Phylogenetic analysis of COVID-19 in Burundi

Harambe, Prosper; Supervised by : Dr. David Niyukuri

2024
UB, FS
https://repository.ub.edu.bi/handle/123456789/1936

UNIVERSITY OF BURUNDI
FACULTY OF SCIENCES
DEPARTMENT OF MATHEMATICS
MATHEMATICS AND PHYSICS RESEARCH CENTER

PHYLOGENETIC ANALYSIS OF COVID-19 IN BURUNDI

By :
HARAMBE Prosper

Thesis presented and defended publicly to obtain the Master's Degree in Fundamental and Applied Mathematics

Supervised by:

Dr. David NIYUKURI

Co-supervised by:

Msc. Cassien NDUWIMANA

Bujumbura, December 14, 2024

Members of Jury

Prof Servat NYANDWI : President of the Jury
Dr Menus NKURUNZIZA : Secretary of the jury
Dr David NIYUKURI : Supervisor
Msc Cassien NDUWIMANA : Co-supervisor

Dedication

To Almighty God ;
to my family ;
to all my loved ones;

I dedicate this memoir.

Acknowledgements

I want to thank God for the privileged for giving me the strength, the courage, patience and health to carry out this work. It is not by my own mighty or power that am here but that of God. My supervisor Dr David NIYUKURI and my co-supervisor Cassien NDUWIMANA have been very helpful towards my research work and am very grateful to them.
I thank the members of the jury who agreed to read this dissertation and participate in its evaluation.

Thanks

Résumé

L'analyse phylogénétique a pris de l'ampleur au cours des 10 dernières années dans l'étude de l'évolution des virus à ARN. Pour comprendre l'évolution du virus, nous utilisons des modèles de substitution de nucléotides. Dans le travail actuel, sur la base des modèles évolutifs utilisés dans la littérature qui expliquent l'évolution moléculaire des virus à ARN, nous sélectionnons le meilleur modèle et décrivons les raisons pour lesquelles il est considéré comme le meilleur. Le modèle $GTR + \Gamma$ est sélectionné comme le meilleur modèle qui explique le processus d'évolution des virus à ARN, en particulier le SARS-COV-2, le virus responsable de la COVID-19. Ensuite, en utilisant la jonction de voisins de l'arbre reconstruit, nous reconstruisons l'arbre phylogénétique en utilisant les paramètres du modèle $GTR + \Gamma$ et les données de séquence du SARS-COV-2. La présente étude visait à étudier le rôle de différentes variantes du SARS-COV-2 dans les vagues successives de COVID-19 vécues au Burundi et l'impact de leur évolution sur le cours de cette pandémie. Au total, avec les échantillons téléchargés sur GISAID, nous avons documenté 12 lignées PANGO dont 5 (BA.1, B.1.617.2, BA.1.13, AY.46 et B.1.1) étaient des variantes préoccupantes (VOC) et représentaient 77,84 % de tous les génomes isolés au Burundi de mai à décembre 2021.

Mots clés: COVID-19, variantes du SARS-COV-2, diversité génétique, modèles d'évolution.

Abstract

Phylogenetic analysis over the last 10 years has taken a rise in the study of evolution of the RNA viruses. To understand the evolution of the virus, we use models of nucleotide substitution. In the current work, based on evolutionary models used in the literature that explain molecular evolution of RNA viruses, we select the best model and outline reasons for it being considered the best. The $GTR + \Gamma$ model is selected as the best model that explains the process of evolution in RNA viruses specifically SARS-COV-2, the virus responsible for COVID-19. Then, using Neighbour Joining of tree reconstruction, we reconstruct the phylogenetic tree using the $GTR + \Gamma$ model parameters and the SARS-COV-2 sequence data. The present study aimed to investigate the role of different SARS-COV-2 variants in the successive COVID-19 waves experienced in Burundi and the impact of their evolution on the course of that pandemic. In total, with the samples uploaded to GISAID we documented 12 PANGO lineages of which 5 (BA.1, B.1.617.2, BA.1.13, AY.46 and B.1.1) were variants of concern (VOCs) and accounted for 77.84% of all the genomes isolated in Burundi from May to December 2021.

Keywords:COVID-19, SARS-COV-2 variants, Genetic diversity, Models of evolution.

Contents

Members of Jury i

Dedication ii

Acknowledgements iii

Résumé iv

Abstract v

Contents vi

List of figures ix

List of Tables x

Abbreviations xi

Foreword xiii

1 Introduction 1

2 Generalities 3

2.1 History and epidemiology of COVID-19 3

2.2 Organization of the SARS-CoV-2 genome 4

2.3 Replication cycle 6

2.4 Diagnosis and Data generation 8

2.4.1 Real-time PCR or qPCR (quantitative PCR) 8

2.4.2 Serological tests 8

2.5 Genomics 9

2.5.1 Definition 9

2.5.2 Historical . . . 9
2.5.3 Sequencing . . . 11
2.5.4 First generation sequencers . . . 11
2.5.5 Second generation sequencers . . . 14
2.5.6 The 3rd generation sequencers . . . 16
2.6 SARS-CoV-2 variants . . . 16
2.6.1 Variants of concern(VOC) . . . 16
2.6.2 Variant under investigation or variant of interest . . . 19
2.6.3 Variants currently being evaluated . . . 19
2.6.4 High Consequence Variants . . . 20

3 Constructing evolutionary models 21
3.1 Introduction . . . 21
3.2 Stochastic process of evolution . . . 21
3.3 Extending to evolutionary models . . . 23
3.4 Models of evolution . . . 25
3.4.1 Likelihood ratio test . . . 26
3.4.2 Akaike information criterion (AIC) . . . 26
3.4.3 The inputs of the models . . . 27
3.4.4 Models and rate of variation among sites . . . 30
3.4.5 Selecting the best model . . . 30
3.4.6 Best fit model . . . 31
3.5 $GTR + \Gamma$ model . . . 33
3.5.1 GTR + Γ model model assumptions . . . 33
3.5.2 The input of $GTR + \Gamma$ model . . . 34
3.5.3 Limitation of the $GTR + \Gamma$ model . . . 35

4 COVID-19 Data in Burundi 37
4.1 Study sites and population . . . 37
4.2 Type of study . . . 37
4.3 Study periode . . . 37
4.4 Sampling . . . 37
4.5 Bioinformatics analysis . . . 38
4.6 Statistical data analysis . . . 38

5 COVID-19 sequence data analysis 39

5.1 Overall results 39

5.1.1 The genetic diversity of the SARS-COV-2 lineages 40

5.1.2 Evolution of SARS-CoV-2 in Burundi 42

5.1.3 Notable mutations detected 42

5.1.4 Phylogenetic analysis 43

6 Discussion 45

Conclusion 47

Recommandations 48

Bibliography 49

List of figures

2.1 *Schematic presentation of the organization of the SARS-CoV-2 genome, and virion structure.* 5

2.2 *Biological cycle of SARS-CoV-2* 7

5.1 Variants of concern detected in BURUNDI during the periode from May to December 2021 40

5.2 *Distribution of the SARS-COV-2 lineage composition of different COVID-19 waves. B.1.1.7 = Alpha variant, B.1.351 = Beta variant, B.1.617.2 = Delta variant, B.1.1.529 = Omicron variant, BA.1,BA.1.13,BA.1.13.1 and BA.1.1 are Omicron sublineages, AY.46 is Delta sublineage* 41

5.3 Distribution of mutations carried by variants identified in Burundi on the SARS-COV-2 genome 43

5.4 *The Neighbour-Joining (NJ) phylogenetic tree of the nucleotide sequences of SARS-COV-2 genomes detected in Burundi in 2021. The B.1.351 genomes (n = 5) are indicated in green, Delta (B.1.617.2) and its sublineages (n = 66) in yellow, the Omicron (B.1.1.529) and its sublineages BA.1 and BA.1.1 genomes (n = 86) in red, and the only one B.1.1.7 genome is in blue.* 44

List of tables

3.1 *Models considered during model test* 27
3.2 Results of Model test using jmodeltest 33

5.1 *Distribution of samples according to sex* 39
5.2 *Distribution of samples according to age* 40
5.3 *Polymorphisms in the SARS-CoV-2 genome* 42

Abbreviations

A	Adenine
ACE2	Angiotensin 2 Converting Receptor
ATP	Adenosine Triphosphate
C	Cytosine
CCD	Charge-Coupled Device
CDC	Centers for Disease Control and Prevention
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleotide Triphosphate
dATP	Deoxyadenosine Triphosphate
dTTP	Deoxythymidine Triphosphate
dCTP	Deoxycytidine Triphosphate
dGTP	Deoxyguanosine Triphosphate
ddATP	Dideoxyadenosine Triphosphate
ddTTP	Dideoxythymidine Triphosphate
ddCTP	Dideoxycytidine Triphosphate
ddGTP	Dideoxyguanosine Triphosphate
ECDC	European Center of Disease and Control
ERGIC	Endoplasmic Reticulum and Golgi Intermediate Compartement
FC	Flow Cell
G	Guanine
GISAID	Global Initiative for Sharing All Influenza Data
GTR	General Time Reversble
HCoV-229E	Human Coronavirus-229E
HCoV-OC43	Human Coronavirus Organ Culture 43
HCoV-NL63	Human Coronavirus Netherlands 63
HCoV-HKU1	Human Coronavirus Hong Kong University 1
MERS-CoV	Middle East Respiratory Syndrome Coronavirus
MCM	Medical Countermeasure
NAAT	Nucleic Acid Amplification Tests
Nsp	Non-Structural Protein
NTD	N Terminal Domain
WHO	World Health Organization
ORF	Open Reading Frame
Pb	Base Pair

PCR	Polymerase Chain Reaction
GDP	Gross Domestic Product
PP1a	Polyprotein 1a
PP1ab	Polyprotein 1ab
RBD	Receiver Binding Domain
RNA	Ribonucleic Acid
RDT	Rapid Diagnostic Tests
SBS	Sequencing by Synthesis
SNP	Single Nucleotide Polymorphism
SOLID	Sequencing by Oligonucleotide Ligation and Detection
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
T	Thymine
TMPRSS	Transmembrane Serine-Type Protease
VOI	Variant Of Interest
VOC	Variant Of Concern

Foreword

The emergence and rapid spread of new severe acute respiratory syndrome coronavirus 2 (SARS- COV-2) variants have challenged the control of the COVID-19 pandemic globally. Burundi was not spared by that pandemic, but the genetic diversity, evolution, and epidemiology of those variants in the country remained poorly understood. The present study sought to investigate the role of different SARS-COV-2 variants in the successive COVID-19 waves experienced in Burundi and the impact of their evolution on the course of that pandemic. We conducted a cross-sectional descriptive study using positive SARS-COV-2 samples for genomic sequencing. Subsequently, we performed statistical and bioinformatics analyses of the genome sequences in light of available metadata.

Chapter 1

Introduction

Coronavirus disease 2019 or COVID-19 is an illness of the upper and lower respiratory tract caused by a virus of the *Betacoronavirus* family [1].Currently, seven (7) species are implicated in human infections, four (4) of which cause mild respiratory infections, namely human coronavirus-229E (HCoV-229 E), human coronavirus Netherland 63 (HCoV-NL63), human coronavirus organ culture 43 (HCoV-OC43), human coronavirus Hong Kong (HCoV-HKU1). The other three (3) are involved in severe forms including the first severe acute respiratory syndrome coronavirus (SARS-CoV-1), the Middle East respiratory syndrome coronavirus (MERS-CoV) and the second respiratory syndrome coronavirus severe acute (SARS CoV-2). Coronaviruses are divided into four genera: *Alphacoronaviruses*, *Betacoronaviruses*, *Gammacoronaviruses* and *Deltacoronaviruses*. The genera *Alphacoronavirus* and *Betacoronavirus* infect mammals and the birds. The *Gammacoronavirus* and *Deltacoronavirus* genera infect birds.

Coronaviruses were first discovered in poultry in 1930. A few years later, the virus crossed the human barrier in the 1960s with the HCoV-229E species first and then other species [2]. SARS-CoV-1, MERS-CoV and SARS-CoV-2 were the three species that strongly attracted global attention for their pathogenicities. Severe acute respiratory syndrome coronavirus-1 (SARS-CoV-1) was first discovered in Guangdong, China in February 2003 with civet cats as the intermediate host. The Middle East respiratory syndrome coronavirus (MERS-CoV) was discovered in Jeddah, Saudi Arabia in 2012 with camels as an intermediate host. The last virus responsible for acute respiratory syndrome is SARS-CoV-2. It was first named nCoV-2019 before the current name SARS-CoV-2 on February 11, 2020 by the international committee on taxonomy of viruses and the disease COVID-19 by the(World Health Organisation(WHO). This name is due to its molecular and structural similarities with the Severe Acute Respiratory Syndrome virus that emerged in China in 2003 [3].

As of 17 December 2023, over 772 million confirmed cases and nearly seven million deaths have been reported globally. In Africa, the first case was reported on February 14, 2020 in Egypt, and by January 2023 the African continent had recorded over 9 million COVID-19 cases and 175,140 deaths. This pandemic did not spare Burundi. The first two cases of COVID-19 were detected in Burundi on March 31, 2020, prompting the public health authorities to put in place control measures. By December 2022, a total of 52,162 COVID-

19 confirmed cases and 15 associated deaths had been reported in Burundi since the first report of two imported COVID-19 cases [4].

This pandemic has strongly affected the global economy. The following measures adopted to curb the spread of the virus have led to a severe economic recession. It is the closure of borders, curfews, confinements, the closure of leisure places, and that of schools, etc...[5].

This pandemic has strongly affected Burundi: heavy drop in economic growth and increase in poverty, indebtedness of future generations and delay of progress for sustainable development, job losses and deterioration of human capital [6]. Current strategies to combat the pandemic are based on means of prevention such as barrier measures, travel restrictions, the use of masks and vaccines. In the absence of effective antiviral treatment, hope rests on vaccines. However, the evolution of the virus with the appearance of new variants constitutes a major challenge in the fight against the coronavirus pandemic.

Different variants have been the cause of waves of the pandemic around the world. Depending on the clinical importance of the mutations they carry, the WHO classifies variants into three categories: variants of concern, variants of interest and variants under observation. There is a fourth category that has never been detected before, which is the high-consequence variant. The variants of concern that have been described to date are the Alpha variant which originated in Great Britain, the Beta variant first detected in South Africa, the Gamma variant from Brazil, the Delta variant which comes from India, and recently the Omicron variant (Botswana/South Africa) [7]. These variants can have negative impacts on transmissibility, the severity of infection and on strategies to combat the pandemic. Thus, it is necessary to establish a surveillance system in each country as recommended by the WHO in order to detect new variants which may be obstacles to controlling the pandemic.

In Burundi, the epidemiological surveillance system for the pandemic is mainly based on the detection of cases by antigenic tests and qualitative RT-PCR. However, monitoring the evolution of variants requires genomics tools consisting of a sequencer of the virus's genetic material followed by sequence analysis. In Burundi, data on circulating variants is very limited. The first genomes from Burundi were sequenced in Uganda and Senegal. Given the difficulties in air transport of biological materials and the need to obtain results quickly, it was necessary to develop local sequencing capacity in order to monitor the evolution of the virus.

Chapter 2

Generalities

2.1 History and epidemiology of COVID-19

Coronaviruses are a large family of positive-sense single-stranded RNA viruses with a spherical envelope, a diameter of 100-160 nm . They belong to the family *Coronaviridae*, subfamily *Coronavirinae*, order *Nidovirales* divided into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus* [2].

Known to date, coronaviruses infected birds with a first case in 1930 in poultry with the avian infectious bronchitis virus (*Gammacoronavirus*) in chickens. In 1968, the term coronavirus was coined by their appearance on electron microscopy showing the appearance of a crown. All cases of human infections all have the source of animals which are the natural and intermediate hosts. The majority of infections due to coronaviruses infecting animals result in gastrointestinal infections including the porcine epidemic diarrhea virus (PEDV) which wreaked havoc on pig farms in America and Asia in 2013.

A few years later, the virus crossed the human barrier with the first species HCoV-229E in 1960 in Chicago. Subsequently, HCoV-0C43 in 1967 in the USA, HCoV-NL63 in 2004 in the Netherlands, HCoV-HKU1 in 2005 in Hong Kong were described in human infections. These four viruses cause mild respiratory infections. The three species responsible for serious human infections are: SARS-CoV-1 (China from 2002-2003); MERS-CoV (Middle East 2012); and SARS-CoV-2 (2019 Wuhan, China). In terms of classification, HCoV-229E and HCoV-OC43 are the classic coronaviruses. HCoV-NL63 and HCoV-HKU1 are class 2A coronaviruses SARS-CoV-1, MERS-CoV and SARS-CoV-2 are classified as class 2B coronaviruses [2].

Some important dates have marked the history of SARS-CoV-2. It all started in China on December 31, 2019 when the WHO office in China was informed of an outbreak of pneumonia in the city of Wuhan, Hubei province. Thanks to advances in sequencing technology and previous knowledge about coronaviruses, within a week the pathogen was identified. A new coronavirus, more precisely a close Betacoronavirus of the coronavirus from the severe acute respiratory syndrome epidemic that appeared in China in 2003 is the pathogen of this pneumonia. Thailand was the first country to report a case after

China on January 13, 2020. The number of cases and deaths caused by the epidemic exceeded in one month those of the SARS-CoV-1 epidemic with 910 deaths and 4,000 cases in China alone. On February 11, 2020, the International Committee on Taxonomy of Viruses named the novel coronavirus SARS-CoV-2, based on its genetic similarity to SARS-CoV-1. The disease caused by this virus was named COVID-19 by the WHO. This nomination is due to its structural and molecular resemblance to SARS CoV-1. On the molecular level, SARS-CoV-2 has a genetic sequence similarity of approximately 80% with SARS-CoV-1. At the structural level, they share four (4) structural proteins with some differences specific to SARS-CoV-2 [1].

In mid-February, the epidemic quickly reached other countries such as Egypt and France. As of the end of February 2020, eleven (11) other European countries have reported cases with 82,000 confirmed infections and 2,800 deaths worldwide. On March 11, 2020, the WHO officially declared the outbreak a pandemic and governments around the world began implementing strategies to slow the spread of infection. The epidemic was quickly brought under control in China by their rigor and respect for virus propagation measures as no cases were reported on March 16, 2020. It continues to evolve in other parts of the world such as Italy which quickly became the new epicenter emerging with a peak in new daily cases reported at 6,557 on March 21, 2020. The United States, with at least 100,000 cumulative cases on March 27 and more than 2,700 deaths. On March 29, Spain recorded 838 new deaths in 24 hours. The first cases of COVID-19 in Africa were reported in February 2020. BURUNDI reported its first case on March 31, 2020 [8].

Containment measures, closures of leisure places, borders and the cessation of several beneficial activities were adopted all over the world with catastrophic repercussions on the global economy. The global number of cases was over 600,000, including over 29,000 deaths as of March 29, 2020. At the end of March, few countries remained with unreported cases. A global drop in cases was reported at the end of April 2020 thanks to response measures against the pandemic. Borders have been opened, lockdowns have been lifted and some activities have resumed but these decisions should be delayed or accompanied by more awareness because the consequences were a second wave. As of July 12, 2020, more than 12.7 million SARS-CoV-2 infections have been confirmed in 213 countries and territories, including more than 560,000 deaths, with the highest proportion of cases and mortality in the United States.

2.2 Organization of the SARS-CoV-2 genome

Coronaviruses have the longest genome among all RNA viruses at 27 to 32 kilobases (kb) [2]. The SARS-CoV-2 genome is divided into three (3) parts: open reading frames 1a (ORF 1a), open reading frames 1b (ORF1b) and structural proteins. During the replication process, ORFs 1a and 1b encode two poly-proteins PP1a and PP1ab which will be used for the synthesis of non-structural proteins (nsps). The genome includes 6 to 11 ORFs with 5' and 3' untranslated regions (UTRs). There are sixteen non-structural

proteins (nsp1-nsp16). Non-structural proteins ensure the proper functioning of replication. Nsp3 encodes papain-like protease, nsp5 encodes chymotrypsin-like protease, nsp12 encodes RNA-dependent RNA polymerase, and nsp13 encodes helicase [9]. Structural proteins are essential for contracting SARS-CoV-2 infection. There are four of these proteins:

- **S spike protein**: in the form of a spike it is present on the entire surface of the virus giving it the appearance of a crown, hence the name coronavirus (Figure 2.1). This protein plays a very important role in the entry of the virus into the target cell. The spike protein has two subunits, the S1 subunit (two subdomains the N terminal domain (NTD) and the C-terminal domain (CTD) allowing it to bind to the entry receptor (ACE-2) in the host cell followed by its priming by membrane proteases (TMPRSS2, Cat B/L). The S2 subunit facilitates the fusion of the virus with cell membranes. Despite the common presence of the S protein in both viruses, SARS -CoV-2 is distinguished from SARS-CoV-1 by the presence of six amino acids in the ACE-2 receptor binding domain giving it a stronger affinity for the receptor than SARS-CoV-1. This protein offers several therapeutic targets because the success of the infection depends on its binding with the receptor.

- **Nucleocapsid proteins (N)**: protect the viral genome from external host cells for the proper functioning of replication.

- **Proteins M and E**: are responsible for the assembly, transmembrane transport, budding and release of virions formed [3].

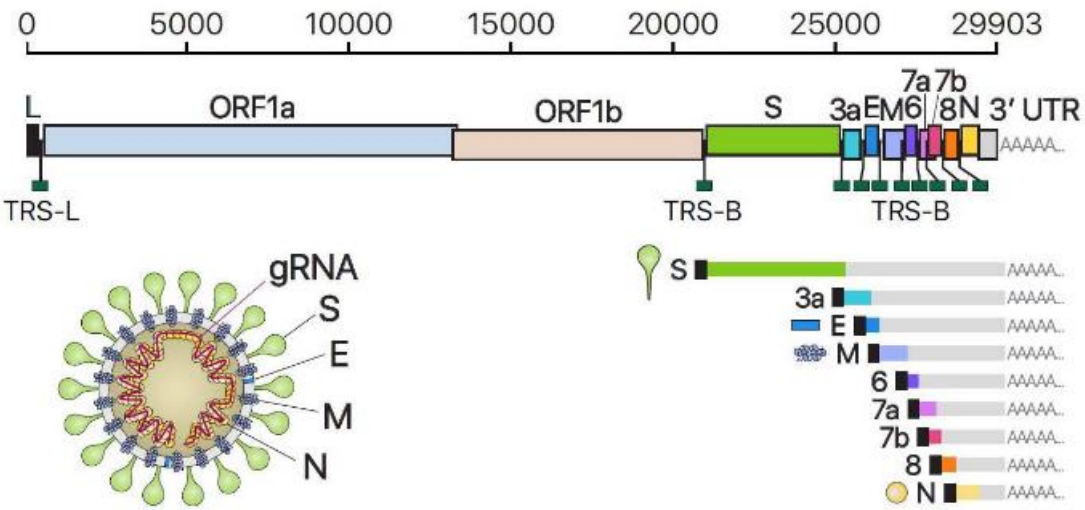

Figure 2.1: *Schematic presentation of the organization of the SARS-CoV-2 genome, and virion structure.*

The SARS-CoV-2 genome in 2/3 contains two open reading frames (ORF 1a and ORF1b) which encode two large poly-proteins (pp1a and pp1ab) and a small part encodes structural proteins (Figure 2.1) [10].

2.3 Replication cycle

Although pangolins have yet to be shed light on, snakes and turtles are all to date considered intermediate hosts of SARS-CoV-2. Bats are the natural reservoirs. The mechanisms of passage of the germ from animal to man or contrary are to be clarified [1]. Infection of the host begins with the projection of respiratory droplets from symptomatic, presymptomatic and, more rarely, asymptomatic people onto the mucous membranes (nasal, ocular, oral, etc.). You can also become infected through contaminated hands, or in case of contact with infected surfaces [11]. The virus attacks several organs expressing the angiotensin-converting enzyme 2 (ACE-2) receptor but with a strong tropism for lung cells. The main frightening characterization of the disease is infection pulmonary. After the attack on the lung, viral multiplication takes place in five (5) stages: cell attachment and entry, viral replicase transcription , genomic transcription and replication, translation of structural proteins, virion assembly and release.

Infection is facilitated by the attachment of the spike (S) protein of SARS-CoV-2 with ACE-2. Initial binding begins with the attachment of the virus to the host cell via its S1 domain to the ACE-2 receptor initiating fusion followed also by the fusion of the S protein via the S2 domain to the cellular transmembrane protein (like furin). This facilitates priming of the virus by the transmembrane serine protease (TMPRSS2). Thus the virus is cleaved and entry into the host cell occurs by endocytosis. The S protein of SARS-CoV-2 has a high binding affinity for the ACE-2 receptor.

The SARS-CoV-2 genome in 2/3 contains two open reading frames (ORF 1a and ORF1b) which encode two large poly-proteins (PP1a and PP1ab) and a small part encodes structural proteins (Figure 2.1).

After successful entry and decoating of the virus, membrane fusion releases the nucleocapsid surrounding the genomic RNA into the cytosol, subsequently the genomic RNA (sgRNA) serves as a transcript and allows cap-dependent translation of ORF1a producing the pp1a polyprotein. Then, a sliding sequence and an RNA pseudoknot towards the end of ORF1a leads 25 to 30% of the ribosomes to undergo a frameshift, hence the continuation of the translation on ORF1b then the production of the longer pp1ab poly-protein. The auto-proteolytic cleavage of PP1a and PP1ab generates sixteen (16) non-structural proteins (nsps) which have specific functions. RNA-dependent RNA polymerase (RdRP) is encoded by nsp12. nsp7 and nsp8 are the polymerase cofactors. Nsp3 and nsp5 respectively encode papain-like protease (PLpro) and main protease (Mpro). The proteins Nsp3, Nsp4 and Nsp6 are specialized in cell membrane rearrangement to form double membrane vesicles (DMVs) [12].

After release, the genomic RNA serves as a model for the synthesis of negative sense RNA using the polymerase (NSP12) by a mechanism not yet elucidated. This negative sense RNA serves as a template for the production of mRNA coding for the different structural proteins and genomic RNAs which will then be encapsidated. The structural proteins S, Envelope (E), Membrane (M) are translated by ribosomes bound to the endoplasmic

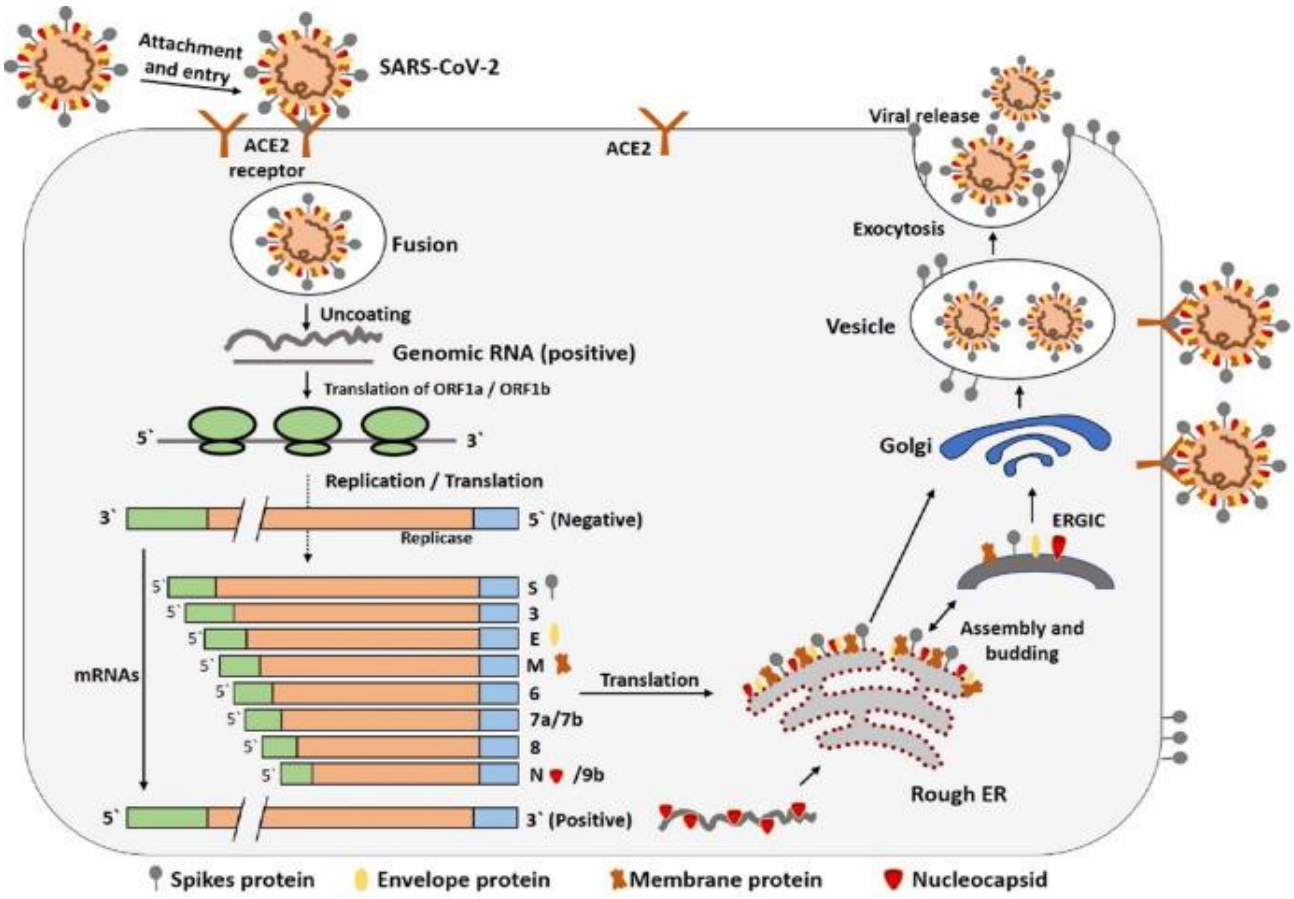

Figure 2.2: *Biological cycle of SARS-CoV-2*

reticulum (ER). The endoplasmic reticulium forms double membrane vesicles (DMV) in which the viral RNA is replicated and protected from innate immune system of the host. Then Nsp3 creates pores through which viral RNA leaves the DMVs for virion assembly.

Once replication is complete, the new genomic RNA formed associates with the N protein to form the nucleocapsid. Then there will be an interaction between the different structural proteins allowing them to associate from the nucleocapsid after different modification processes to form the membrane and then certain non-structural proteins (Nsp3, Nsp4 and Nsp6) specialized in cellular rearrangement will intervene to complete the assembly. The entire replication process takes place in the cytoplasm and assembly in the endoplasmic reticulum and Golgi intermediate compartment (ERGIC) (Figure 2.2)[13]. Once assembly is complete, the double-membrane vesicles are transported and released to the surface of the cells via the Golgi apparatus and then the virions are released by exocytosis, ready to attack other cells [10].

2.4 Diagnosis and Data generation

Three types of tests are used for the diagnosis of COVID-19, namely PCR, antigen tests and serological tests. In addition to these tests, there is sequencing for in-depth diagnosis of the germ, making it possible to follow the evolution of the virus [3].

2.4.1 Real-time PCR or qPCR (quantitative PCR)

The reference SARS-CoV-2 screening test is based on the detection of viral RNA by qPCR on nasopharyngeal samples, the performance of which would be better than on oropharyngeal samples. qPCR on saliva samples can be considered under certain conditions. The virus has also been detected in various clinical samples, such as bronchoalveolar lavage fluid, sputum, nasal swabs, fibrobronchoscope brush biopsy samples, throat swabs, feces, urine and the blood. With nasopharyngeal samples, qPCR has high sensitivity in the first week of infection from the start of the appearance of symptoms [14].

2.4.1.1 Antigen tests

Antigen tests detect proteins specific to SARS-CoV-2. These tests can be carried out on nasopharyngeal samples and samples from the lower respiratory tract. Like qPCR tests, they ensure early diagnosis of the disease from the acute phase. Their poor performance in cases of low viral load limits their recommendations for clinical use.
Like nucleic acid detection tests (NAAT), RDTs are more effective after five (5) days from the start of infection to the appearance of symptoms [15].
Their main advantage is the result delivery time of around a few minutes (10-15min). They allow faster detection of patients in the event of a high viral load.
They are less sensitive in cases of low viral load. They have a sensitivity of less than 70%. Antigen detection tests have lower sensitivity than NAATs.

2.4.2 Serological tests

Serological tests allow the detection of specific IgG and IgM antibodies (Ab) produced by the body and directed against SARS-CoV-2. These tests are carried out on blood samples and could be used to identify patients who have developed immunity to SARS-CoV-2 whether they have been symptomatic or not. These tests are more efficient seven (7) days after infection [16].

These tests can be very useful to know whether or not patients have been infected with SARS-CoV-2 and to know the serological status of exposed people (health professionals for example). Finally, these tests could also be useful in collecting epidemiological data related to COVID-19. The combination of these tests with qPCR will make a complete diagnosis of the infection.

They can provide information on the clinical course of the disease. IgM type antibodies appear from D7 and IgG type antibodies from D10 [14]

They have limited sensitivity because antibodies are only detected at a late stage of infection. Non-exhaustive test as a diagnostic test because a positive serological test result only indicates previous infection, and a negative test for antibodies cannot exclude active SARS-CoV-2 infection [16].

2.5 Genomics

2.5.1 Definition

Genomics is the exhaustive study of genomes and in particular of all genes, their arrangement on chromosomes, their sequence, their function and their role. The genomes of living organisms have considerable sizes ranging from a few thousand to billions of nucleotides (3 billion for the human genome) [17].

2.5.2 Historical

The history of genomics is linked to that of genetics, which began with the work of Gregor Mendel (Hungarian monk 1822-1884) on the transmission of characteristics on peas. Following these experiments, Mendel established three laws explaining the phenomenon of the transmission of characters: the law of uniformity of characters in the first generation, segregation of characters and the law of independent association of characters. In 1865 he presented his work at a conference and published articles. But nevertheless, this work was ignored and not considered until certain discoveries a few years later.

1901: Discovery of the phenomenon of mutations by three researchers (Hugo Devris 1848-1935, Carl Correns 1864-1933, Erich Tschermad 1871-1962) to explain the appearance of new characters. This discovery made a considerable contribution to explaining and understanding Mendel's laws. Hugo explains that mutations are the main mechanism of evolution. The combined discoveries of Gregor Mendel (laws of heredity) and Hugo were proofs that could explain the Darwinian theory of evolution: appearance (mutations) and transmission (heredity) of variations within individuals of a population. According to this theory, species (animal and plant) to evolve had to change (mutations) or adapt to environmental variations. The same year, Bateson published an English translation of Mendel's article: "Experiments in plant hybridization date";

1915: Morgan and his students provide the explanation that the element at the basis of these developments is located on a chromosome (element of the cell made up of macromolecules), support of hereditary character (chromosomal theory of heredity) called gene. The word gene was introduced in 1907 by Wilhelm Johannsen (1866-1945) to designate a

mathematical unit of heredity. The definition of the word gene was gradually refined during the twentieth century with progress in knowledge and the development of experimental approaches (genetics, molecular biology, computer science, etc.)

1957 : Discovery of proteomics by Crick;

1961 : Elucidation of the genetic code detailing the different stages of replication;

1965 : Holley and his collaborators sequenced the first two nucleic acids in history: the alanine tRNA of Escherichia coli, then that of yeast;

1971 : Sequencing of the first DNA molecule. These first sequences were obtained using specific chemical reactions, such as depurination. These methods made it possible to obtain sequences of 10 to 20 nucleotides [18].

Subsequently, the history of genomics continues with that of PCR and sequencing.

DNA sequencing consists of determining the order of sequence or succession of the nucleotides composing it. This technique is still undergoing improvements with the appearance new sequencers [19].

PCR is an enzymatic (in vitro) reaction which, in the presence of the actors (DNA, primers, deoxyribonucleotides (dntps), buffer, Taq-polymerase and water) makes it possible to select and then amplify in a very large quantity a particular nucleic acid fragment (DNA or RNA) present in very small quantities at the start, among several fragments [20].

Knowledge of these techniques is recent in relation to the history of humanity but nevertheless encompasses discoveries dating back more than a century, the main ones of which are:

The discovery of DNA: isolated for the first time by the Swiss biochemist Friedrich Miescher in 1871 of whom he had little information on the said substance but he succeeded in demonstrating that it contained phosphorus (new for researchers) and that She was sour. It was therefore named deoxyribonucleic acid.

The founding of radio crystallography in 1912 by the German physicist Max Von Laue allowed the British W. Astbury to study DNA and demonstrate its structure as a long filament and its constitution by a succession of bases.

In 1919 Phoebus Levène also worked on this constitution and specified the constitution of purine and pyrimidine bases of nucleic acids (DNA or RNA).

Griffith's work on bacterial transformation in search of a vaccine for pneumococcus allowed Oswald Avery to demonstrate that genetic information is stored on DNA.

1947-1950 Erwin Chargaff discovers the proportions of the bases adenine-thymine (30%) and cytosine-guanine (20%) in humans. He symbolized them by $A = T$ and $G \equiv C$

showing that the proportion of adenine is equal to that of thymine and that of cytosine to guanine.

Rosalind Franklin, James Watson and Francis Crick discovered the double helix structure of DNA in 1953.

Already in the sixties the necessary information on the functioning of DNA was available (macromolecules at the basis of phenotypic and genotypic characteristics). Researchers are therefore looking for ways to intervene on the structure of DNA in order to correct certain situations such as genetic diseases, drug resistance, etc. in the hope of replacing defective genes.

The discovery of enzymes called restriction endonucleases in 1965 by W. Arber, D. Nathans and H. Smith, which have the property of cutting DNA at specific sequences, facilitated this intervention and made molecular biology operational.
In 1971 the first sequence comprising a foreign gene was created and qualified at the time by the term genetic manipulation.

Immediately in 1983, Karry Mulis invented PCR, making it possible to amplify DNA exponentially using *Taq polymerase* (thermo-stable enzyme).

In 1980 Oslon and D.Burke succeeded in cloning a large quantity of DNA (250kb) and constructed artificial yeast chromosomes. They opened the way for improvement work which subsequently made it possible to clone up to two million base pairs.

From 1990, genome sequencing work on numerous species was undertaken (viruses, bacteria, nematodes, etc.). The first genome sequenced was that of *Hemophilus influenzae*.

The human genome sequencing project began in 1985 and its completion was announced in April 2003. On the same date, BLAST and GRAIL alignment software was developed for gene identification [21].

2.5.3 Sequencing

By definition, sequencing makes it possible to determine the order of succession of the nucleic acids constituting a DNA or RNA nucleotide sequence. With technological advances in this field combined with bioinformatics tools, the complete genome of the virus was identified just after the detection of the first cases. These techniques make it possible to detect new variants of the virus as they evolve [3].

In addition to other techniques such as the scanner,CT scanning, Doppler ultrasound etc. are also used to monitor certain complications of the disease [3].

2.5.4 First generation sequencers

These sequencing tools are based on two diametrically opposed principles of twentieth-century sequencing. These are that of Frederick Sanger based on enzymatic synthesis and that of Maxam and Gilbert based on chemical degradation. The Sanger technique

has had the most success in use thanks to the evolution of chemical techniques enzymatic (invention of PCR by Kary Mulis in 1985, discovery of Taq polymerase, possible industrial use, easy automation) while that of Maxam and Gilbert requires toxic chemical reactions and is limited by the resolution of the analysis and its automation is more and more difficult.

Another early generation is pyrosequencing introduced in 1988 by Ronaghi, Uhlen et al. 1998, Hyman et al. 1988). It is mainly based on the addition of a single nucleotide which is revealed in real time by luminescence detection.

2.5.4.1 Sanger method

This method is classically used to carry out small one-off sequencing (kilo-base pair). Developed by Sanger between 1975-77 in Cambridge, Great Britain called the dideoxyribonucleotide method.
The principle of this method consists of initiating DNA polymerization using a small oligonucleotide (primer) complementary to part of the DNA fragment (template) to be sequenced. The elongation of the primer hybridized to the template strand is carried out in the presence of the four deoxyribonucleotide triphosphates (dATP, dTTP, dGTP, dCTP), monomers used by the polymerase, and a dideoxyribonucleotide analogue (ddATP, ddCTP, ddGTP, ddTTP) which plays the role of chain terminator. Due to the specific incorporation of the analog by the polymerase, a mixture of fragments is obtained which selectively terminate at positions corresponding to the chosen nucleotide. Four reactions are thus carried out in parallel, each with one of the four ddNTPs, and the fragments obtained are separated by electrophoresis. In order to be able to identify the DNA fragments synthesized by the polymerase and in particular to be able to distinguish them from the template DNA, they are marked with a fluorescent tracer. This is attached to one of its two ends, either at 5' on the sequencing primer, or at 3' on the terminating dideoxyribonucleotide.

Modern automatic sequencers use an in situ detection system during electrophoresis. The beam of a laser emitting in the absorption band of the fluorophore passes through the gel. During migration, when a band of DNA passes in front of the beam, a fluorescence signal is emitted. This is captured by a photodiode located opposite the gel. The signal is amplified then transmitted to the control computer and analyzed by specialized software. Under favorable conditions, this technique makes it possible to read up to 1000 nucleotides per sequenced fragment. Routinely, the average is around 500 to 800 nucleotides per experiment [22].

2.5.4.2 Maxam and Gilbert method

This method is based on chemical degradation of DNA and uses the different reactivities of the four bases A, T, G and C, to carry out selective cuts. By reconstructing the order

of the cuts, we can go back to the sequence of nucleotides of the corresponding DNA. This chemical sequencing can be broken down into six successive steps:

- **Marking**: The ends of the two DNA strands to be sequenced are marked with a radioactive tracer (32P). This reaction is generally carried out using radioactive ATP and polynucleotide kinase.

- **Isolation of the DNA fragment to be sequenced**: This is separated by means of electrophoresis on a polyacrylamide gel. The DNA fragment is cut from the gel and recovered by diffusion.

- **Strand separation**: The two strands of each DNA fragment are separated by thermal denaturation and then purified by further electrophoresis.

- **Specific chemical modifications**: Single-stranded DNAs are subject to specific chemical reactions of different base types. Walter Gilbert developed several types of specific reactions, carried out in parallel on a fraction of each labeled DNA strand: for example, a reaction for G (alkylation with dimethyl sulfate), a reaction for G and A (depurination), a reaction for C, as well as a reaction for C and T (alkaline hydrolysis). These different reactions are carried out under very careful conditions, so that on average each DNA molecule carries only zero or one modification.

- **Cutting**: After these reactions, the DNA is cleaved at the modification by reaction with a base, piperidine.

- **Analysis**: For each fragment, the products of the different reactions are separated by electrophoresis under denaturing conditions and analyzed to reconstitute the DNA sequence.

This analysis is similar to that carried out for the Sanger method [22].

2.5.4.3 Pyrosequencing: Non-Sanger method of sequencing

Pyrosequencing is by far the most successful non-Sanger technique. This DNA sequencing technique introduced since 1988, by Hyman and colleagues (Hyman 1988), and improved by a Swedish group (Ronaghi, Karamohamed et al. 1996; Ronaghi, Pettersson et al. 1998; Ronaghi, Uhlen et al. 1998) by introduction of PCR. This is sequencing by synthesis (SBS) and is characterized by real-time revelation. of DNA polymerase activity (real time sequencing) which adds a single non-fluorescent nucleotide at a time.

Pyrosequencing takes place in 5 steps:

Step 1 : Consists of preparing the reaction mixture, with the key enzymes and the different substrates.

Step 2 : Here, the nucleotides are not added all together as in a normal sequencing reaction but one after the other. If the nucleotide added to the reaction medium corresponds to that expected by the polymerase, it is incorporated into the strand being synthesized (elongated) and releases a pyrophosphate.

Step 3 : ATP-sulfurylase then transforms this pyrophosphate (PPi) into ATP which is then used, coupled with a Luciferin, by a Luciferase. We then produce oxyluciferin and a light signal

Step 4 : Apyrase degrades excess nucleotides.

Step 5 : The light signal is captured by a CCD sensor then reproduced in the form of a peak on the pyrogram. The height of this peak depends on the intensity [23].

2.5.5 Second generation sequencers

First generation techniques have been improved to increase the resolution and sequencing time. Since 2005, these machines have been on the market called next generation sequencing (NGS) due to the low throughput and high throughput of first generation sequencers. Four second-generation sequencing platforms are currently available on the market, offering different versions of machines: Roche 454 (2005), Illumina/Solexa (2006), SoliD (life technologies Applied Biosystems 2007), Ion Torrent (life technologies Ion Torrent 2010). However, they are all broken down into 4 main stages: the preparation of the libraries which contains a PCR amplification stage, the sequencing reaction cycles, the image taking after each of these cycles to determine the corresponding nucleotide, then the data analysis. The advantage of these new generations of machines is their ability to analyze large genomes at high resolution thanks to the parallelization of reactions [22].

2.5.5.1 Roche 454

The 454 Life Sciences laboratory was the first to market these devices in 2005. Then purchased by Roche, hence the name Roche 454. The technology requires a bead PCR step to enrich the fragments to be sequenced in order to obtain several copies of the same DNA fragment. The DNAs are combined with beads, then the beads are deposited into wells on a solid support to perform pyrosequencing. We obtain up to 900 Mb of data in around ten hours, or 15,000 times more than with the first generation of sequencers. This

technology has the largest read size (up to 700 bp compared to 100 bp in its early days) and its high accuracy makes it suitable for de novo sequencing.

2.5.5.2 Illumina/Solexa

The first successful sequencer was marketed in 2006 (the Genome Analyzer GA). The specificity of this technology is based on bridge amplification (bridge PCR) of the fragments to be sequenced. It takes place on a glass surface called a "flow cell"(FC) similar to a microscope slide. The fragments of the library to be sequenced have adapters at their ends. These will allow them to attach randomly to the FC, by hybridization on the primers which cover the surface. A new strand is then synthesized by a polymerase: it is covalently attached to the FC. The original strand is then removed by denaturation, and the free end of the remaining strand hybridizes to an adjacent primer to form a bridge. The polymerase re-synthesizes the complementary strand to form a double-stranded DNA bridge, then both copies are released by denaturation. The bridging amplification cycle begins again to eventually form a grouping of clonal DNA into an area called a cluster. Sequencing is carried out on hundreds of millions of clusters simultaneously, thanks to reversible terminator chemistry: blocked nucleotides marked by fluorescence are added, when one of them is incorporated, the fluorescence emitted is recorded then the fluorophore and the blocker are cleaved allowing the addition of a new nucleotide. At each incorporation cycle, a base can be determined. This chemistry has the advantage of correctly sequencing the homopolymers. The Genome Analyzer however, has the disadvantage of reading few bases, which makes it suitable for the analysis of genomes with good annotation. More modern instruments on the market include the MiSeq, iSeq, NextSeq, and NovaSeq.

2.5.5.3 Ion Torrent

Founded in 2007 by Jonathan M. Rothberg bought in 2010 by Life Technologies, which will market its first sequencer, the PGM (Personal Genome Machine), in December 2010. The technology is based on the natural release of an H+ ion after incorporation of a nucleotide by a polymerase. This phenomenon causes a change in pH that can be detected by a semiconductor silicon chip composed of several million transistors.

This technology is called PostLight because no light intermediate is used unlike the methods mentioned previously: it is a chemical modification which leads to the creation of the signal. The absence of fluorescent marking and optical detection system or the use of standard microchips explains the low cost of this machine.

2.5.5.4 Sequencing by Oligonucleotide Ligation and Detection(SOLiD)

The third next-generation sequencing platform, it has been marketed by Applied Biosystems (currently Life Technologies) since 2007. The technology is also based on bead emulsion PCR. Sequencing is not carried out by synthesis as on previous platforms but

by ligation. A universal sequencing primer attaches to the adapter then degenerate 8-base oligonucleotides, labeled by fluorescence, are added. As soon as one of them corresponds to the sequence adjacent to the primer, the ligase fixes it and fluorescence is emitted, making it possible to identify the fixed oligonucleotide and thus to interrogate its first two bases. The number of ligation, detection and cleavage cycles thus determines the read length. Each base is read twice with this technology, which explains its high precision and which makes it suitable for resequencing or the analysis of polymorphisms. However, the complexity of operation of this technology is one disadvantage since it involves heavy analysis work.

2.5.6 The 3rd generation sequencers

The major difference between the second and third generation of sequencers lies in their ability to directly sequence individual DNA molecules without any prior amplification. We can therefore also sequence RNA without having to first convert it into cDNA.

Pacific Biosciences and Oxford Nanopore Technologies are the two major players in third-generation technologies. Other technologies, such as 10x Genomics can be considered third generation technology [24].

2.6 SARS-CoV-2 variants

Variants are subtypes of viruses that have undergone one or more mutations in the nucleotide sequence of their genome compared to the initial strain. These mutations can be silent or lead to a replacement , insertion or deletion of one or more amino acids in viral proteins. These mutations are considered when they alter the functions of the main proteins involved in the infection. SARS-CoV-2 has a strong capacity for mutation [25].

The variants are identified by comparison of their genomes with that of the initial strain of the virus detected in Wuhan in December 2019. They are distinguished from the latter by several mutations in the genome giving them other characteristics (for example greater transmissibility).
A lineage or clade (a set of viruses descended from the same ancestral viral strain) of SARS-CoV-2 carries several mutations. There are many known amino acid polymorphisms in COVID-19 variation but some have attracted attention such as E484K, N501, K417N, D614G. The D614G mutation has been associated with reduced vaccine effectiveness, strong host immune stimulation and virus contagiousness.
The WHO distinguishes three categories of variants with a fourth never yet detected:

2.6.1 Variants of concern(VOC)

Are variants for which an increase in transmissibility or an unfavorable impact on the epidemiology of COVID-19 has been demonstrated by comparing with one or more reference

viruses; an increase in severity or a change in clinical presentation or a reduction in the effectiveness of the control measures put in place (prevention measures, diagnostic tests, vaccines, therapeutic molecules). Five (5) of these variants are identified and are subject to enhanced surveillance.

2.6.1.1 Alpha variant or lineage B.1.1.7

Identified on September 20, 2020 and reported to the WHO on December 14, 2020 by the British authorities a new variant of SARS-CoV-2 called VOC 202012/01 for "Variant Of Concern, year 2020, month 12, variant 01". It is characterized by 24 mutations in its genome. The B.1.1.7 variant would be 1.4 to 1.8 times more transmissible and 1.1 to 1.7 times more virulent (risk of hospitalization or death) than common variants. It could cause infections of longer duration and is associated with a higher viral load in the upper respiratory tract, which could contribute to its increased transmissibility. However, it does not bring any change in the clinical manifestations of COVID-19 or the probability of reinfection compared to common variants. To date, validated vaccines and monoclonal antibody treatments remain effective on this subtype.
The characteristic mutations on the spike protein associated with this variant are position N501Y and D614G (first mutation of concern described from the initial strain) with the possibility of increasing propagation, contagiousness, virulence, detection by PCR.
This variant has been reported in 172 countries [26].

2.6.1.2 Beta variant or 501.V2, 20C/501Y.V2 of B.1.351 lineage

Detected for the first time on December 18, 2020 by South African national authorities, a new variant of SARS-CoV-2 named 501.V2,20C/501Y.V2 or B.1.351 lineage is spreading rapidly in three South African provinces. The B.1.351 variant would be 1.5 times more transmissible than common variants. It has been associated with an increase in lethality in the said country. This variant has been associated with a reduction in the effectiveness of vaccines (such as that of AstraZeneca/Oxford), antibody treatments, and the sensitivity of diagnostic tests. The mutations carried by the B.1.351 variant do not affect the performance of NAATs for SARS-CoV-2 screening because the 69-70 deletion is not found in this variant. In Quebec, the N501Y mutation is the only target of screening tests used to detect presumptive cases of the B.1.351 variant in samples positive for SARS-CoV-2. Eventually, a second target, the E484K mutation, would be added for some laboratories. So this variant must be monitored in the common interest of defeating the virus because this is a phenomenon that can destroy vaccine efforts for the disease. The mutations involved are K417N, E484K, N501Y, D614G with the possibility of conferring traits such as rapid spread, high contagiousness and immune evasion. This variant has been reported in 120 countries.

2.6.1.3 P.1 lineage gamma variant

The P.1 variant is part of the B.1.1.28 lineage detected on March 5, 2020 in Brazil by the Adolfo Lutz institute. The P.1 variant was detected for the first time in December 2020 in the Manaus region of Brazil but notified to the WHO on January 9, 2021. This variant carries thirty-five (35) mutations with three mutations of concern in the RBD (receptor binding domain) of the spike protein: K417N/T, E484K and N501Y.

The P.1 variant would be 1.8 to 2.5 times more transmissible and 1.1 to 1.8 times more virulent (risk of death) than common variants. It would be associated with infections with a higher viral load compared to those caused by common variants. A low risk of reinfection was estimated with this variant (6.4%) for people who were already infected during the first wave. Studies on the transmission and impacts of this variant on clinical manifestations, hospitalizations, deaths, reinfections and vaccination are limited. Due to mutations shared with the B.1.351 variant, a reduction in the effectiveness of vaccines and monoclonal antibody treatments is expected for the P.1 variant. The mutations found in the P.1 variant do not affect the performance of NAATs for SARS-CoV-2 screening. Studies are underway to evaluate the effectiveness of vaccines against the P.1 variant. However, since this variant has the same mutations that impact the biological properties of the B.1.351 variant (N501Y, E484K and K471T/N in the spike protein), a decrease in the effectiveness of plasma neutralizing antibodies convalescents or people vaccinated for COVID-19 is expected. The same goes for the effectiveness of monoclonal antibodies found in certain treatments against COVID-19.

The mutations involved are K417T/ E484K/ N501Y and D614G with the possibility of conferring traits such as rapid propagation, high contagiousness and immune evasion.
This variant has been reported in 72 countries [26].
Currently two variants of interest have joined the three variants above, bringing the list to five:

2.6.1.4 Delta variant or B.1.617.2 or G/452R lineage

Discovered for the first time in October 2020 in India, in the Nagpur region. The WHO classified this variant among the concerns on May 10, 2021 because it would be 40 to 60% more contagious than the alpha variant, which is itself already much more contagious than the historical strain of SARS-CoV-2. The Delta variant is characterized by several mutations throughout its structure, but the most important of which are located on the S protein: L19R, a deletion at position 157/158, L452R, T478K, all in the N-terminal domain; and D950N in the receptor binding domain (RBD). The L452K mutation is looked for in screening tests. It is necessary to confirm the presence of the Delta variant in a sample [27]

2.6.1.5 Omicron variant or B.1.1.529 lineage

The Omicron variant of lineage B.1.1.529 was first reported to the WHO in South Africa on November 24, 2021. It carries sixty-two (62) mutations on its genome with seventy-five percent (75%) on structural proteins and twenty-five percent (25%) on non-structural proteins. It carries thirty-six (36) mutations on the S protein. Among these mutations, five can increase its affinity with the receptor. ACE2: N501Y, G339D, T478K, N440K and S477N.

Three mutations (H655Y, N699K, P681H) at the furin cleavage site may facilitate cell penetration. The high percentage of mutations in structural proteins suggests that two antiviral drugs (lagevrio and paxlovid) will remain effective on omicron due to their targets on non-structural proteins [27].

In terms of effectiveness of neutralization by antibodies, a study showed that powerful neutralization of the omicron variant would be obtained after several doses of vaccines (Pfizer and moderna). People who have received more doses are better protected than those naïve to infections [28]. It is true that the omicron variant is known for its high transmissibility than the previous variants, however the virus seems less virulent. As with all other variants, vaccines do not prevent contracting but protect against serious forms. The effectiveness of antiviral drugs requires even more data [29].

2.6.1.6 Epsilon variant

B.1.427 and B.1.429 which emerged in the United States are now considered VOC by the American CDC due to higher transmissibility than the reference viruses but probably lower than that of the alpha variant. They are currently considered VOI by the WHO and ECDC [27].

2.6.2 Variant under investigation or variant of interest

Are variants that are characterized by a phenotypic change compared to a reference virus, by mutations that lead to changes in amino acids associated with confirmed or suspected phenotypic implications and responsible for community transmission or multiple confirmed cases or clusters.
To date these variants are VOI: Lambda and Mu, kappa, Iota, Eta [30].
These variants have been associated with a reduction in the effectiveness of treatments with monoclonal antibodies, treatments with convalescent serum and certain drug treatments [27].

2.6.3 Variants currently being evaluated

These are variants characterized by the absence of virological, epidemiological or clinical evidence supporting an impact on public health, despite the presence of mutations found

in one or more variants of interest to be followed [27].

2.6.4 High Consequence Variants

A high-consequence variant has clear evidence that prevention measures or medical countermeasures (MCMs) have significantly reduced effectiveness compared to variants that were previously circulating.
Possible attributes of a high consequence variant:

- Impact on medical countermeasures (MCM)
- Demonstrated failure of diagnostic tests
- Evidence suggesting a significant reduction in vaccine effectiveness, a disproportionately high number of vaccine breakthrough cases, or very low induced protection by the vaccine against a serious illness
- Significantly reduced sensitivity to several approved therapeutic products
- More serious clinical illness and increased hospitalizations

A variant of high significance would require notification to the WHO under the International Health Regulations, notification to the CDC, announcement of strategies to prevent or contain transmission, and recommendations for updated treatments and vaccines. Currently, there are no SARS-CoV-2 variants that have reached a high level of consequence.

Chapter 3

Constructing evolutionary models

3.1 Introduction

It has been observed that the evolution of the SARS-COV-2 virus is a process where nucleotide bases change from one base to another over time. In order to understand how a DNA sequence would have evolved in future times, modelling under various assumptions has to be conducted. In this work,we begin by studying a stochastic process $X(t)$, continuous Markov chains and use the results to refer to the continuous Markov chain evolutionary models of nucleotide substitution.

3.2 Stochastic process of evolution

Let a stochastic process X be a collection of random variables $\{X(t) : t \in T\}$ from a state space S such that $\{X(t) : t \in T\}$ is defined on the same probability space (Ω, F, $\mathbf{P}$).
The index t $\in$ T can be discrete that is X $= \{X_t, t = 0, 1, 2, 3...\}$ or continuous such that $X = X_t$, $0 \leq T < \infty$.
When X has a Markov property (3.1) then it is a Markov chain.

$$P(X_{t+1} = x|X_1 = x_1, X_2 = x_2...X_t = x_t) = P(X_{t+1} = x|X_t = x_t) \tag{3.1}$$

$X(t) : t \geq 0$ where S is discrete is called continuous time Markov process if;

(i) property (3.1) holds

(ii) $P_{ij} = P(X(t + s) = j|X(s) = i)$

A continuous Markov model is defined as a model of the form

$$X = (S, Q, F, \varphi) \tag{3.2}$$

where $S = \{s_1, s_2, ...s_n\}$ state space, Q is transition matrix that is $S \times S$ whose properties are such that

(i) $0 \leq -q_{ii} < \infty \forall i$

(ii) $0 \leq q_{ii} \forall i \neq j$

(iii) $\Sigma_j q_{ij} = 0 \forall i$

F is finite set of outputs and $\varphi : S \to F$ is the output function [31].
Making a transition at state s_j in time dt comes with a probability of making a transition to state s_k where $k \neq j$ which is given by $a_{jk}dt$.
This is the base that is used to formulate stochastic differential equations.
Important properties of Markov chains include;

(i) $X(t)$ is Time Homogeneous or stationary if:

$$P(X_{t+1} = x | X_t = y) = P(X_t = x | X_{t-1} = y) \forall t \tag{3.3}$$

(ii) $X(t)$ is memoryless if:

$$P(X_{t+1} = x_{t+1} | X_t = x_t, X_{t-1} = x_{t-1}, ..., X_0 = x_0) = P(X_{t+1} | X_t = x_t) \tag{3.4}$$

(iii) If there exists a probability distribution π then $X(t)$ is stationary if:

$$\pi Q = 0 \tag{3.5}$$

or

$$\pi = \pi \mathbf{P} \tag{3.6}$$

where

Q is jump rate matrix and $\mathbf{P}$ is the probability transition matrix.

(iv) $X(t)$ is time reversible if there is π probability distribution such that

$$\pi_i P(X_{t+1} = j | X_t = i) = \pi_j P(X_{t+1} = i | X_t = j). \tag{3.7}$$

π_i, π_j being the equilibrium frequency of i and j.

(v) if $X(t)$ can move from every state to every state then it is ergodic.

(vi) when a markov chain is irreducible, ergodic then: $\lim_{n\to+\infty} P_{ij}$ exits and is not affected by i.

The whole process is defined by P_{ij}:$i, j \in S$ the transition probabilities contained in transition probability matrix $P(t)$. In one step $P_{ij} = P(X_1 = j|X_0 = i)$ and by Chapman-Kolmogorov; given $X(t)$ has the homogeneity property and is discrete we can jump from state i to j in n steps

$$P_{ij}^n = P(X_{k+n} = j|X_k = i) \tag{3.8}$$

since

$$P_{ij} = P(X_{k+1} = j|X_k = i) \tag{3.9}$$

which gives

$$P_{ij}^{t+s} = \Sigma_{k\in S} P_{ik}^t P_{kj}^s \tag{3.10}$$

in transition matrix we get

$$\mathbf{P}^{t+s} = \mathbf{P}^t\mathbf{P}^s \tag{3.11}$$

In continuous Markov chains, by Chapman-Kolmogorov the probability transition over $t + s$ seconds is given as

$$\mathbf{P}(t + s) = \mathbf{P}(t)\mathbf{P}(s) \tag{3.12}$$

3.3 Extending to evolutionary models

Let $X(t)$ be a stochastic process describing nucleotide base change. This process has random events of substitution that are found in the state space $S = \{A, T, C, G\}$.

Since substitution events are continuous then the process $X(t)$ is such that
$X = X_t, 0 \leq T < \infty$.

Let N be a site on a DNA sequence and let $x, y \in S$ be any DNA elements in S.

By the nature of evolution each site N is independent of other sites and there exists the same process for all sites. The properties and definitions stated above of continuous time Markov chain hold for each site N.

Since each mutation outcome is defined by the transition probabilities, we then find the most general form of a transition matrix is given by $\mathbf{P}(t) = e^{Qt}$.

proof. Given that $S = (A, G, C, T)$. Let $P(t) = (P_A(t), P_G(t), P_C(t), P_T(t))$ be probabilities of the states in the state space at time t for the site N .

For $x, y \in S$, let μ_{xy} be the transition rate from state x to state y.

For any $x \in S$, let $\mu_x = \sum_{y\neq x} \mu_{xy}$
which
$P_x(t)$ =probability of nucleotide x in S at time t. After time Δt, the change in probability is given by $P_x(t + \Delta t)$ is the probability of nucleotide x in S at time $t + \Delta t$. To find $P_x(t + \Delta t)$, consider the following,

(i) the frequency of the nucleotide at time $t + \Delta t$

(ii) the frequency of the nucleotide at time t,

(iii) the lost of the nucleotide during Δt,

(iv) the gain of the nucleotide during Δt.

Thus for a nucleotide $x \in S$ the change in probability is given as:

$$P_x(t + \Delta t) = P_x(t) - P_x(t)\mu_x \Delta t + \sum_y P_y(t)\mu_{yx}\Delta t, y \in S \tag{3.13}$$

Explicitly this becomes:

$$\begin{cases} P_A(t+\Delta t) &= P_A(t) - P_A(t)\mu_A\Delta t + P_G(t)\mu_{GA}\Delta t + P_C(t)\mu_{CA}\Delta t + P_T(t)\mu_{TA}\Delta t \\ P_G(t+\Delta t) &= P_G(t) - P_G(t)\mu_G\Delta t + P_A(t)\mu_{AG}\Delta t + P_C(t)\mu_{CG}\Delta t + P_T(t)\mu_{TG}\Delta t \\ P_C(t+\Delta t) &= P_C(t) - P_C(t)\mu_C\Delta t + P_A(t)\mu_{AC}\Delta t + P_G(t)\mu_{GC}\Delta t + P_T(t)\mu_{TC}\Delta t \\ P_T(t+\Delta t) &= P_T(t) - P_T(t)\mu_T\Delta t + P_A(t)\mu_{AT}\Delta t + P_G(t)\mu_{GT}\Delta t + P_C(t)\mu_{CT}\Delta t \end{cases} \tag{3.14}$$

We can rewrite equation (3.14) in matrix form as:

$$\begin{pmatrix} P_A(t+\Delta t) \\ P_G(t+\Delta t) \\ P_C(t+\Delta t) \\ P_T(t+\Delta t) \end{pmatrix} = \begin{pmatrix} P_A(t) \\ P_G(t) \\ P_C(t) \\ P_T(t) \end{pmatrix} + \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_G & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_C & \mu_{AC} & -\mu_C & \mu_{TC} \\ \mu_T & \mu_{AT} & \mu_{GT} & -\mu_T \end{pmatrix} \times \begin{pmatrix} P_A(t) \\ P_G(t) \\ P_C(t) \\ P_T(t) \end{pmatrix} \Delta t \tag{3.15}$$

$$\Rightarrow \mathbf{P}(t + \Delta t) = \mathbf{P}(t) + Q\mathbf{P}(t)\Delta t \tag{3.16}$$

$$\frac{\mathbf{P}(t + \Delta t) - \mathbf{P}(t)}{\Delta t} = Q\mathbf{P}(t) \tag{3.17}$$

$$\Rightarrow \lim_{\Delta t \to 0} \frac{\mathbf{P}(t + \Delta t) - \mathbf{P}(t)}{\Delta t} = Q\mathbf{P}(t) \tag{3.18}$$

$$\Rightarrow \mathbf{P}\prime(t) = Q\mathbf{P}(t) \tag{3.19}$$

solving the ordinary differential equation(ODE) using separation of variables method we get:

$$\mathbf{P}(t) = e^{Qt} \tag{3.20}$$

where:

$$\mathbf{P}(t) = \begin{pmatrix} p_{AA}(t) & p_{GA}(t) & p_{CA}(t) & p_{TA}(t) \\ p_{AG}(t) & p_{GG}(t) & p_{CG}(t) & p_{TG}(t) \\ p_{AC}(t) & p_{GC}(t) & p_{CC}(t) & p_{TC}(t) \\ p_{AT}(t) & p_{GT}(t) & p_{CT}(t) & p_{TT}(t) \end{pmatrix} \tag{3.21}$$

cqfd. if Q is diagonalisable such that $Q = vAv^{-1}$ then $\mathbf{P}$(t)=$ve^{At}v^{-1}$

3.4 Models of evolution

Having derived how models of nucleotide substitution are modelled, we take an inventory at some of the widely used models by researchers who study phylogenetics.

The most widely and generally useful nucleotide substitution models are those from the General Time-Reversible (GTR) family [32]. Software specialised in phylogenetic analysis like Molecular Evolutionary Genetics Analysis (MEGA) [33], Phylogenetic Analysis Using Parsimony (PAUP) [34], PHYLogeny Inference Package (PHYLIP) [35] or PHYlogenetic inferences using Maximum Likelihood (PHYML) [36] make use of these models. The GTR family of models consists of about 203 models all varying in assumption. Despite having such large number of models, the most common and widly used of the 203 models include Jukes Cantor (JC69) [37], Kimura 2-parameter (K80 or K2P) [38], Felsenstein (F81) [39], Hasegawa-Kishino-Yano (HKY 85) [40], Tamura-Nei (TN 93) [41], and GTR [42]. Debates as to which model is best has always depended on which model describes the process of nucleotide evolution on the dataset accurately. The chosen model determines the results of the phylogenetic analysis. We avoid biased inferences by avoiding models that fit the data poorly. Bollback in his study uses Bayesian method to evaluate the overall adequacy of DNA models used in sequence evolution [32]. He considered GTR, HKY 85, K2P and JC69 as the commonly used models in phylogenetic literature.

The general time-reversible (GTR) model for the last decade has been the most used model of nucleotide substitution. For various applications used to test for best model, this model has been at the top of these tests. All GTR family models assumes homogeneous rates among site and when rate variation is taken into account, the gamma model G model is added to GTR model giving another good model called $GTR + \Gamma$. Deciding on adding an invariant site to the process of substitution gives the best model called $GTR + \Gamma + I$ model [43].

Due to how important nucleotide models are in phylogenetic analysis, it has now become a criterion before performing any phylogenetic study that a model test is conducted. To pick the best model,researchers conduct model tests that give the best model that fit the data being studied. Maximum likelihood ratio test and Akaike information criterion are commonly used statistical measures in model choice.

3.4.1 Likelihood ratio test

The likelihood ratio is given by

$$\delta = 2(\ln L_1 - \ln L_0) \tag{3.22}$$

whereby L_1 is the complex model maximum likelihood giving the alternative hypothesis and L_0 is the simple model maximum likelihood giving the null hypothesis. This method is the most widely used method for comparing two competing models [44]. This method compares nested models where the model with more parameters is the complex model and the simple model has less parameters. The likelihood functions are conditional probabilities of the sequences (Data) given the model of evolution with its parameters and the tree i.e;

$$L_1(T, \varphi) = Prob(D|T, \varphi) = prob(sequences|tree, model\ of\ evolution) \tag{3.23}$$

$$L_2(T, \varphi) = Prob(D|T, \varphi) = prob(sequences|tree, model\ of\ evolution) \tag{3.24}$$

we make the likelihood functions to have the largest values by getting maximum estimates for T, φ i.e

$$\hat{T}, \hat{\varphi} = max(T, \varphi) \tag{3.25}$$

δ is χ^2 distributed with the number of difference in free parameters between the two models being the degree of freedom.

3.4.2 Akaike information criterion (AIC)

A different approach in comparing competing models is to compare all competing models minimum theoretical Akaike information criterion (AIC) given by

$$AIC = -2 \ln L + 2n \tag{3.26}$$

Where L gives the maximum likelihood for a model being considered using n independently adjusted parameters [44].

When n $<$ 40, corrected AIC is used i.e;

$$AIC_c = -2(lnL) + 2n + \frac{2n(n+1)}{k-n-1} \tag{3.27}$$

k:sample size [45]
But note that as k increases, $\frac{2n(n+1)}{k-n-1} \approxeq 0$ and $AIC_c \approx AIC$
AIC does not support a model with lots of parameters thus models that fit the data well

but with less parameters are selected by it. Smaller values of AIC are required since they indicate better models.

Log-likelihood can be used to measure overall fit of a model where the larger value gives the better model. The model test to use depends on the goal being achieved.

The models that are given as results from a model test include Jukes-Cantor,Kimura Felsenstein , Hasegawa-Kishino-Yano and GTR [44]. To each of these models, a gamma distribution and invariant sites can be added. There are various software that can be used to perform the model test the prominent two being **jmodeltest** and **MEGA**. The common models considered in MEGA are considered in jmodeltest but what makes these software differ is the total number of models considered. The Table 3.1 gives description of the most used models considered in both MEGA and jmodeltest.

Tables 3.1: *Models considered during model test*

Model	free Parameters	Base frequencies	substitution rates
HKY	4	unequal	AC=AT=CG=GT;AG=CT
JC	0	equal	AC=AG=AT=CG=CT=GT
K81	3	unequal	AC=AG=AT=CG=CT=GT
K80	1	equal	AC=AT=CG=GT;AG=CT
TNef	2	equal	AC=AT=CG=GT;AG;CT
TN	5	unequal	AC=AT=CG=GT;AG;CT
TPM1	2	equal	AC=GT;AT=CG;AG=CT
TPM1uf	5	unequal	AC=GT;AT=CG;AG==CT
TVM B	7	unequal	AC;AT;CG;GT;AG=CT
SYM	5	equal	AC;AG;AT;CG;CT;GT
GTR	8	unequal	AC;AG;AT;CG;CT;GT

3.4.3 The inputs of the models

When modelling evolution, nucleotide substitutions have generally been assumed to follow a stochastic process, $X(t)$ that holds true to the Markov chain property i.e., $P(X_{t+1} = x|X_1 = x_1, X_2 = x_2...X_t = x_t) = P(X_{t+1} = x|X_t = x_t)$ such that $X(t)$: t ≥ 0 [46]. The Markov chain property states that for a given site in a sequence alignment, the probability that it will change from one state to another is conditioned only on the current state and is not affected by any previous state [47].Thus most models of evolutionary change, including those discussed in this thesis, are part of the continuous-in-time (i.e.,$X = X_t$, $0 \leq T < \infty$) Markov chain family of models where nucleotide bases have one of four states (Adenine, Guanine, Cytosine and Thymine), $K = 4$ i.e., in any given mutation event a change can occur from any of the four bases (A, G, C, or T) to any of the other three bases in the state space where K is the state space [48].

The stochastic process $X(t)$ at each site, N, is determined by the rate matrix, $\mathbf{R}$, (equation (3.29)) where each entry in matrix, $\mathbf{R}$, is defined as r_{ij} the relative rate of change from state j to state i over a period of time, t, where $r_{ii} = \sum_{i \neq j} r_{ij}$ and Q the instantaneous

rate matrix that includes $\pi = (\pi_1, \pi_2, \pi_3, \pi_4)$ a frequency distribution matrix representing the frequency of each of the four nucleotide states at equilibrium with the condition that $0 \leq \pi_i \leq 1$; $\forall i$ and $\sum_i \pi_i = 1$ for all four nucleotides.

$$\mathbf{R} = \{r_{ij}\} = \begin{pmatrix} r_{AA} & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & r_{CC} & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & r_{GG} & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & r_{TT} \end{pmatrix} \tag{3.28}$$

$$\mathbf{Q} = \{\mathbf{R}\pi_i\} = \begin{pmatrix} r_{AA}\pi_A & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{CA}\pi_A & r_{CC}\pi_C & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{GA}\pi_A & r_{GC}\pi_C & r_{GG}\pi_G & r_{GT}\pi_T \\ r_{TA}\pi_A & r_{TC}\pi_C & r_{TG}\pi_G & r_{TT}\pi_T \end{pmatrix} \tag{3.29}$$

Mathematically, the rate matrix, Q, must satisfy two conditions to be considered valid i.e.:

$$q_{ij} > 0 \forall i \neq j \tag{3.30}$$

$$q_{ii} = -\sum_{j \neq i} r_{ij} \tag{3.31}$$

The instantaneous matrix, Q, and the frequency distribution matrix, π, together form the probability transition matrix, $P(t) = e^{Qt}$. This matrix defines the probability of changing from one base to another at a site, N. Solving the differential equation of the probability transition matrix $P^{'}(t) = QP(t)$ at an initial condition of $P(0) = i$ results in $P(t) = e^{Qt}$; a probability transition matrix were the position in the ijth entry is the probability that a state i will mutate to a state j during the evolutionary time interval of length, t [49].

Due to the complexity in the processes that lead to mutations, the construction of nucleotide substitution models has relied on simplifying assumptions about these processes [50]. These assumptions can range from extremely strong (such as mutational processes result in all substitutions occurring at exactly the same rate) to very relaxed (such as variations in mutational processes for different nucleotides result in different substitutions happening at different rates).

3.4.3.1 The Jukes-Cantor model

Starting off with the simplest model, Jukes-Cantor model equation , that assumes that the frequencies of the four different nucleotides A, C, G and T are equal on any given genome sequence such that $\pi = (0.25, 0.25, 0.25, 0.25)$ and that the relative rates of all possible nucleotide substitutions are the same: i.e. given i and j are bases, the rate of i to j will be the same for all base changes (i.e., $r_{ij} = \alpha$; $\forall i, j$; Equation (3.32)) [51]

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & \alpha\pi & \alpha\pi & \alpha\pi \\ \alpha\pi & - & \alpha\pi & \alpha\pi \\ \alpha\pi & \alpha\pi & - & \alpha\pi \\ \alpha\pi & \alpha\pi & \alpha\pi & - \end{pmatrix} \tag{3.32}$$

3.4.3.2 The Kimura 2 parameter model

The Kimura 2 parameter model enables transversion substitutions and transition substitutions to occur at different rates (equation (3.33)). In reality transitions do in fact generally occur more commonly than transversions. The Kimura 2 parameter model maintains an equal frequency distribution matrix $\pi = (0.25, 0.25, 0.25, 0.25)$ as does the JC model.

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & \beta\pi & \alpha\pi & \beta\pi \\ \beta\pi & - & \beta\pi & \alpha\pi \\ \alpha\pi & \beta\pi & - & \beta\pi \\ \beta\pi & \alpha\pi & \beta\pi & - \end{pmatrix} \tag{3.33}$$

3.4.3.3 The Felsenstein model (F81)

A more parameterized of the Kimura 2 parameter model, F81 Felsenstein (F81) in equation (3.34), not only permits transitions and transversions to occur at different frequencies but also allows for differences in the equilibrium frequencies for different nucleotides (i.e. all nucleotides are not constrained to remain at a frequency of 0.25) [52].

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & - & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & - & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & - \end{pmatrix} \tag{3.34}$$

3.4.3.4 The TN93 model (Tamura and Nei 1993)

TN93, the Tamura and Nei 1993 model, distinguishes between the two different types of transition; i.e. $(A \leftrightarrow G)$ is allowed to have a different rate to $(C \leftrightarrow T)$. Transversions are all assumed to occur at the same rate, but that rate is allowed to be different from both of the rates for transitions.

TN93 also allows unequal base frequencies ($\pi_A \neq \pi_G \neq \pi_C \neq \pi_T \pi_A \neq \pi_G \neq \pi_C \neq \pi_T$).

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & \alpha\pi_G & \pi_C & \pi_T \\ \alpha\pi_A & - & \pi_C & \pi_T \\ \pi_A & \pi_G & - & \beta\pi_T \\ \pi_A & \pi_G & \beta\pi_C & - \end{pmatrix} \tag{3.35}$$

3.4.3.5 The GTR model

Multiple models have been formulated to further relax the assumptions of the K80 and F81 models in pursuit of greater realism. To date the most realistic models that are in widespread use are those of the general time- reversible (GTR) family. As with F81, GTR allows nucleotide frequencies to vary, but GTR additionally allows the changes between the different nucleotide states to all occur at different rates. So,for two given states, i and j, the two states have the same probability or rate of mutating from one to the other. For example, the rate, m, of A changing to G is equal to the rate, n, of G changing to A. This property is referred to as time-reversibility and yields a symmetrical rate matrix, Q (Equation (3.36)) [53].

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & d\pi_T \\ a\pi_A & - & c\pi_G & e\pi_T \\ b\pi_A & c\pi_C & - & f\pi_T \\ d\pi_A & e\pi_C & f\pi_G & - \end{pmatrix} \tag{3.36}$$

Additional modifications have been made to GTR and other simpler models to, for example, account for (1) the occurrence of invariant nucleotide sites (which are common in datasets of closely related sequences) and (2) rate heterogeneity where different nucleotides sites across an analysed set of sequences are evolving at different rates [52].

3.4.4 Models and rate of variation among sites

Sites have been assumed to evolve independently of other sites but identical. It is important to consider the fact that sites on a sequence do not always evolve at the same rate as others thus it will be important to add the rate variation among sites to all the model of use since all the above models assume homogeneity evolutionary process among sites. The mutation rate is considered as a variable who's values have a gamma distribution with parameters α and β.α governs the shape that the gamma distribution will have which ranges from bell shaped for larger values of alpha and L-shaped for smaller values. The main importance of considering rate variation is to avoid biasedness in the rates of change from one nucleotide to another. By conducting model tests, one can determine if the inclusion of rate variation is necessary for the given data set.

3.4.5 Selecting the best model

Having looked at models used in phylogenetic analysis by other authors in retrovirology, we picked what we consider the best five models and did an in depth study about them so as to choose the best. Below is the summary of the models.

Jukes-Cantor (**JC69**): This is the simplest model for it has only one parameter μ which is the overall substitution rate and the model has zero free parameters. It assumes that all

the bases occur by the same frequency hence equal equilibrium probabilities with a uniform stationary distribution. It assumes equal mutation rates, sites evolve independently and are time reversible. Due to the number of parameters and assumptions made, **JC69** is computationally the easiest model but then it is not realistic.

Kimura 2 parameter (**K80**): This model assumes that all the bases occur by the same number hence equal equilibrium probabilities with a uniform stationary distribution. When it comes to the rate of substitution (mutation rates), this model assumes that the substitutions between bases from the same nitrogenous base group has a different rate with substitutions between bases from different nitrogenous base group. **K80** model has one free parameter thus computationally easy. Despite **K80** taking into account the difference between transition and transversion substitutions, its assumption on base frequencies makes it unrealistic.

Felsentein (**F81**): The model allows for base frequencies to differ (flexible stationary distribution) but still keeps equal rates of substitution. Thus the model becomes unrealistic for considering equal rates of substitution.

Felsentein (**F84**): The model has four parameters, it allows for base frequencies to differ (flexible stationary distribution) and when it comes to the rate of substitution (mutation rates) it assumes different rates for transition and transversions. Though its rate matrix is more complicated than other models, mathematically it is tractable.

Hasegawa-Kishono-Yano (**HKY85**): It allows for base frequencies to differ (flexible stationary distribution) and When it comes to the rate of substitution (mutation rates), this model assumes that the substitutions between bases from the same nitrogenous base group has a different rate with substitutions between bases from different nitrogenous base group.**F84** and **HKY85** are closely related.

Tamura-Nei 93 (**TN93**): allows for base frequencies to differ (flexible stationary distribution). When it comes to mutation rates, it assumes that the transitions has two different rates i.e $A <> G, T <> G$ and there is also one rate for transversions.

From the study of literature, we pick the GTR model as the best model that best explains the process of nucleotide evolution. To this model, we disqualify the assumption of homogeneous evolution across all sites and apply rate heterogeneity among sites. In the next section, we conduct an in-depth study of the $GTR + \Gamma$ model examining its assumptions and limitations.

3.4.6 Best fit model

So far, we have worked under the assumption that the $GTR + \Gamma$ model is the best model that best estimates the process of evolution. The next step is to use the sequence to test for the best model. We conducted a model test using jmodeltest. The Table (3.2) displays results of 16 measured models using jmodeltest. For AIC, each of the models above was measured relative to all other models so as to find the relative quality of each

model. In the column containing AIC scores, as the number of parameters reached 52, the estimate of the amount of information lost was minimial in the $GTR + \Gamma$ model. Despite $GTR + I$ having parameters 52, its AIC number is high. Based on the AIC values, the $GTR + \Gamma$ model offers the least amount of information lost during the process of modelling evolution. The models in the table above can be nested into one another and using likelihood ratio test one can calculate whether the more parametrised model is necessary. From the many pairs of nested models that can be made, of importance is the nested pair $GTR+\Gamma$ and $GTR+\Gamma+I$ for these two models both have lowest loglikelihood values. Using

$$\delta = 2(lnL_1 - lnL_0)$$

we calculate the likelihood ratio which is χ^2 distributed with degrees of freedom equal to 1.

$$\delta = 2(-71760.5899 + 71757.3231) = -6.5336 \tag{3.37}$$

with significance level $\alpha = 0.05, df = 1$ and critical value = 3.841.

Since the critical value is greater than the calculated, we have insufficient evidence to reject the null hypothesis thus the higher model $GTR+\Gamma+I$ does not significantly fit the data better than $GTR+\Gamma$. This means the best model under likelihood ratio in jmodeltest is $GTR + \Gamma$. Using the corrected AIC values, the model which best approximates how the sequences evolved by looking at the data we have is $GTR + \Gamma$.

Using jmodeltest as shown the Table (3.2) the top four models based on overall fit of the models estimator -lnL in descending order are $GTR+\Gamma$, $TN+\Gamma$, $GTR+I+\Gamma$, $HKY+\Gamma$.

It has been observed that the model that fits our data using log likelihood is $GTR + \Gamma$ but using AIC scores the best model is $F81 + \Gamma$ putting $GTR + \Gamma$ at number three. From the two combined results, $GTR + \Gamma$ model was declared the best model.

Tables 3.2: Results of Model test using jmodeltest

Model	-lnL	Parameters	AIC
F81 + Γ	71761.0511	48	143619.9674
HKY + Γ	71760.8332	49	143621.6101
GTR + Γ	71757.3231	53	143622.9203
TN + Γ	71760.5246	50	143623.073
F81 + I + Γ	71762.9897	49	143625.923
HKY + I + Γ	71762.5518	50	143627.1274
GTR + I + Γ	71760.5899	54	143631.5407
TN + I + Γ	71763.8224	51	143631.7505
K80 + Γ	72942.195	46	145978.1032
K80 + I + Γ	72942.4466	48	145981.2017
JC + I + Γ	72948.729	46	145991.1712
JC + Γ	72953.6541	45	145998.9479
SYM + Γ	72814.9284	50	145731.8806
SYM + I + Γ	72815.1815	51	145734.4685
GTR	72955.9807	52	146018.1505

3.5 $GTR + \Gamma$ model

The models of evolution as described above vary based on assumptions being held. $GTR+\Gamma$ model unlike other models above, has the most number of assumptions and thus is considered the most advanced model. GTR is also the model with the highest number of parameters, it contains six different rates of substitution making a 12 parameter rate matrix and that is why its transition probabilities cannot be expressed algebraical but rather is only solved numerically. Below are the assumptions that are considered when evolution is being simulated using the $GTR + \Gamma$ model.

3.5.1 GTR + Γ model model assumptions

(i) Each site on a DNA sequence evolves independently of other sites, it is independent of its own history and has the same Markov process of substitution as other sites.

(ii) The model holds under time reversibility that is for the GTR process X(t);$\pi_i Q_{ij} = \pi_j Q_{ji}$.

(iii) The model assumes that the frequencies of the four bases(A,G,C,U) are different and this is due to the differences in the physiochemical properties of each nucleotide.

(iv) The GTR model incorporates different rates of substitution for every change. It not only acknowledges that transitions are likely to occur than trans-versions but also considers that rates are different among transitions as well as among trans-versions.

This is due to the fact that in molecular biology replacement of similar structure is likely to occur.

(v) Stationarity: This means that for the $GTR + \Gamma$, it has a certain stationary distributions such that once it gets to the level of this equilibrium distribution, it will stay at that distribution throughout in the future.

(vi) The GTR model assumes that all sites evolves at the same rate which has proved to not be true thus the nesting of the gamma distribution to account for rate heterogeneity among sites. For $GTR + \Gamma$ model the assumption is that sites do not evolve at the same rate.

3.5.2 The input of $GTR + \Gamma$ model

In order to find the probabilities at which one base changes into another, the model requires the following;

(i) Base frequencies: $\pi = (\pi_A, \pi_C, \pi_T, \pi_G)$

(ii) rates: $q_{AT} = \beta, q_{AG} = \eta, q_{AC} = \delta, q_{CG} = \epsilon, q_{CT} = \alpha, q_{GT} = \gamma$

(iii)

$$Q = \begin{pmatrix} -(\alpha\pi_C + \beta\pi_A + \gamma\pi_G) & \alpha\pi_C & \beta\pi_A) & \gamma\pi_G \\ \alpha\pi_T & -(\epsilon\pi_T + \delta\pi_A + \epsilon\pi_G) & \delta\pi_A & \epsilon\pi_G \\ \beta\pi^T & \delta\pi_C & -(\beta\pi_T + \delta\pi_C + \eta\pi_G) & \eta\pi_G \\ \gamma\pi_T & \epsilon\pi_C & \eta\pi_A & -(\gamma\pi_A + \epsilon\pi_C + \eta\pi_A) \end{pmatrix} \tag{3.38}$$

The rate of change from a state i to i is negative so as to show that the nucleotide will take a change away from itself at any time. In most studies, off diagonal parameters are usually normalised to 1 The equilibrium frequencies for all bases must add up to one thus we can work on estimating only three frequency parameters. For the six rate parameters in the rate matrix, by normalization of the matrix we get to estimate only five rates. Thus this leaves the GTR model with eight free parameters.

Since we are considering rate variation among sites, we have to estimate two more parameters which are the α for the shape and $\gamma$for the scale. Maximum Likelihood methods are used to estimate the amount of rate variation among the sites which thus gives us the shape parameter.

The rate matrix Q i.e. $Q = Ve^AV^{-1}$ is diagonalisable with eigenvalues and eigenvectors which cannot be expressed in an algebraic form. Thus we can say we are only able to

numerically compute the eigenvalues and eigenvectors. We will have to estimate all inputs that is stationary probabilities,the rate parameters and the time taken by using maximum likelihood.

To obtain the transition matrix for the GTR model, it is important to note that due to its many number of parameters it is not possible to have an algebraic expression of its probability transition matrix but rather we are only able to numerically compute it using $e^{Qt} = Ve^{A}V^{-1}$.

3.5.3 Limitation of the $GTR + \Gamma$ model

Despite the $GTR + \Gamma$ model being the model that best explains the evolution of SARS-COV-2, there are two main problems that arise out of using the model. The first flaw is a mathematical problem due to lack of closure under matrix multiplication for GTR rate matrices while the second one is a problem that arise from the dynamics of evolution. Below, these flaws have been discussed in detail.

3.5.3.1 GTR and lack of closure under matrix multiplication

A homogeneous Markov process is an assumption in Markov chains. But for evolutionary processes the substitution rates can not be the same at all times thus evolution is considered heterogeneous. This means given the process $X(t)$ Consider time $t = 0$ to t_1 governed by Q_1 and $P_1(t_1) = e^{Q_1 t_1}$. Assuming the process takes a rest and continues over time t_2 such that its rate matrix is different Q_2 and $P_2(t_2) = e^{Q_2 t_2}$

From chapman-kolmogorov property, $t = 0$ to $t = t_1 + t_2$ can be expressed as

$$(P_1(t_1) = e^{Q_1 t_1})(P_2(t_2) = e^{Q_2 t_2}) = \bar{P}$$

We can consider a single rate matrix $\bar{Q} = Q_1 Q_2$ such that $\bar{P} = e^{\bar{Q}(t_1+t_2)}$. For rate matrices in the class GTR $\bar{Q} = Q_1 Q_2$ may not always be true thus there is lack of closure under matrix multiplication for the GTR model. As a result, chapman-kolmogorov property does not hold since the product of two probability transition matrices does not give a GTR probability substitution matrix.

3.5.3.2 The evolutionary dynamics of SARS-COV-2 not captured by $GTR+\Gamma$

Note that the $GTR + \Gamma$ model does not take into account all evolutionary dynamics, namely recombination. Retroviral recombination occurs when a cell simultaneously contains two genetically different variants and, therefore, the new variant produced contains the genetic information of both variants. Therefore, the phylogenetic tree constructed from the simulated data will give erroneous conclusions if recombination is not taken into account at all. A tree leaf will be considered to have evolved from a common ancestor with other leaves when in reality it is an entirely new variant formed by two different

variants exchanging genetic information. Since there is always the assumption that all variants studied have a unique history, failing to take recombination into account will result in incorrect phylogenetic inference.

Chapter 4

COVID-19 Data in Burundi

4.1 Study sites and population

Our study focused on positive SARS-CoV-2 samples collected in different provinces of Burundi including the municipality of Bujumbura.
The study population comprised COVID-19 cases aged $\geq$ 1 year who tested positive for the SARS-COV-2 at the screening or control test performed at different testing facilities across the country.

4.2 Type of study

We conducted a cross-sectional descriptive study on samples collected from different COVID-19 screening sites.

4.3 Study periode

We used the positive SARS-CoV-2 confirmed by samples collected in Burundi from May 2021 to December 2021.

4.4 Sampling

We downloaded the Burundian sequences available on the GISAID[54] database on the date of 24 December,2023.

4.5 Bioinformatics analysis

The viral genomes used in this study were sequenced by using MinIon at Uganda Virus Research Institute (UVRI), and Illumina at Institut Pasteur de Dakar (IPD), Senegal.
The generated FASTA sequences were fed into the PANGOLIN tool [55] and Nextclade [56] to assign the SARS-COV-2 PANGO lineages and clades.
Furthermore, the consensus sequences generated in this study, multiple sequence alignment was performed using MUSCLE tool [57] against the reference SARS-COV-2 genome. The sequence alignment was used to infer a neighbour joining (NJ) phylogenetic tree using MEGA11 software [58] and MEGA6 to determine the best-fit nucleotide substitution model. We used an Ultrafast Bootstrap of 1000 replicates to assess the NJ tree branch supports [59]. The generated NJ phylogenetic tree (in a Newick file format) was then fed into ITOL v5 software [60] for genome visualization. Moreover, the amino acid mutation profiles of SARS-COV-2 genomes identified in Burundi were investigated using the Nextclade tool.

4.6 Statistical data analysis

The statistical data from this study were parsed and analyzed in R software. The proportions of different lineages were calculated in R and visualised using the ggplot2 package.

Chapter 5

COVID-19 sequence data analysis

5.1 Overall results

Overall, 158 SARS-COV-2 genomes were successfully sequenced in the reference sequencing laboratory of Uganda (Uganda Virus Research Institute) and Senegal (Institut Pasteur de Dakar), from May 2021 to December 2021.

The age of the study participants ranged between 1 and 76 years and 46.2% (73/158) were males while females represented 53.8% (85/158). They were distributed across different provinces of the country including the municipality of Bujumbura.

The female gender was the most represented with 53.8% (Table 5.1).

Fourty-seven (47%) of the patients were aged over 30 years (Table 5.2).

Tables 5.1: *Distribution of samples according to sex*

Sex		
Male n(%)	Female n(%)	Total n(%)
73(46.2%)	85(53.8%)	158(100%)

Tables 5.2: *Distribution of samples according to age*

Age	Effectif(%)
≤ 30	84(53.16%)
> 30	74(46.84%)
Total	158(100%)

5.1.1 The genetic diversity of the SARS-COV-2 lineages

The 158 genomes sequenced in Burundi during the study period (from January 2021 to December 2021) were classified into 12 SARS-COV-2 PANGO lineages. Of the 12 PANGO lineages, 5 accounted for 77.84% of all the SARS-COV-2 genomes sequenced in Burundi during this study period. These lineages included BA.1 (7.59% ,12/158), B.1.617.2 (Delta) lineage (17.08%, 27/158), BA.1.13(24.05%, 38/158), AY.46(18.98%, 30/374), and B.1.1(10.12%,16/158). All 5 common PANGO lineages were variants of concern (VOCs) or their sublineages (Figure 5.1).

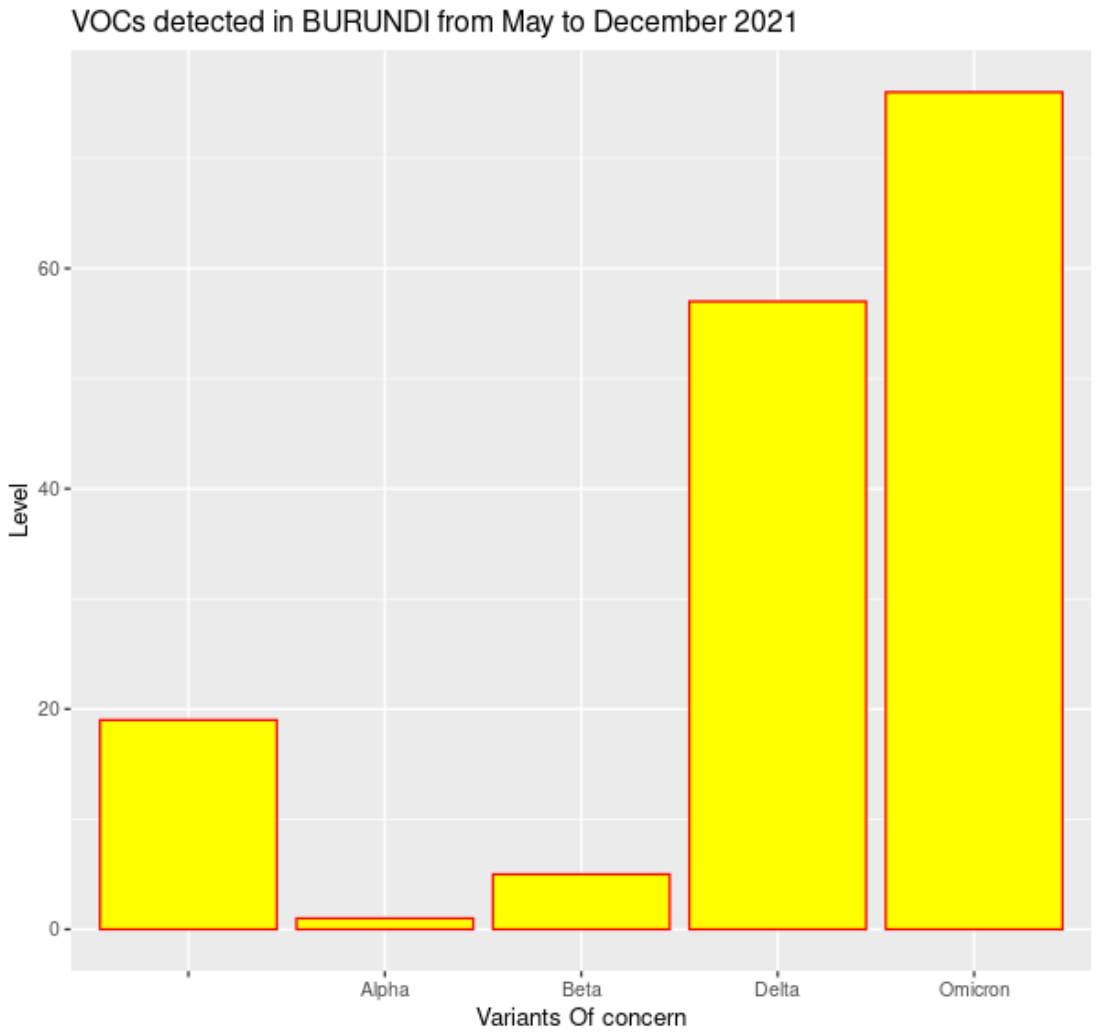

Figure 5.1: Variants of concern detected in BURUNDI during the periode from May to December 2021

Two different peaks of COVID-19 cases – from May to September 2021 and November to December 2021 respectively – predominated by different SARS-COV-2 lineages were observed during this study period (Figure 5.1). The SARS-COV-2 lineage B.1.617.2 (Delta) was predominant during the wave that occurred from May to September 2021 and accounted for 36.07% (n = 57) of all the SARS-COV-2 genomes isolated in Burundi during that period. On the other hand, the peak of cases observed from November to December 2021 was predominated by the lineage B.1.1.529 (Omicron) and its sublineages BA.1 and BA.1.1, accounting for 48.1% of all the SARS-COV-2 genomes identified in Burundi during that period (Figure 5.1). Clearly, the Omicron variant tended to replace the Delta (B.1.617.2) lineage which represented only 1% of all the SARS-COV-2 lineages identified in November and December 2021. The Omicron lineage (B.1.1.529) and its sublineages BA.1 and BA.1.1 were first identified in Burundi from samples collected on November 24, 2021.

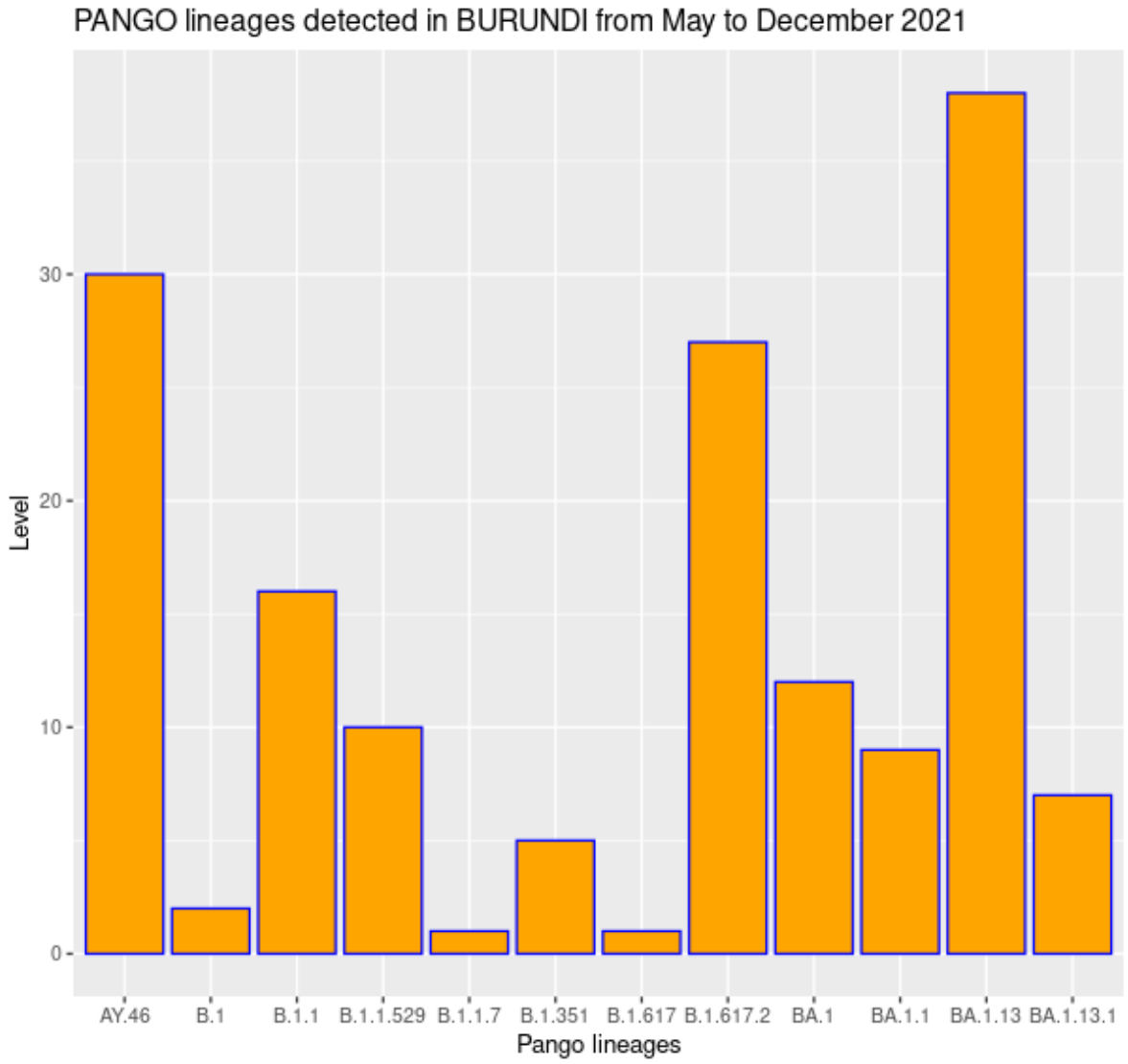

Figure 5.2: *Distribution of the SARS-COV-2 lineage composition of different COVID-19 waves. B.1.1.7 = Alpha variant, B.1.351 = Beta variant, B.1.617.2 = Delta variant, B.1.1.529 = Omicron variant, BA.1,BA.1.13,BA.1.13.1 and BA.1.1 are Omicron sublineages, AY.46 is Delta sublineage*

5.1.2 Evolution of SARS-CoV-2 in Burundi

In this subsection, we present the results on the evolution of the virus in Burundi. Overall, compared to the reference genome, we observed a total of four thousand nine hundred and eighty-three (4983) mutations across all of our samples with an interval of seventeen (17) to forty-eight (48), or an average of 31.53 mutations/genome, or one mutation every 948.3 bp on the genome (Table 5.3).

Tables 5.3: *Polymorphisms in the SARS-CoV-2 genome*

Mutations	Values
Interval	17-48
Average	31.53
Density mutations	943.3 pb
Total	4983

The SARS-CoV-2 genome can be divided into three large parts (ORF1a, ORF1b and the structural proteins joined together by accessory proteins). Mutations in each essential part of the genome can impact the infectivity of the virus and impact strategies to combat the pandemic (antiviral drugs, vaccines, diagnostic tests and neutralizing antibodies).

5.1.3 Notable mutations detected

Some mutations are known to be associated with a change in characteristics in the virus. This makes control strategies less effective. These mutations have been named notable non specific mutations. The most common mutations carried by the different SARS-COV2 variants identified in Burundi mainly concerned the surface glycoprotein, N, and different non-structural proteins(Figure 5.3). The D614G mutation in the spike protein was shared by almost all SARS-COV-2 genomes identified in the present study, regardless of lineage. Additionally, all Omicron genome sequences (B.1.1.529, BA.1 and BA.1.1) displayed mutations D796Y, N764K, N856K, T547K and N696K. The G142D mutation was rather shared with the Delta variant (B.1.617.2). The following mutations have been identified in the Delta variant spike protein: A22V, L452R, D950N, E156G, G142D, P681R, S698L and T19R. Concerning the Alpha and Beta variants, their surface glycoprotein harbored the following mutations: D614G, N501Y, D215G, E484K, K417N, A701V, P681H, D1118H, S982A, L18F, T716I and L54F.

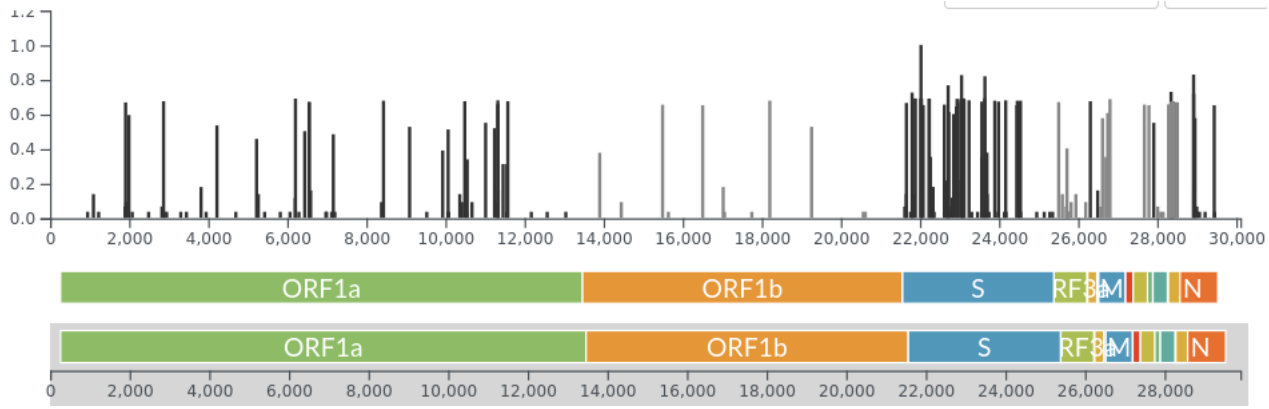

Figure 5.3: Distribution of mutations carried by variants identified in Burundi on the SARS-COV-2 genome

5.1.4 Phylogenetic analysis

The phylogenetic analysis of the sequences from SARS-COV-2 lineages identified in Burundi showed that B.1.1.7 and B.1.351 lineages clusters did not expand largely through community transmission. The B.1.1.7 lineage from this study is represented by one genome detected from a sample collected on May 27, 2021, in the municipality of Bujumbura. Despite the relatively small proportion of the B.1.351 lineage detected in this study, the 5 corresponding sequences (indicated in green on Figure 5.4) formed two phylogenetically separate clusters, suggesting two different introductions events. They were all generated from COVID-19 samples collected at different COVID-19 routine testing sites located in the municipality of Bujumbura, during the period from May 27 to May 31, 2021. The SARS-COV-2 genomes detected in Burundi during that period were mostly dominated by the B.1.1.529 lineage.

The Delta variant (highlighted in yellow) was first detected in Burundi from samples collected on May 26, 2021, during a campaign of mass COVID-19 testing(Figure 5.4). The sequences of Delta lineages from this study formed separate clusters based on different sublineages.
Furthermore, the sequences of Delta genomes identified in this study formed a separate cluster, suggesting a different introduction of that variant in Burundi.

However, a close genetic relatedness was observed between the sequences from the Delta genome identified in the municipality of Bujumbura and other provinces of Burundi.
Phylogeny reconstruction demonstrated that the COVID-19 cases observed in municipality of Bujumbura during the first wave as of May to September 2021 were interlinked and resulted in the expansion of initial cases, suggesting a community transmission of the pandemic in the country.

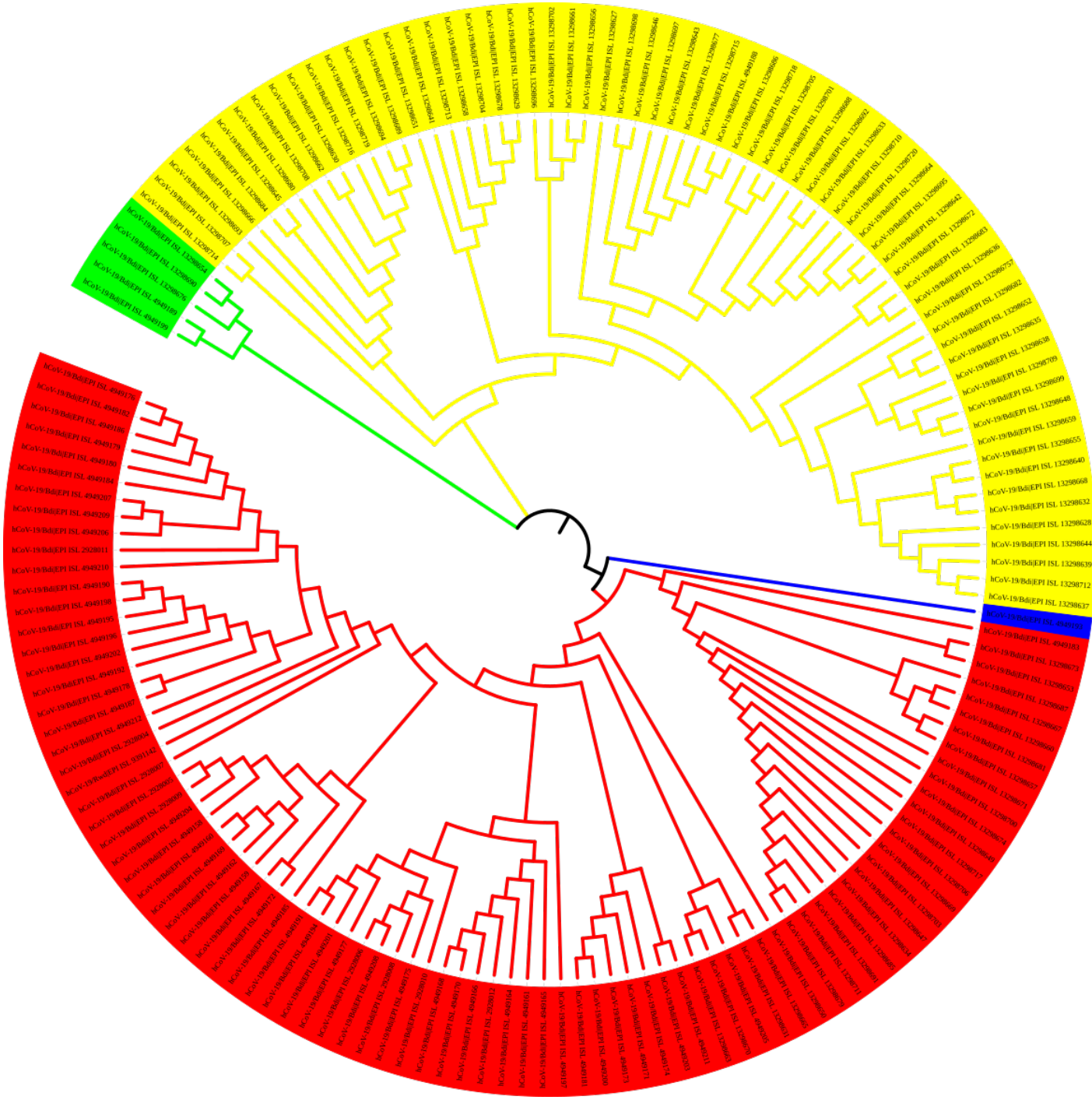

Figure 5.4: *The Neighbour-Joining (NJ) phylogenetic tree of the nucleotide sequences of SARS-COV-2 genomes detected in Burundi in 2021. The B.1.351 genomes (n = 5) are indicated in green, Delta (B.1.617.2) and its sublineages (n = 66) in yellow, the Omicron (B.1.1.529) and its sublineages BA.1 and BA.1.1 genomes (n = 86) in red, and the only one B.1.1.7 genome is in blue.*

Chapter 6

Discussion

The present study aimed at investigating the genetic diversity, evolution, and molecular epidemiology of SARS-COV-2 variants in Burundi. It allowed to demon- strate the great genetic diversity of SARS-COV-2 line- ages that circulated in Burundi during the period from May 2021 to December 2021, with 12 lineages in total, of which 5 were VOCs. Furthermore, the findings from this study demonstrated that different peaks of COVID-19 cases that followed phases of quasi-controlled COVID-19 pandemic in Burundi involved different SARS-COV-2 variants; suggesting multiple introductions of different SARS-COV-2 lineages from other countries. This is the case of the peak of COVID-19 cases observed from May to October September 2021 and the peak associated with the Omicron variant of December 2021.

The introduction of B.1.617.2 (Delta) in Burundi – first detected in the country in May 2021 – and the subsequent resurgence of COVID-19 cases accompanied the lifting of movement restrictions and the reduction in quarantine duration for international travelers from 7 to 4 days in April 2021. Similar epidemiological trends after the relaxation of travel restrictions have been described else- where [61]. This lineage carried amino acid mutations such as T19R, and G142D in the N-Terminal Domain (NTD) known to be a point of recognition for vaccines and the immune response [62]. Moreover, other mutations such as E484K, D614G, and L452R known to have a substantial impact on the virus infectivity were identified in the Receptor-Binding Domain (RBD) of Delta genome sequences [63].

In addition, the present study demonstrated how the introduction of new SARS-CoV-2 lineages challenged the public health interventions and social measures put in place to tackle the COVID-19 pandemic in Burundi. Similar impacts of the emergence of new variants were reported elsewhere [64]. The Omicron variant B.1.1.529 was associated with the greatest number of COVID-19 cases in Burundi and tended to replace the Delta variant which predominated in previous peaks. A similar situation was reported elsewhere and was suggested to be associated with immune escape from this variant rather than intrinsic higher transmissibility. The Omicron variant genomes harbor several mutations in the spike protein which increased its ability to escape the immune response and the infectivity [65]. Some of these mutations were shared with Delta and included D614D,and G142D in the spike protein (S).

The phylogenetic analysis of SARS-COV-2 genomes offered an opportunity to visualize the clustering patterns of different SARS-COV-2 genomes identified in Burundi and the evolutionary relationships linking the SARS-COV-2 genomes detected in the community and those imported from other countries. In this regard, B.1.617.2 (Delta) and its sub-lineage AY.46 on one hand, and B.1.1.529 (Omicron) and its sublineage BA.1 on the other hand formed several separate clusters suggesting multiple introductions in Burundi. However, a close evolutionary relationship was observed between locally detected and imported genomes from travelers and national citizens returning from the Democratic Republic of Congo, Rwanda, Uganda, and Canada among others. This suggests that this variant was imported from other countries that had experienced it before Burundi.

The Delta variant was first reported in India in March 2021 from December 2020 samples [66]. This lineage was first detected in Burundi from samples collected in the municipality of Bujumbura as of May 27, 2021. Further cases associated with B.1.617.2 VOC and its sublineage AY.46 were subsequently detected in other provinces of Burundi as a result of inter-province transmission, as suggested by the evolutionary relationships documented through the phylogeny reconstruction.

The B.1.617.2 (Delta) lineage was gradually replaced by Omicron (B.1.1. 529) starting in December 2021. Omicron caused the highest number of COVID-19 cases ever observed in Burundi since the initial phase of that pandemic. That trend was consistent with the anticipations made by public health experts about a rapid expansion of COVID-19 cases associated with the Omicron variant, based on the genomic properties of the virus [67].

As a complement, the phylogenetic analysis allowed us to visualize cases of community transmission of SARS-COV-2 lineages and the involvement of imported SARS-COV-2 lineages.However, this situation needs to be interpreted cautiously, especially outside the municipality of Bujumbura as insufficient sampling in those provinces might mask possible inter-province transmission events and separate introductions of SARS-COV-2.

Conclusion

This study contributed at investigating the genetic diversity, evolution, and molecular epidemiology of SARS-COV-2 variants in Burundi. It allowed to demonstrate the genetic diversity of SARS-COV-2 lineages during the period from May 2021 to December 2021, with 12 lineages in total, of which 5 were VOCs. Furthermore, the close evolutionary relationships between imported and community-isolated SARS-COV-2 lineages in Burundi imply COVID-19 transmission between Burundi and other countries. We demonstrated that new peaks of COVID-19 cases were mainly associated with newly introduced SARS-COV-2 VOCs. Furthermore, the present study highlighted the implications of travel restrictions relaxation and the virus mutations on the introduction and the subsequent spread of SARS-COV-2 variants in Burundi.

Recommandations

In order to fight against the SARS-CoV-2, we recommand to the health authorities and the government to strengthen the epidemiological monitoring on SARS-COV-2, enhance the protection by increasing the SARS-COV-2 vaccine coverage, and support the competent laboratories in reagents, devices and consumables so that they can participate in the epidemiological assessment. It is also important for the researchers to continue the research on SARS-COV-2 for a better understanding of the disease.

Bibliography

[1] Jatin Machhi, Jonathan Herskovitz, Ahmed M Senan, Debashis Dutta, Barnali Nath, Maxim D Oleynikov, Wilson R Blomberg, Douglas D Meigs, Mahmudul Hasan, Milankumar Patel, et al. The natural history, pathobiology, and clinical manifestations of sars-cov-2 infections. *Journal of Neuroimmune Pharmacology*, 15:359–386, 2020.

[2] Ariane Bonnin. *Caractérisation de la protéine S du coronavirus humain 229E*. PhD thesis, Université de Lille, 2018.

[3] Dhama Kuldeep, Sharun Khan, R Tiwari, S Sircar, S Bhat, YS Malik, KP Singh, and W Chalcumpa. Update on covid-19. *Clinical Microbiology Reviews*, 33(4):1–48, 2020.

[4] Organisation mondiale de la Santé. Test de diagnostic de la covid-19 dans le contexte des voyages internationaux: document d'information scientifique, 16 décembre 2023. Technical report, Organisation mondiale de la Santé, 2023.

[5] François Chesnais. L'état de l'économie mondiale au début de la grande récession covid-19: repères historiques, analyses et illustrations. *A l'encontre*, 2020.

[6] Sara Mbago-Bhunu, Habte-Selassie Dagmawi, and Deirdre Mc Grenra. Republic of burundi country strategic opportunities programme. 2022.

[7] Dinah V Parums. Revised world health organization (who) terminology for variants of concern and variants of interest of sars-cov-2. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 27:e933622–1, 2021.

[8] World Health Organization et al. Covid-19 weekly epidemiological update, edition 134, 17 december 2023. 2023.

[9] Swatantra Kumar, Rajni Nyodu, Vimal K Maurya, and Shailendra K Saxena. Morphology, genome organization, replication, and pathogenesis of severe acute respiratory syndrome coronavirus 2 (sars-cov-2). In *Coronavirus Disease 2019 (COVID-19)*, pages 23–31. Springer, 2020.

[10] Rajanish Giri, Taniya Bhardwaj, Meenakshi Shegane, Bhuvaneshwari R Gehi, Prateek Kumar, Kundlik Gadhave, Christopher J Oldfield, and Vladimir N Uversky. Understanding covid-19 via comparative analysis of dark proteomes of sars-cov-2,

human sars and bat sars-like coronaviruses. *Cellular and Molecular Life Sciences*, 78:1655–1688, 2021.

[11] Imane Jamai Amir, Zina Lebar, Mustapha Mahmoud, et al. Covid-19: virologie, épidémiologie et diagnostic biologique. *Option/Bio*, 31(619):15, 2020.

[12] Nicholas A Wong and Milton H Saier Jr. The sars-coronavirus infection cycle: a survey of viral membrane proteins, their functional interactions and pathogenesis. *International journal of molecular sciences*, 22(3):1308, 2021.

[13] Muhammad Adnan Shereen, Suliman Khan, Abeer Kazmi, Nadia Bashir, and Rabeea Siddique. Covid-19 infection: Emergence, transmission, and characteristics of human coronaviruses. *Journal of advanced research*, 24:91–98, 2020.

[14] Sébastien Hantz. Diagnostic biologique de l'infection à sars-cov-2: stratégies et interprétation des résultats. *Revue Francophone Des Laboratoires*, 2020(526):48–56, 2020.

[15] Organisation mondiale de la Santé. Test de diagnostic de la covid-19 dans le contexte des voyages internationaux: document d'information scientifique, 16 décembre 2020. Technical report, Organisation mondiale de la Santé, 2020.

[16] Jean Luc Gala, Omar Nyabi, Jean-François Durant, Nawfal Chibani, and Mostafa Bentahir. Méthodes diagnostiques du covid-19. *Louvain Med*, 139(05-06):228–235, 2020.

[17] A Padane. Evolution of variants of concern of sars-cov-2 during three waves in senegal. *Journal of Public Health in Africa*, pages 15–16, 2022.

[18] Grant Murewanhema, Faith Mutsigiri-Murewanhema, and Edward Kunonga. Enhancing sars-cov-2 surveillance through regular genomic sequencing is an essential element of covid-19 control in resource-limited settings. *Pan African Medical Journal*, 41(1), 2022.

[19] Frederic Raymond. *bio-informatique pour la génomique et le diagnostic des maladies infectieuses*. PhD thesis, Université Laval, 2011.

[20] Hanane Bahouq, Madiha Bahouq, and Abdelmajid Soulaymani. Overview of genomic surveillance related to severe acute respiratory syndrom coronavirus 2 (sars-cov-2). In *E3S Web of Conferences*, volume 319, page 01043. EDP Sciences, 2021.

[21] Organisation mondiale de la Santé. Séquençage génomique du sars-cov-2 à des fins de santé publique: orientations provisoires, 8 janvier 2021. Technical report, Organisation mondiale de la Santé, 2021.

[22] M Ahakoud. Le sequençage d'acide désoxyribonucléique: prinicpe téchnique, indications medicales et experience du chu hassane ii de fés. *Univ. SIDI MOHAMMED BEN ABDELLAH, 159p*, 2015.

[23] Elmostafa EL FAHIME and Mly Mustapha ENNAJI. Évolution des techniques de séquençage. *Les technologies de laboratoire*, 2(5), 2007.

[24] Pierre Morisse. *Correction de données de séquençage de troisième génération*. PhD thesis, Normandie Université, 2019.

[25] Comité sur l'immunisation du Québec. Utilisation du vaccin astrazeneca contre la covid-19 dans le contexte du signal de thromboses avec thrombocytopénie suite à la vaccination. 2021.

[26] France Info. Covid-19: le variant anglais détecté dans 68% des tests positifs à dunkerque, 2021.

[27] Talha Burki. Understanding variants of sars-cov-2. *The Lancet*, 397(10273):462, 2021.

[28] Wilfredo F Garcia-Beltran, Kerri J St Denis, Angelique Hoelzemer, Evan C Lam, Adam D Nitido, Maegan L Sheehan, Cristhian Berrios, Onosereme Ofoman, Christina C Chang, Blake M Hauser, et al. mrna-based covid-19 vaccine boosters induce neutralizing immunity against sars-cov-2 omicron variant. *Cell*, 185(3):457–466, 2022.

[29] Rita Feghali, Georgi Merhi, Aurelia Kwasiborski, Veronique Hourdel, Nada Ghosn, and Sima Tokajian. Genomic characterization and phylogenetic analysis of the first sars-cov-2 variants introduced in lebanon. *PeerJ*, 9:e11015, 2021.

[30] Mhamad Abou-Hamdan, Kassem Hamze, Ali Abdel Sater, Haidar Akl, Nabil El-Zein, Israa Dandache, and Fadi Abdel-Sater. Variant analysis of the first lebanese sars-cov-2 isolates. *Genomics*, 113(1):892–895, 2021.

[31] Adnan Aziz, Kumud Sanwal, Vigyan Singhal, and Robert Brayton. Verifying continuous time markov chains. In *Computer Aided Verification: 8th International Conference, CAV'96 New Brunswick, NJ, USA, July 31–August 3, 1996 Proceedings 8*, pages 269–276. Springer, 1996.

[32] Jonathan P Bollback. Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, 19(7):1171–1180, 2002.

[33] Sudhir Kumar, Koichiro Tamura, and Masatoshi Nei. Mega: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics*, 10(2):189–191, 1994.

[34] David L Swofford. Phylogenetic analysis using parsimony. 1998.

[35] Bernard R Baum. Phylip: phylogeny inference package. version 3.2, 1989.

[36] Stéphane Guindon, Frédéric Delsuc, Jean-François Dufayard, and Olivier Gascuel. Estimating maximum likelihood phylogenies with phyml. *Bioinformatics for DNA sequence analysis*, pages 113–137, 2009.

[37] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[38] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16:111–120, 1980.

[39] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17:368–376, 1981.

[40] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *Journal of molecular evolution*, 22:160–174, 1985.

[41] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526, 1993.

[42] Simon Tavaré. Some probabilistic and statistical problems on the analysis of dna sequence. *Lecture of Mathematics for Life Science*, 17:57, 1986.

[43] Jeremy G Sumner, Peter D Jarvis, Jesús Fernández-Sánchez, Bodie T Kaine, Michael D Woodhams, and Barbara R Holland. Is the general time-reversible model bad for molecular phylogenetics? *Systematic biology*, 61(6):1069–1074, 2012.

[44] David Posada, Keith A Crandall, and David M Hillis. Phylogenetics of hiv. In *Computational and Evolutionary Analysis of HIV Molecular Sequences*, pages 121–160. Springer, 2000.

[45] D Anderson and K Burnham. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63(2020):10, 2004.

[46] Korbinian Strimmer, Arndt von Haeseler, Anne-Mieke Salemi, et al. Nucleotide substitution models. *The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny*, pages 72–100, 2003.

[47] Lars S Jermiin, Vivek Jayaswal, Faisal M Ababneh, and John Robinson. Identifying optimal models of evolution. *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution*, pages 379–420, 2017.

[48] Ziheng Yang. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal of Molecular evolution*, 39:306–314, 1994.

[49] James E Allen and Simon Whelan. Assessing the state of substitution models describing noncoding rna evolution. *Genome biology and evolution*, 6(1):65–75, 2014.

[50] Shiran Abadi, Dana Azouri, Tal Pupko, and Itay Mayrose. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature communications*, 10(1):934, 2019.

[51] María José Domper Arnal, Ángel Ferrández Arenas, and Ángel Lanas Arbeloa. Esophageal cancer: Risk factors, screening and endoscopic treatment in western and eastern countries. *World journal of gastroenterology: WJG*, 21(26):7933, 2015.

[52] José Ramón Pardos-Blas, Iker Irisarri, Samuel Abalde, Carlos ML Afonso, Manuel J Tenorio, and Rafael Zardoya. The genome of the venomous snail lautoconus ventricosus sheds light on the origin of conotoxin diversity. *Gigascience*, 10(5):giab037, 2021.

[53] Korbinian Strimmer, Arndt von Haeseler, and Marco Salemi. Genetic distances and nucleotide substitution models. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*, pages 111–141, 2009.

[54] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, et al. Gisaid's role in pandemic response. *China CDC weekly*, 3(49):1049, 2021.

[55] Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus evolution*, 7(2):veab064, 2021.

[56] Ivan Aksamentov, Cornelius Roemer, Emma B Hodcroft, and Richard A Neher. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of open source software*, 6(67):3773, 2021.

[57] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[58] Koichiro Tamura, Glen Stecher, and Sudhir Kumar. Mega11: molecular evolutionary genetics analysis version 11. *Molecular biology and evolution*, 38(7):3022–3027, 2021.

[59] Bui Quang Minh, Minh Anh Thi Nguyen, and Arndt Von Haeseler. Ultrafast approximation for phylogenetic bootstrap. *Molecular biology and evolution*, 30(5):1188–1195, 2013.

[60] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*, 49(W1):W293–W296, 2021.

[61] Idowu B Olawoye, Paul E Oluniyi, Judith U Oguzie, Jessica N Uwanibe, Tolulope A Kayode, Testimony J Olumade, Fehintola V Ajogbasile, Edyth Parker, Philomena E Eromon, Priscilla Abechi, et al. Emergence and spread of two sars-cov-2 variants of interest in nigeria. *Nature communications*, 14(1):811, 2023.

[62] Abeer Asif, Iqra Ilyas, Mohammad Abdullah, Sadaf Sarfraz, Muhammad Mustafa, and Arif Mahmood. The comparison of mutational progression in sars-cov-2: A short updated overview. *Journal of Molecular Pathology*, 3(4):201–218, 2022.

[63] Bette Korber, Will M Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, Elena E Giorgi, Tanmoy Bhattacharya, Brian Foley, et al. Tracking changes in sars-cov-2 spike: evidence that d614g increases infectivity of the covid-19 virus. *Cell*, 182(4):812–827, 2020.

[64] Anise N Happi, Chinedu A Ugwu, and Christian T Happi. Tracking the emergence of new sars-cov-2 variants in south africa. *Nature Medicine*, 27(3):372–373, 2021.

[65] Sarah Cherian, Varsha Potdar, Santosh Jadhav, Pragya Yadav, Nivedita Gupta, Mousumi Das, Partha Rakshit, Sujeet Singh, Priya Abraham, Samiran Panda, et al. Sars-cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of covid-19 in maharashtra, india. *Microorganisms*, 9(7):1542, 2021.

[66] Alfredo Parra-Lucares, Paula Segura, Verónica Rojas, Catalina Pumarino, Gustavo Saint-Pierre, and Luis Toro. Emergence of sars-cov-2 variants in the world: how could this happen? *Life*, 12(2):194, 2022.

[67] Taha Menasria and Margarita Aguilera. Genomic diversity of sars-cov-2 in algeria and north african countries: what we know so far and what we expect? *Microorganisms*, 10(2):467, 2022.

[68] Mart M Lamers and Bart L Haagmans. Sars-cov-2 pathogenesis. *Nature reviews microbiology*, 20(5):270–284, 2022.

[69] Di Wu, Tiantian Wu, Qun Liu, and Zhicong Yang. The sars-cov-2 outbreak: what we know. *International journal of infectious diseases*, 94:44–48, 2020.

[70] Ben Hu, Hua Guo, Peng Zhou, and Zheng-Li Shi. Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, 19(3):141–154, 2021.

[71] Allen G Rodrigo and Gerald H Learn. *Computational and evolutionary analysis of HIV molecular sequences*. Springer Science & Business Media, 2001.

[72] Cassien Nduwimana, Néhémie Nzoyikorera, Armstrong Ndihokubwayo, Théogène Ihorimbere, Célestin Nibogora, Adolphe Ndoreraho, Oscar Hajayandi, Jean Claude Bizimana, Idrissa Diawara, Dionis Niyonizigiye, et al. Genomic surveillance of severe acute respiratory syndrome coronavirus 2 in burundi, from may 2021 to january 2022. *BMC genomics*, 24(1):1–10, 2023.

[73] Organisation mondiale de la Santé. Recherche des contacts et placement en quarantaine dans le contexte du variant omicron du sars-cov-2: orientations provisoires, 17 février 2022. Technical report, Organisation mondiale de la Santé, 2022.

[74] Pettersson Nyrén, Bertil Pettersson, and Mathias Uhlén. Solid phase dna minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical biochemistry*, 208(1):171–175, 1993.

[75] WHO Covid. Dashboard. 2020. *Coronavirus Disease (COVID-19)–World Health Organization [Internet] https://www. who. int/emergencies/diseases/novel-coronavirus-2019 [cited 29 August 2020]. Available from:[Google Scholar]*, 19.

[76] Roger Tsafack Nanfosso and Juliana Hadjitchoneva. La théorie économique face à la covid-19: de joseph schumpeter à robert solow.

[77] Bourema Kouriba, Angela Dürr, Alexandra Rehn, Abdoul Karim Sangaré, Brehima Y Traoré, Malena S Bestehorn-Willmann, Judicael Ouedraogo, Asli Heitzer, Elisabeth Sogodogo, Abderrhamane Maiga, et al. First phylogenetic analysis of malian sars-cov-2 sequences provides molecular insights into the genomic diversity of the sahel region. *Viruses*, 12(11):1251, 2020.

[78] Ibrahim M Hezam. Covid-19 global humanitarian response plan: An optimal distribution model for high-priority countries. *ISA transactions*, 124:1–20, 2022.

[79] Ivica Letunic and Peer Bork. Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic acids research*, 47(W1):W256–W259, 2019.

[80] Sanjeet Bagcchi. Covid-19 and measles: double trouble for burundi. *The Lancet Microbe*, 1(2):e65, 2020.

[81] World Health Organization et al. Report on the strategic response to covid-19 in the who african region–1 february 2021 to 31 january 2022. 2022.

[82] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[83] Organisation mondiale de la Santé. Test de diagnostic de la covid-19 dans le contexte des voyages internationaux: document d'information scientifique, 16 décembre 2022. Technical report, Organisation mondiale de la Santé, 2022.