confirmed Covid-19 cases in Cluster 29
Confirmed cases
Time
Egypt
France
Brazil
USA
England
Poland
Croatia
Switzerland
Germany
Bulgaria
Spain
Netherlands
Austria
Northern Ireland

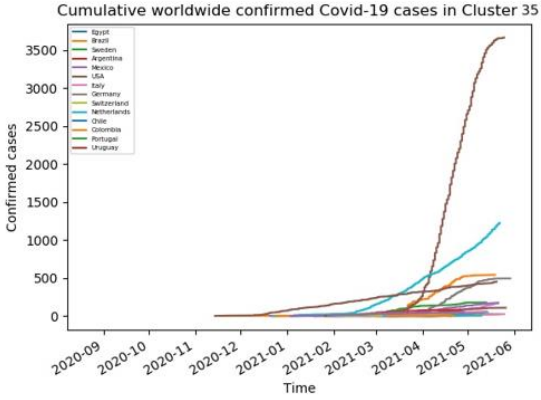

Cumulative worldwide confirmed Covid-19 cases in Cluster 30
Confirmed cases
Time
Egypt
Brazil
Paraguay
Spain
USA
Bolivia
England
Slovenia
France
Poland
Switzerland
Norway
Italy
Germany

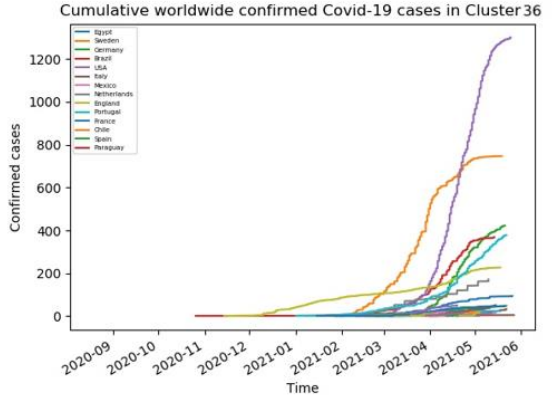

Cumulative worldwide confirmed Covid-19 cases in Cluster 31
Confirmed cases
Time
Egypt
Argentina
Brazil
Jordan
Germany
England
France
Scotland
Sweden
Switzerland
Norway
USA
Belgium
Canada

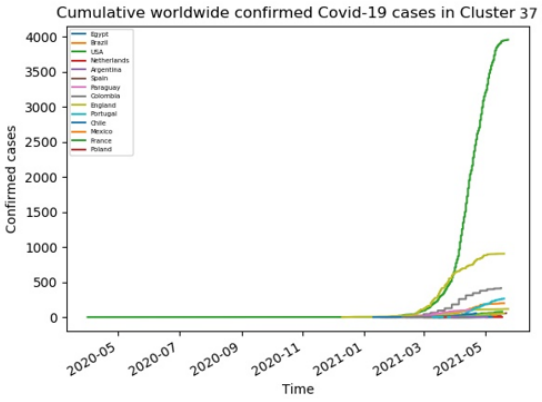

Cumulative worldwide confirmed Covid-19 cases in Cluster 32
Confirmed cases
Time
Egypt
Brazil
Colombia
Argentina
USA
Spain
Portugal
French Guiana
Suriname
Sweden
Chile
England
France
Wales

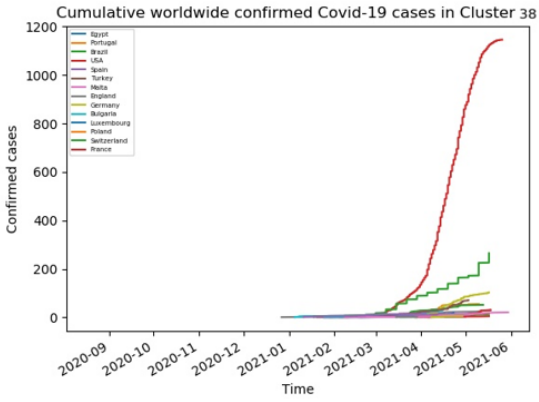

Cumulative worldwide confirmed Covid-19 cases in Cluster 33
Confirmed cases
Time
Egypt
Brazil
Japan
USA
Chile
Portugal
Sweden
Switzerland
England
Germany
Poland
Luxembourg
Denmark
Spain

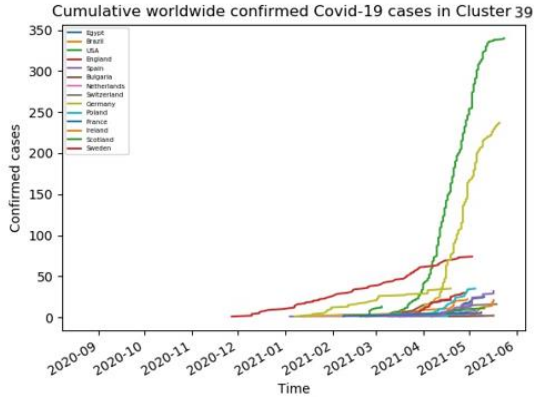

Cumulative worldwide confirmed Covid-19 cases in Cluster 34
Confirmed cases
Time
Egypt
France
Switzerland
Brazil
Italy
Chile
USA
Mexico
Uruguay
Spain
Sweden
Netherlands
England
French Guiana

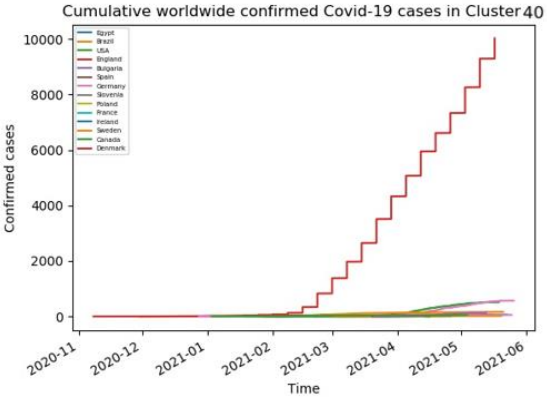

Cumulative worldwide confirmed Covid-19 cases in Cluster 35
Confirmed cases
Time
Egypt
Brazil
Sweden
Argentina
Mexico
USA
Italy
Germany
Switzerland
Netherlands
Chile
Colombia
Portugal
Uruguay

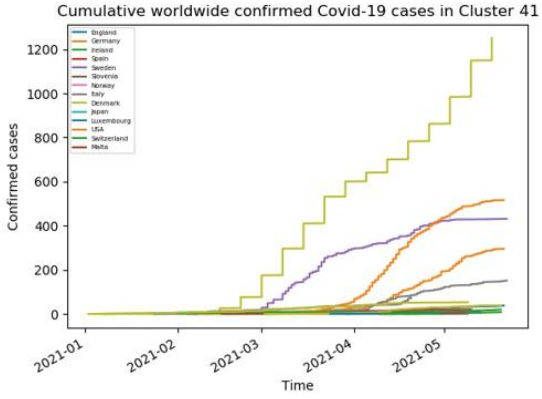

Cumulative worldwide confirmed Covid-19 cases in Cluster 36
Confirmed cases
Time
Egypt
Sweden
Germany
Brazil
USA
Italy
Mexico
Netherlands
England
Portugal
France
Chile
Spain
Paraguay

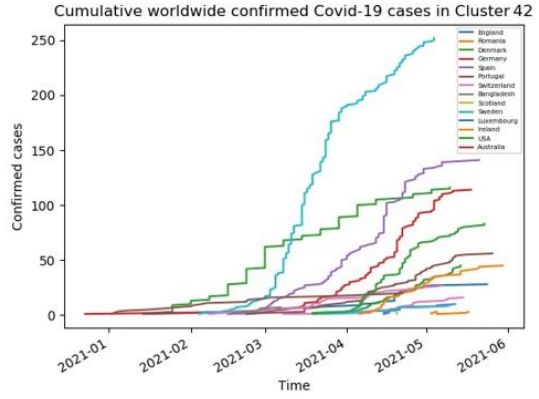

Cumulative worldwide confirmed Covid-19 cases in Cluster 37
Confirmed cases
Time
Egypt
Brazil
USA
Netherlands
Argentina
Spain
Paraguay
Colombia
England
Portugal
Chile
Mexico
France
Poland

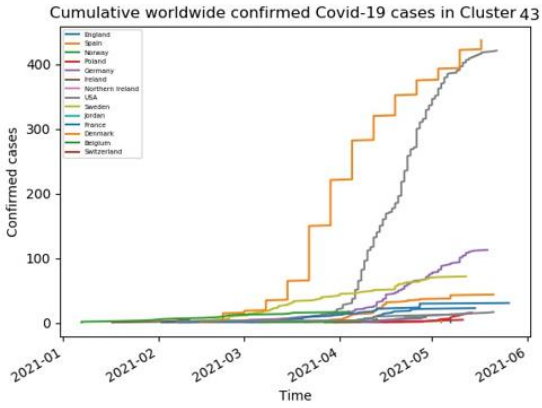

Cumulative worldwide confirmed Covid-19 cases in Cluster 38
Confirmed cases
Time
Egypt
Portugal
Brazil
USA
Spain
Turkey
Malta
England
Germany
Bulgaria
Luxembourg
Poland
Switzerland
France

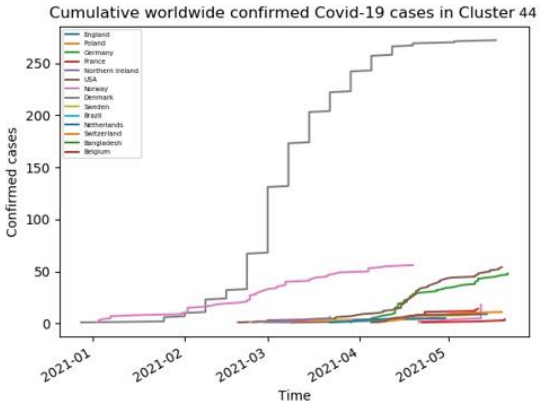

Cumulative worldwide confirmed Covid-19 cases in Cluster 39
Confirmed cases
Time
Egypt
Brazil
USA
England
Spain
Bulgaria
Netherlands
Switzerland
Germany
Poland
France
Ireland
Scotland
Sweden

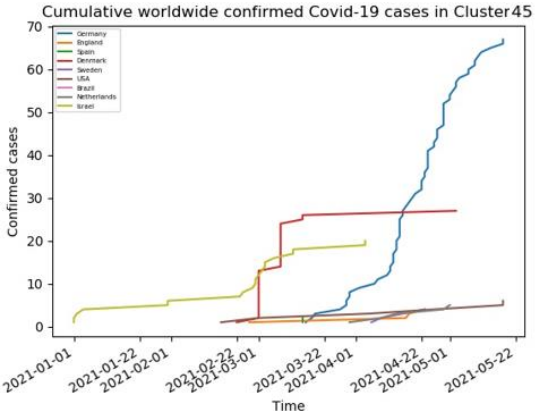

Cumulative worldwide confirmed Covid-19 cases in Cluster 40
Confirmed cases
Time
Egypt
Brazil
USA
England
Bulgaria
Spain
Germany
Slovenia
Poland
France
Ireland
Sweden
Canada
Denmark

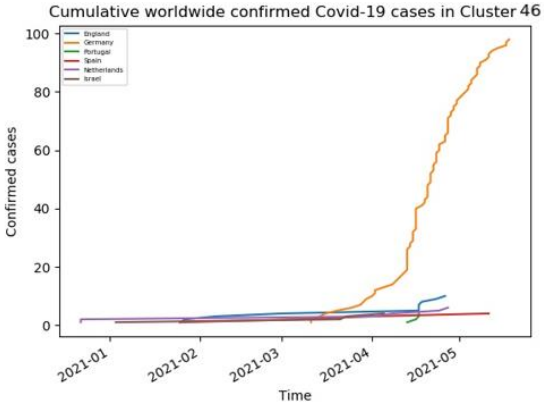

Cumulative worldwide confirmed Covid-19 cases in Cluster 41
Confirmed cases
Time
England
Germany
Ireland
Spain
Sweden
Slovenia
Norway
Italy
Denmark
Japan
Luxembourg
USA
Switzerland
Malta

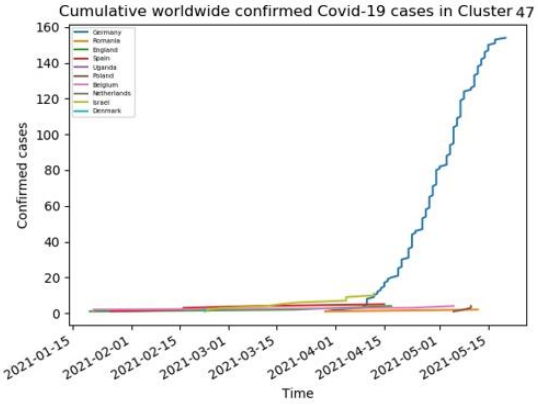

Cumulative worldwide confirmed Covid-19 cases in Cluster 42
Confirmed cases
Time
England
Romania
Denmark
Germany
Spain
Portugal
Switzerland
Bangladesh
Scotland
Sweden
Luxembourg
Ireland
USA
Australia

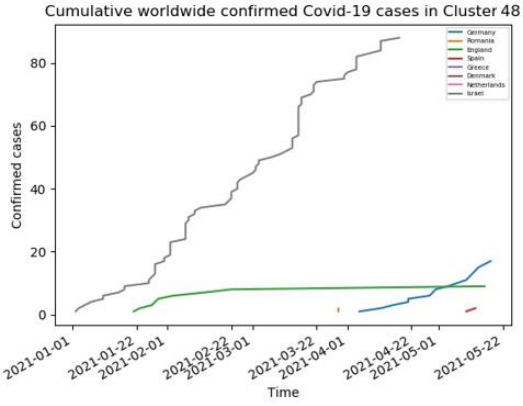

Cumulative worldwide confirmed Covid-19 cases in Cluster 43
Confirmed cases
Time
England
Spain
Norway
Poland
Germany
Ireland
Northern Ireland
USA
Sweden
Jordan
France
Denmark
Belgium
Switzerland

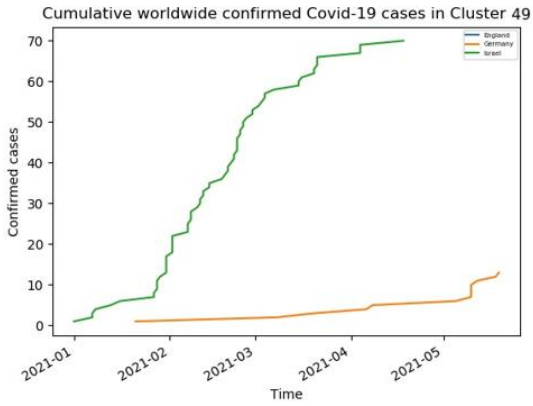

Cumulative worldwide confirmed Covid-19 cases in Cluster 44
Confirmed cases
Time
England
Poland
Germany
France
Northern Ireland
USA
Norway
Denmark
Sweden
Brazil
Netherlands
Switzerland
Bangladesh
Belgium

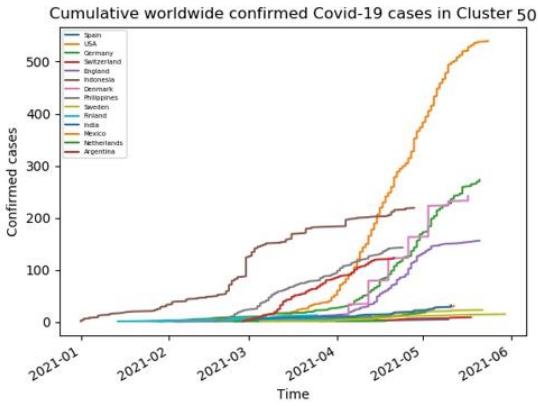

Cumulative worldwide confirmed Covid-19 cases in Cluster 45
Confirmed cases
Time
Germany
England
Spain
Denmark
Sweden
USA
Brazil
Netherlands
Israel

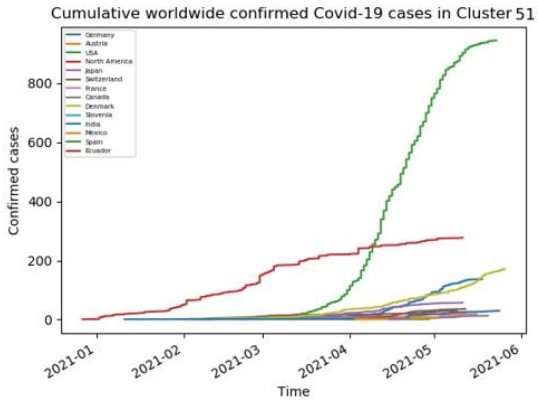

Cumulative worldwide confirmed Covid-19 cases in Cluster 46
Confirmed cases
Time
England
Germany
Portugal
Spain
Netherlands
Israel

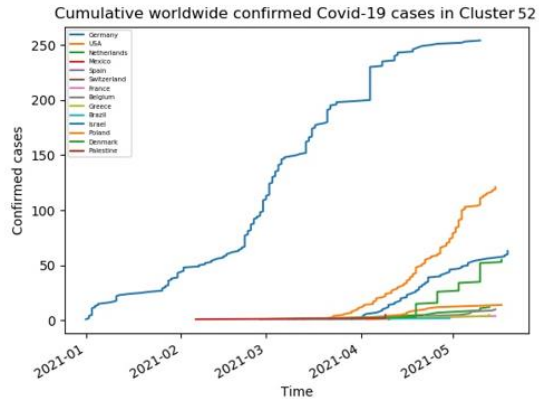

Cumulative worldwide confirmed Covid-19 cases in Cluster 47
Confirmed cases
Time
Germany
Romania
England
Spain
Uganda
Poland
Belgium
Netherlands
Israel
Denmark

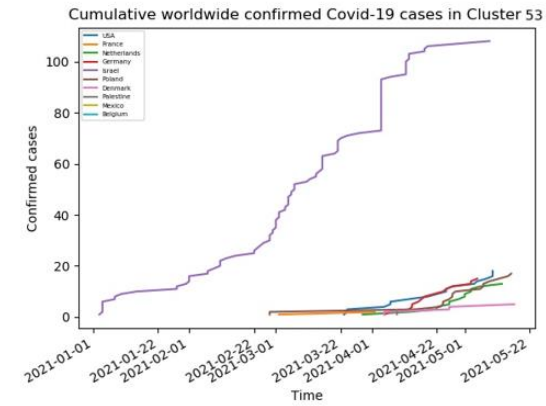

Cumulative worldwide confirmed Covid-19 cases in Cluster 48
Confirmed cases
Time
Germany
Romania
England
Spain
Greece
Denmark
Netherlands
Israel

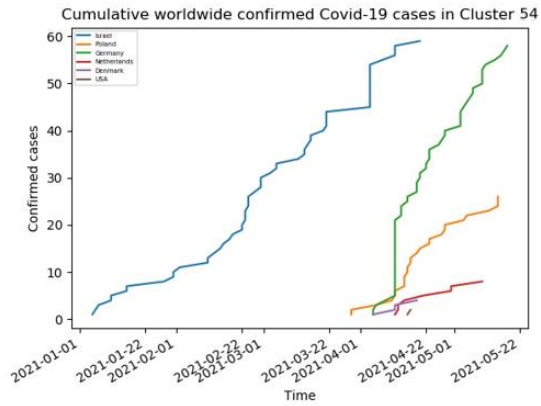

Cumulative worldwide confirmed Covid-19 cases in Cluster 49
Confirmed cases
Time
England
Germany
Israel

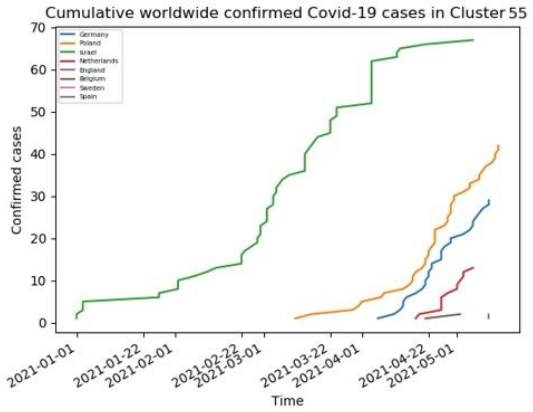

Cumulative worldwide confirmed Covid-19 cases in Cluster 50
Confirmed cases
Time
Spain
USA
Germany
Switzerland
England
Indonesia
Denmark
Philippines
Sweden
Finland
India
Mexico
Netherlands
Argentina

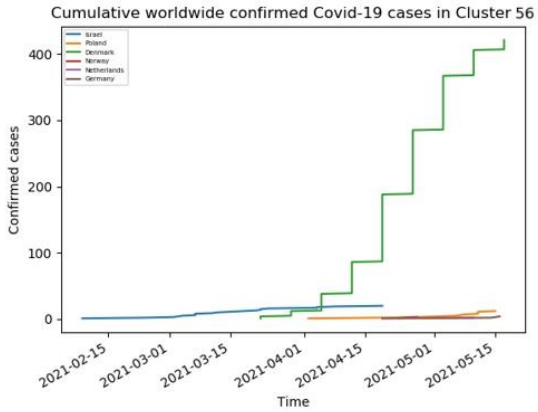

Cumulative worldwide confirmed Covid-19 cases in Cluster 51
Confirmed cases
Time
Germany
Austria
USA
North America
Japan
Switzerland
France
Canada
Denmark
Slovenia
India
Mexico
Spain
Ecuador

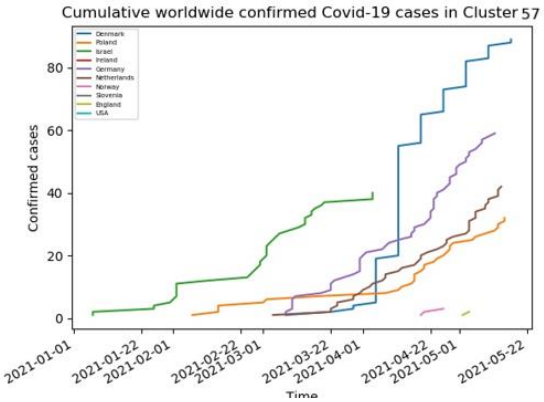

Cumulative worldwide confirmed Covid-19 cases in Cluster 52
Confirmed cases
Time
Germany
USA
Netherlands
Mexico
Spain
Switzerland
France
Belgium
Greece
Brazil
Israel
Poland
Denmark
Palestine

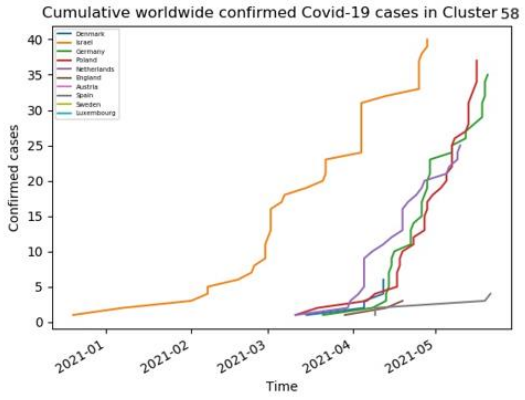

Cumulative worldwide confirmed Covid-19 cases in Cluster 53
Confirmed cases
Time
USA
France
Netherlands
Germany
Israel
Poland
Denmark
Palestine
Mexico
Belgium

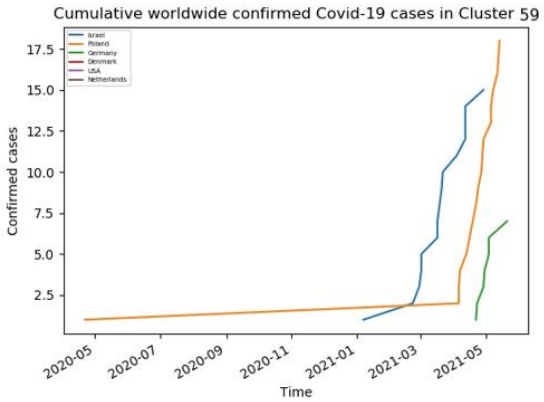

Cumulative worldwide confirmed Covid-19 cases in Cluster 54
Confirmed cases
Time
Israel
Poland
Germany
Netherlands
Denmark
USA

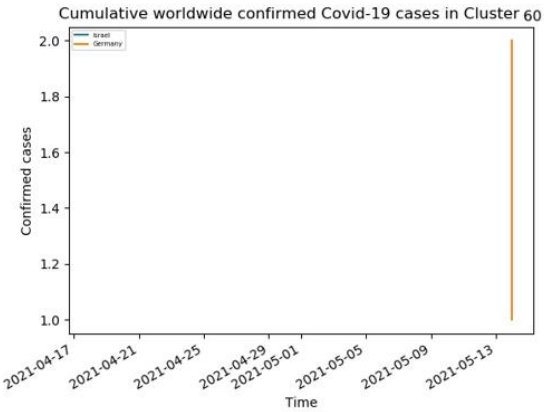

Cumulative worldwide confirmed Covid-19 cases in Cluster 55
Germany
Poland
Israel
Netherlands
England
Belgium
Sweden
Spain
Confirmed cases
Time

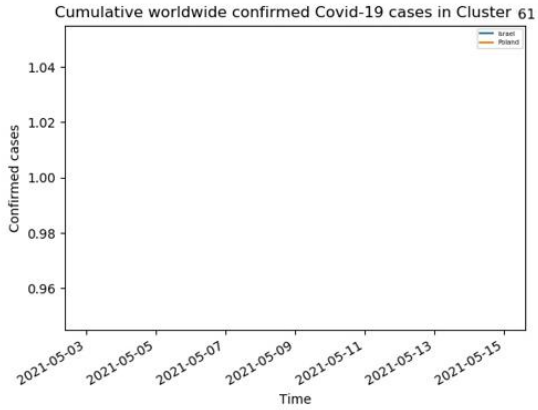

Cumulative worldwide confirmed Covid-19 cases in Cluster 56
Israel
Poland
Denmark
Norway
Netherlands
Germany
Confirmed cases
Time

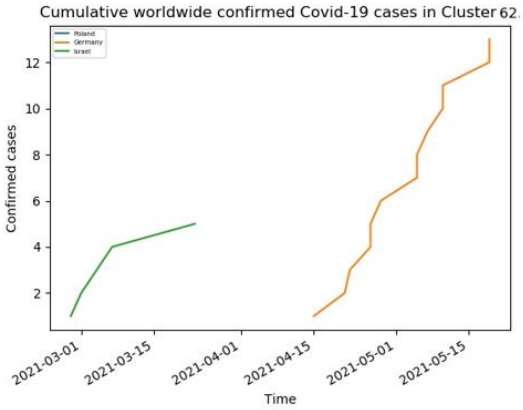

Cumulative worldwide confirmed Covid-19 cases in Cluster 57
Denmark
Poland
Israel
Ireland
Germany
Netherlands
Norway
Slovenia
England
USA
Confirmed cases
Time

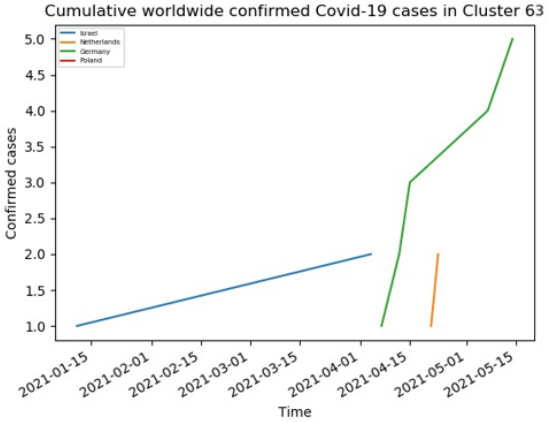

Cumulative worldwide confirmed Covid-19 cases in Cluster 58
Denmark
Israel
Germany
Poland
Netherlands
England
Austria
Spain
Sweden
Luxembourg
Confirmed cases
Time


Cumulative worldwide confirmed Covid-19 cases in Cluster 59
Israel
Poland
Germany
Denmark
USA
Netherlands
Confirmed cases
Time

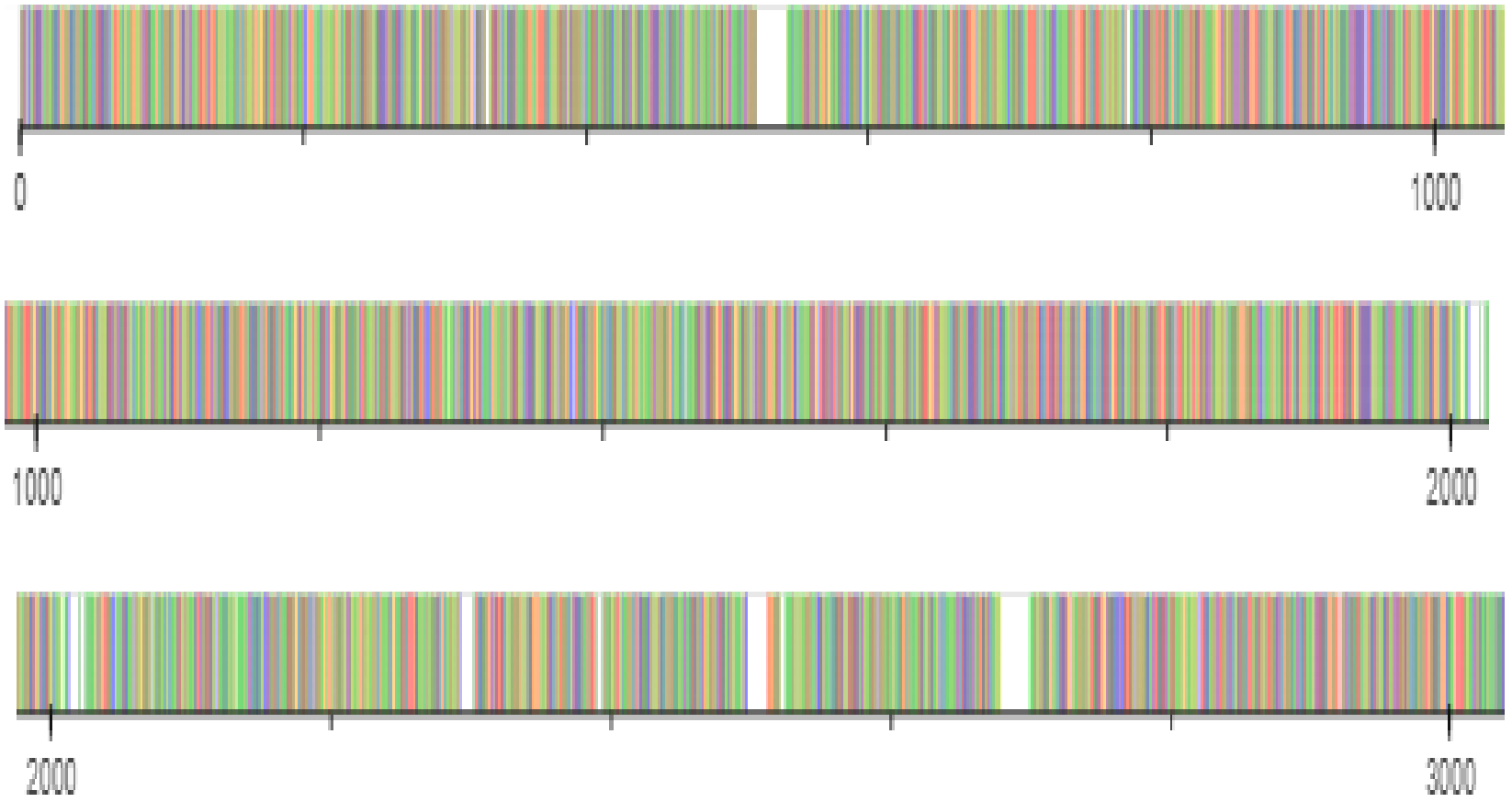

Cumulative worldwide confirmed Covid-19 cases in Cluster 60
Israel
Germany
Confirmed cases
Time

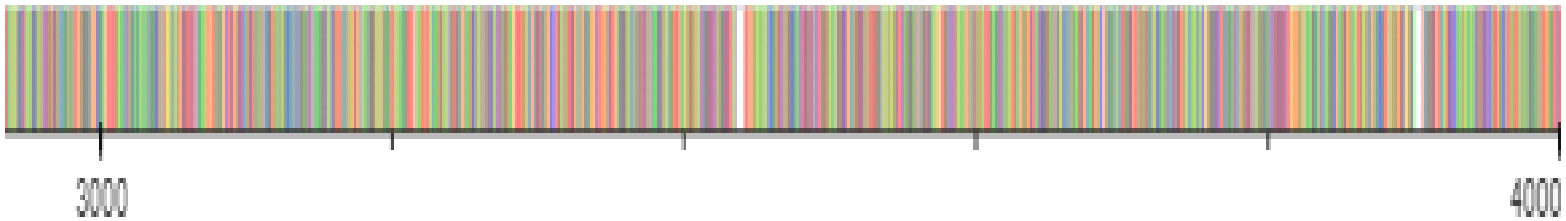

Cumulative worldwide confirmed Covid-19 cases in Cluster 61
Israel
Poland
Confirmed cases
1.04
1.02
1.00
0.98
0.96
2021-05-03
2021-05-05
2021-05-07
2021-05-09
2021-05-11
2021-05-13
2021-05-15
Time


Cumulative worldwide confirmed Covid-19 cases in Cluster 62
Poland
Germany
Israel
Confirmed cases
12
10
8
6
4
2
2021-03-01
2021-03-15
2021-04-01
2021-04-15
2021-05-01
2021-05-15
Time


Cumulative worldwide confirmed Covid-19 cases in Cluster 63
Israel
Netherlands
Germany
Poland
Confirmed cases
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
2021-01-15
2021-02-01
2021-02-15
2021-03-01
2021-03-15
2021-04-01
2021-04-15
2021-05-01
2021-05-15
Time

Appendix D: Sequence Alignment Viewer


0
1000
1000
2000
2000
3000


3000
4000

Appendix E: Summary of quality control checks for sequences from British Columbia

Table 0-1. Summary of quality control checks for sequences from British Columbia sequenced by Nanopore.

Sample Name| accession number	% Dups	% GC	Read Length	M Seqs
Nanopore_report | SRR14621182	64.2%	49%	515 bp	0.6
Nanopore_report | SRR14621183	64.4%	47%	619 bp	0.7
Nanopore_report | SRR14621184	75.6%	52%	468 bp	1.1
Nanopore_report | SRR14621185	78.4%	52%	436 bp	1.0
Nanopore_report | SRR14621186	62.0%	49%	634 bp	0.8
Nanopore_report | SRR14621187	76.1%	49%	576 bp	0.8
Nanopore_report | SRR14621188	72.8%	49%	531 bp	0.9
Nanopore_report | SRR14621189	68.5%	51%	575 bp	1.4
Nanopore_report | SRR14621190	72.3%	50%	593 bp	1.3
Nanopore_report | SRR14621192	71.7%	49%	621 bp	1.2
Nanopore_report | SRR14621193	72.7%	56%	408 bp	0.7
Nanopore_report | SRR14621194	65.4%	50%	618 bp	1.0
Nanopore_report | SRR14621195	72.9%	48%	465 bp	0.9
Nanopore_report | SRR14621196	68.4%	50%	565 bp	0.9
Nanopore_report | SRR14621197	72.3%	48%	597 bp	0.8
Nanopore_report | SRR14621198	76.9%	58%	395 bp	0.6
Nanopore_report | SRR14621199	72.5%	57%	427 bp	0.9
Nanopore_report | SRR14621200	71.3%	50%	634 bp	1.1
Nanopore_report | SRR14621201	72.4%	56%	411 bp	0.7
Nanopore_report | SRR14621203	67.1%	51%	518 bp	0.8
Nanopore_report | SRR14621204	67.2%	52%	463 bp	0.8
Nanopore_report | SRR14621205	69.5%	50%	499 bp	0.8
Nanopore_report | SRR14621206	73.4%	49%	533 bp	1.2
Nanopore_report | SRR14621207	65.0%	51%	461 bp	0.9

Nanopore_report | SRR14621208	43.9%	47%	463 bp	0.1
Nanopore_report | SRR14621209	56.7%	46%	447 bp	0.1
Nanopore_report | SRR14621210	58.3%	49%	435 bp	0.3
Nanopore_report | SRR14621211	60.4%	46%	443 bp	0.3
Nanopore_report | SRR14621212	62.9%	45%	453 bp	0.3
Nanopore_report | SRR14621214	68.4%	45%	488 bp	0.6
Nanopore_report | SRR14621215	67.4%	45%	489 bp	0.8
Nanopore_report | SRR14621216	59.5%	46%	466 bp	0.2
Nanopore_report | SRR14621217	67.3%	44%	448 bp	0.2
Nanopore_report | SRR14621218	71.4%	45%	496 bp	0.8
Nanopore_report | SRR14621219	58.2%	46%	516 bp	0.2
Nanopore_report | SRR14621220	66.4%	44%	465 bp	0.1
Nanopore_report | SRR14621221	65.7%	47%	477 bp	0.3
Nanopore_report | SRR14621222	69.6%	43%	463 bp	0.5
Nanopore_report | SRR14621223	66.8%	42%	470 bp	0.4
Nanopore_report | SRR14621225	58.9%	47%	481 bp	0.2
Nanopore_report | SRR14621226	70.0%	47%	464 bp	0.2
Nanopore_report | SRR14621227	64.8%	49%	467 bp	0.4
Nanopore_report | SRR14621228	71.6%	49%	472 bp	0.7
Nanopore_report | SRR14621229	69.9%	49%	478 bp	0.3
Nanopore_report | SRR14621230	70.8%	49%	503 bp	0.7
Nanopore_report | SRR14621231	68.5%	47%	474 bp	0.7
Nanopore_report | SRR14621232	64.4%	48%	485 bp	0.3
Nanopore_report | SRR14621233	70.0%	48%	464 bp	0.2
Nanopore_report | SRR14621234	72.8%	46%	505 bp	0.6
Nanopore_report | SRR14621236	70.5%	49%	523 bp	0.9
Nanopore_report | SRR14621237	68.6%	48%	485 bp	0.9

Nanopore_report | SRR14621238	61.7%	50%	490 bp	0.4
Nanopore_report | SRR14621239	70.1%	50%	508 bp	0.2
Nanopore_report | SRR14621240	71.1%	47%	486 bp	0.8
Nanopore_report | SRR14621241	73.0%	56%	443 bp	0.6
Nanopore_report | SRR14621242	64.7%	56%	438 bp	0.4
Nanopore_report | SRR14621243	71.7%	50%	515 bp	0.9
Nanopore_report | SRR14621244	63.7%	50%	501 bp	0.7
Nanopore_report | SRR14621245	63.1%	51%	448 bp	0.5
Nanopore_report | SRR14621247	62.2%	48%	497 bp	0.5
Nanopore_report | SRR14621248	65.5%	47%	508 bp	0.7
Nanopore_report | SRR14621249	68.8%	57%	411 bp	0.5
Nanopore_report | SRR14621250	64.9%	55%	438 bp	0.4
Nanopore_report | SRR14621251	74.5%	49%	508 bp	0.4
Nanopore_report | SRR14621252	62.7%	52%	455 bp	0.4
Nanopore_report | SRR14621253	63.7%	51%	480 bp	0.3
Nanopore_report | SRR14621254	66.8%	49%	476 bp	0.3
Nanopore_report | SRR14621255	71.7%	49%	435 bp	0.4
Nanopore_report | SRR14621256	70.6%	56%	440 bp	0.4
Nanopore_report | SRR14621258	67.9%	50%	414 bp	0.5
Nanopore_report | SRR14621259	69.2%	49%	457 bp	0.3
Nanopore_report | SRR14621260	55.7%	50%	447 bp	0.1
Nanopore_report | SRR14621261	66.0%	49%	441 bp	0.1
Nanopore_report | SRR14621262	71.2%	48%	452 bp	0.4
Nanopore_report | SRR14621263	67.8%	50%	494 bp	0.8
Nanopore_report | SRR14621264	64.0%	50%	478 bp	0.6
Nanopore_report | SRR14621265	65.6%	50%	456 bp	0.1
Nanopore_report | SRR14621266	69.7%	48%	497 bp	0.7

Nanopore_report | SRR14621267	58.7%	51%	471 bp	0.3
Nanopore_report | SRR14636801	62.7%	56%	487 bp	0.1
Nanopore_report | SRR14636802	65.3%	57%	497 bp	0.2
Nanopore_report | SRR14636803	62.7%	53%	432 bp	0.1
Nanopore_report | SRR14636804	62.8%	51%	442 bp	0.4
Nanopore_report | SRR14636805	69.0%	49%	482 bp	1.1
Nanopore_report | SRR14636806	70.0%	49%	468 bp	0.2
Nanopore_report | SRR14636807	74.1%	50%	520 bp	1.6

Table 0-2. Summary of quality control checks for sequences from British Columbia sequenced by Illumina.

Sample Name| accession number	**% Dups**	**% GC**	**Read Length**	**M Seqs**
Illumina_report | SRR14621181	89.1%	55%	302 bp	1.7
Illumina_report | SRR14621191	88.3%	53%	302 bp	1.5
Illumina_report | SRR14621202	89.5%	60%	302 bp	1.4
Illumina_report | SRR14621213	89.8%	53%	302 bp	1.4
Illumina_report | SRR14621224	85.2%	54%	302 bp	0.6
Illumina_report | SRR14621235	88.9%	54%	302 bp	1.7
Illumina_report | SRR14621246	88.2%	60%	302 bp	1.5
Illumina_report | SRR14621257	84.7%	59%	302 bp	1.4
Illumina_report | SRR14621268	91.1%	53%	302 bp	1.8
Illumina_report | SRR14621269	89.8%	54%	302 bp	1.7
Illumina_report | SRR15465142	91.5%	49%	502 bp	1.2
Illumina_report | SRR15465143	87.3%	49%	502 bp	1.2
Illumina_report | SRR15465144	89.3%	49%	502 bp	1.3
Illumina_report | SRR15465145	91.4%	49%	502 bp	2.2
Illumina_report | SRR15465146	86.9%	49%	502 bp	1.0

Illumina_report | SRR15465147	93.5%	44%	502 bp	1.1
Illumina_report | SRR15465148	93.2%	50%	502 bp	1.1
Illumina_report | SRR15465149	87.9%	48%	502 bp	1.6
Illumina_report | SRR15465150	88.0%	49%	502 bp	1.1
Illumina_report | SRR15465151	89.6%	50%	502 bp	1.5