Morocco

Mohammed Vth University

Faculty of Medicine and Pharmacy of
Rabat

YEAR: 2021 **THESIS N°: 11/21/CSVS**

Doctoral Centre of Life Sciences and Health

Department: Medical biology, Human Pathology,

Experimental and Environment

Option: Bioinformatics

DOCTORAL THESIS

Mycobacterium tuberculosis and SARS-CoV-2 genomic analysis: input into outbreaks and surveillance investigations

Presented and publicly defended by:

Mariem LAAMARTI

June 18th, 2021

COMMITEE MEMBERS

Pr. Abdallah BADOU Faculty of Medicine and Pharmacy, Hassan II University, Casablanca	**Chair**
Pr. Azeddine IBRAHIMI Faculty of Medicine and Pharmacy, Mohammed Vth University, Rabat	**Advisor**
Pr. Samir SIAH Faculty of Medicine and Pharmacy, Mohammed Vth University, Rabat	**Advisor**
Pr. Rachid ELJAOUDI Faculty of Medicine and Pharmacy, Mohammed Vth University, Rabat	**Reporter**
Pr. Laila SBABOU Faculty of sciences, Mohammed Vth University, Rabat	**Reporter**
Pr. Mohammed EL AZAMI EL IDRISSI Faculty of Medicine and Pharmacy, Université Sidi Mohamed Ben Abdellah, Fes	**Reporter**
Pr. Lahcen BELYAMANI Faculty of Medicine and Pharmacy, Mohammed Vth University, Rabat	**Examiner**
Pr. Mouna OUADGHIRI Faculty of Medicine and Pharmacy, Mohammed Vth University, Rabat	**Examiner**

Declaration

I, **Laamarti Mariem**, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

The thesis work was conducted from September 2015 to March 2020 at the BioMedical Labo-ratory, Medecine and Pharmacy School Rabat, Morrocco, under the supervision of Pr. **Azeddine Ibrahimi** (Medbiotech, Rabat), and **Pr. Samir Siah** (HMIMV).

Date

Signed

Acknowledgement

First and foremost, I would like to praise and thank God, the almighty, whose many blessing have made me who I am today.

Many people have contributed in this thesis and supported me during its development. Herewith I would like to take this opportunity to thank them all.

First of all, I wish to thank each committee member for their assistance and agreeing to evaluate my work. I would particularly like to thank the three reporters who had to read this manuscript and write their report during the month of Ramadan.

I would like to express my deep sense of gratitude to **Pr. Azeddine Ibrahimi**, my supervisor, for all your wise advice and for trusting me entirely during these 6 years !! I also thank you for your inexhaustible optimism and guidance through every stage of the research. I thank **Pr. Samir Siah**, who accepted me as Ph.D student and for his support and valuable discussions.

I wish to express my sincere appreciation to **Pr. Mona Ouadghiri** and **Pr. Tarek Aaniz** for their pieces of advice and the many thrilling discussions.

I reserve my deepest appreciation for all my colleagues who have helped me during my Ph.D - **Lamiaa Lahlou, Tarek ALouane, Souad Kartti, Walid chemao, Mohammed Hakmi, Abdelmounim Saber, Nasma Boumajid, Houda Bendani, Oussama Benhrif and Kaoutar Karfi**. You allowed me to challenge myself on different levels during these last 6 years. Thank you for all the advice and good Birkbeck times.

I would like to acknowledge the support of **Pr. Elouness Moustapha** within the HMIMV, especially for his academic input and constructive criticism. In the same terms, I acknowledge the contribution of all our collaborators.

My deepest gratitude goes to my family for all the support they have given me on the road to get where I am today - My mother **Naima Benyelloul**, aunts **Khadija & Rkia Laamarti** and father **Mohamed Laamarti**. I'm very grateful for my siblings, the guys who have been there since the start of my research Journey, **Mohamed Laamarti** and **Jaouad Laamarti**, for always being there for me throughout my Ph.D. Immense love and gratitude go to my sister **Rokia Laamarti** for funding my PhD journey and for the great model you have always been to me.

I am also thankful to all those who directly or indirectly helped me during this entire work.

Abstract

Title: *Mycobacterium Tuberculosis* and SARS-CoV-2 genomic analysis: input into outbreaks and surveillance investigations
Author: Laamarti Mariem

Keywords: *Mycobacterium Tuberculosis*, SARS-Cov-2, Genomic analysis, Phylogeny, drug resistance, phylodynamic, sequencing.

Comparative microbial genomics is increasingly used for high-resolution epidemiological investigation of infectious agents' sources, transmission dynamics and antimicrobial resistance.

In **Chapter II**, We performed the sequencing and genomic characterization of *M. Tuberculosis* strains from Morocco to get insight into their genomic diversity, drug resistance, population structure and identify potential mutations associated with drug resistance. We conducted a whole-genome analysis of nine Morrocan *M. tuberculosis* isolates; we identified 25 known mutations and 14 novel mutations in drug-associated genes and provided experimental support for them. We found that all resistance and susceptible strains clustered with LAM9 and Haarlem, respectively, belonging to the Euro-American clade. The modelling of GyrA/GyrB mutations showed a decrease in the binding affinity with levofloxacin.

Chapter III addresses the comparative genomic of SARS-CoV-2 from Morocco to identify genetic variants as a crucial step in evaluating the spread in Morocco. This study revealed 108 mutations in their genomes. The analysis haplotype network suggests different sources of SARS-CoV-2 infection in Morocco.

In **Chapter IV**, we collected SARS-CoV-2 genomes isolated from 80 countries.The results showed genotypes specific to geographic location. Moreover, evolution over time has demonstrated a mechanism of mutation co-accumulation, which might affect the severity and spread of the SARS-CoV-2 suggesting that a universal vaccine is more likely to be efficient for all strains.

On the other hand, the selective pressure analysis revealed negatively selected residues that could be considered therapeutic targets. We have also created an inclusive unified database that lists all of the genetic variants of the SARS-CoV-2 genomes found in this study.

Résumé

Title: Analyse génomique de Mycobacterium tuberculosis et du SRAS-CoV-2: contribution aux épidémies et aux enquêtes de surveillance
Name: Laamarti Mariem
Mot-clés: *Mycobacterium Tuberculosis*, SARS-Cov-2, Analyse génomique, Phylogénie, résistance aux antibiotique, phylodynamique, séquençage

La génomique microbienne comparative permet l'investigation épidémiologique des agents infectieux, afin de cerner leur origine, leur dynamique de transmission, ainsi que larésistance aux antibiotiques.

Chapter II, Nous avons effectué le séquençage et la caractérisation génomique des souches de *M. Tuberculosis* du Maroc pour avoir un aperçu de leur diversité génomique, la résistance aux antibiotiques, la structure de la population et identifier les mutations associées à la résistance aux médicaments. Nous avons identifié 25 mutations connues et 14 nouvelles mutations dans les gènes associés à la résistance au antibiotiques et nous leur avons fourni un soutien expérimental. Nous avons constaté que toutes les souches résistantes et sensibles étaient regroupées avec LAM9 et Haarlem, respectivement, appartenant au clade euro-américain. La modélisation des mutations GyrA/GyrB a montré une diminution de son affinité à la lévofloxacine.

Le **Chapter III**, aborde la génomique comparative du SARS-CoV-2 au Maroc pour identifier les variantes génétiques. Cette étude a révélé 108 mutations dans le génome des virus. Le réseau d'haplotypes suggère différentes sources d'infection au Maroc.

Dans le **Chapter IV**, nous avons collecté des génomes du SARS-CoV-2 isolés dans 80 pays. Les résultats ont montré des génotypes spécifiques à l'emplacement géographique. De plus, l'évolution au fil du temps a mis en évidence un mécanisme de co-accumulation de mutations, qui pourrait affecter la propagation du SARS-CoV-2. suggérant que qu'un vaccin universel est plus susceptible d'être efficace pour toutes les souches
D'autre part, l'analyse de pression sélective a révélé des résidus sélectionnés négativement qui pourraient être considérés comme des cibles thérapeutiques. Nous avons également créé une base de données qui répertorie toutes les variantes génétiques des génomes du SARS-CoV-2 trouvés dans cette étude.

ملخص

العنوان: تحليل الجينوم من المتفطرة السلي SARS-CoV-2 : المدخلات في تفشي المرض و تتبعه

الاسم :لعمارتي مريم

الكلمات المفتاحية: المتفطرة السلية ، SARS-CoV-2 ، التحليل الجينومي ، النشوء والتطور ، المقاومة ، الديناميكا النباتية ، التسلسل.

تستخدم مقارنة الجينومات بشكل متزايد في الاستقصاء الوبائي للعوامل المعدية، من أجل تحديد أصلها، وديناميكيات انتقالها، وكذلك استراتيجيات مقاومتها للمضادات الحيوية.

في **الفصل ٢**، أجرينا التسلسل والتوصيف الجيني للسلالات المتفطرة السلية من المغرب، لاكتساب نظرة شاملة حول تنوعها الجيني، مقاومتها للمضادات الحيوية، التركيبة النسبية وايضا حول الطفرات المرتبطة بمقاومة الأدوية. لذلك أجرينا تحليل جينوم كامل لتسع عزلات مغربية من المتفطرة السلية، تمكنا من خلاله تحديد 25 طفرة معروفة و 14 طفرة جديدة في الجينات المرتبطة بمقاومة المضادات الحيوية، و دعمنا هذا تحليل بدراسة النمط الظاهري لهذه العزلات.

وجدنا أيضا أن جميع السلالات المقاومة والحساسة تم تجميعها مع LAM9 وHaarlem، على التوالي، والتي تنتمي إلى الفرع الأوروبي الأمريكي. بينما أظهرت نمذجة طفرات في المركب GyrA\GyrB انخفاضًا في تقارب هذا الأخير مع الليفوفلوكساسين.

يناقش **الفصل ٣** مقارنة جينومية لـ SARS-CoV-2 في المغرب لتحديد المتغيرات الجينية، كخطوة حاسمة في تقييم انتشار الفيروس. وقد كشفت هذه الدراسة عن 108 طفرة في جينوم الفيروسات المدروسة بينما اقترحت شبكة النمط الفرداني مصادر مختلفة لعدوى SARS-CoV-2 في المغرب.

في **الفصل ٤**، قمنا بجمع جينومات SARS-CoV-2 المعزولة من 80 دولة. أظهرت نتائج هذه دراسة أنماط وراثية خاصة بالموقع الجغرافي. بالإضافة إلى ذلك، أظهر التطور بمرور الوقت آلية التراكم المشترك للطفرات، والتي يمكن أن تؤثر على شدة وانتشار الڤيروس. من ناحية أخرى، كشف تحليل الضغط الانتقائي عن مواضع جينية مختارة سلبًا يمكن اعتبارها أهدافًا علاجية. كما تشير هذه النتائج إلى أن لقاح عالمي من المرجح أن يكون فعال لجميع السلالات

. في النهاية، أنشأنا أيضًا قاعدة بيانات شاملة موحدة والتي تسرد جميع المتغيرات الجينية لجينومات SARS-CoV-2الموجودة في هذه الدراسة.

Acronyms

ACE2	Angiotensin-converting enzyme 2
AIDS	acquired immunodeficiency syndrome
AM	Alveolar Macrohages
BCG	bacille Calmette-Guerin,
Blast	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
CCDC	Chinese Center for Disease Contrôle
CCL2	Macrophage Chemokine
COG	Cluster of orthologous group
COVID-19	Corona Virus Disease 2019
DNA	Deoxyribonucleic Acid
EAI	East African-Indian
EMB	Ethambutol
ENA	European Nucleotide Archive
FEL	Fixed Effect Likelihood
FQL	Fluoroquinolone
Fubar	FAST Uncpnstraoned Bayesian ApprRoximation
GISAID	Global initiative on sharing all influenza data
GTR	General time reversible
H	Harleem
HIV	human immunodeficiency virus
HR	Heaptad Repeat
IDR	Intra Dermal Reaction
IgG	Immunoglobulin G
IgM	Immunoglobulin M
INH	Isoniazid
LAM	Latin American – Mediterranean
LTBI	Latent Tuberculosis Infection
MAF	Wester West Africa
MDR-TB	Multi Drug Resistance tuberculosis
MEME	Mixed Effects Model of Evolution
MERS-Cov	Middle East respiratory syndrome
MGIT	Mycobacteria Growth Indicator Tube
MIRU-VNTR	Mycobacterial Interspersed Repetitive Unit-Variable Number

ACE2	Angiotensin-converting enzyme 2
Mtb	Mycobacterium Tuberculosis
MTBC	Mycobacteriu Tuberculosis Complexe
NCBI	National Center for Biotechnology Information
NSP	Non-Structural Protein
ORF	Open Reading Frame
Pamp	Pathogens activated molecular patterns
PDB	Protein Data Bank
PDIM	Phthiocerol dimycoceroserate
RAST	Rapid Annotation using Subsystems Technologyæ
RBD	Receptor Binding Domain
RIF	Rifampicin
RNA	Ribonucleic Acid
RT-LAMP	Reverse-Transcriptase Loop-Mediated Isothermal Amplification assay
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SLAC	Single Likelihood Ancestor Counting
SM	Streptomycin
SNP	Single Nucleotid Polymorphism
TB	Tuberculosis
TLR	Toll ike Receptor
USAID	U.S. Agency for International Development
WGS	Whole Genome Sequencing
WHO	World Health Organization
XDR	Extensively Drug Resistance

Introduction

Several unique features characterize infectious diseases - they rely on a single agent as a cause. They can be transmitted from one person to another, cause epidemics, and have a strong impact on human evolution [1][2]. Besides, infectious diseases can be eradicated, but also new ones may emerge, creating a dynamic stage for human-infection interplay [3]. Clinical symptoms of pulmonary tract infections are highly variable and cannot identify the etiologic agent or agents. Although most respiratory tract infections are caused by viruses (69%; [4]), antibiotics are widely prescribed to treat symptoms. Antibiotics, which are of little or no benefit for viral infections, can contribute to the emergence and spread of resistant bacteria and to higher costs for health care [5][6].

The outcome of an acute respiratory tract infection depends on the organism's virulence and the inflammatory response in the lung. Innate immune responses to microbes in the lungs determine the outcome of infection; an insufficient response can result in life-threatening infection, but an excessive response can lead to life-threatening inflammatory injury.

Thus, the interest in the aftermath of superimposing viral pandemics (SARS-CoV-2) over long-standing diseases, such as tuberculosis (TB), which remains a significant disease for public health worldwide and especially in emerging economies. Tuberculosis and COVID-19 are among the four priority list for research and development. Recent advances in sequencing technologies have allowed genomics analysis to play an essential role in understanding outbreaks of pathogens. While techniques for generating and analyzing genomics data continue to evolve, the fundamental questions during disease outbreaks — where did the infectious agent originate, how is it changing, and where will it go next — remain the same. The subsequent chapters demonstrated the use of many of these methods to analyze infection outbreaks of two different pathogens. Chapter 2, describe our investigation of *Mycobacterium tuberculosis* before the COVID-19 pandemic in Morocco, with an emphasis on how we identify new drug resistance-associated mutations. Chapter 3, explain how we used genomics to understand the spread of SARS-Cov-2 in Morocco. Chapter 4 describes the continuity of the work on SARS-Cov-2, this time focusing on analyzing large scale genomics data in order to understand viral transmission worldwide. The final chapter is the culmination of many of the lessons learned in the preceding chapters and emphasizes the added value of genomics in public health investigations and the power of combining genomics and epidemiological data for outbreak response. Hence, detailing how new pandemic such as COVID-19 can affect ongoing epidemic like tuberculosis to provide a resource that may assist the study of future infectious disease outbreaks.

Contents

Chapter 1

Background Theory

1.1 *Mycobacterium tuberculosis*

1.1.1 Tuberculosis Burden wordwild

Tuberculosis (TB) is an ancient disease caused by a "*Mycobacterium Tuberculosis*" that most frequently affects the lungs; Humanity in its recorded history and prehistory has been affected by TB through large epidemics and continues to be a major global health concern [12]. Tuberculosis is one of the world's ten causes of death worldwide and the primary infection source from a single infectious agent [13]. Despite the sheer numbers, TB has been over shadowed for too long by HIV and malaria, and currently by the COVID-19. The World Health Organization (WHO) claims that 1/3 of the worldwide population is now infected with tuberculosis. About 10% of this population are predicted to develop active TB at some stage in their lifetime [14].

While TB disease seemed "conquered" by the end of the 1990s, two events cooperated with a sharp new increase in TB incidence. In the first place, the HIV epidemic has spread worldwide, making millions of people more vulnerable to active TB infection and latent TB infection reactivation (LTBI) [15][16]. Secondly, the Soviet Union's disintegration resulted in the collapse of Eastern Europe's health system and an outbreak of TB, especially MDR-TB [17]; causing a synergistic epidemic. However, HIV prevention and treatment programs helped decrease the number of cases from 33.7% before 2000 to 25.7% after 2010. From this perspective we raise a serious concern for a third outbreak with the emergence of the new virus [18] such as whether the COVID-19 cases would boost the development of active TB, or vice versa [19][20][21]. Globally, between 2015 and 2019, the incidence of TB decreased by around 9%; this was less than halfway to the 20% reduction estimated by the "End TB Strategy milestone between 2015 and 2020" [22].

In 2019, 10 million individuals worldwide (56% males, 32% females, and 12% of children) became ill with tuberculosis, from which 14% died [13]. More than 95% of TB deaths occurred in low and middle-income countries (**Figure 1.1**). The lowest incidence rate was mainly reported in high-income countries such as Western Europe, Canada, America, Australia, Saudi Arabia, and New Zealand. Countries with the highest inci-

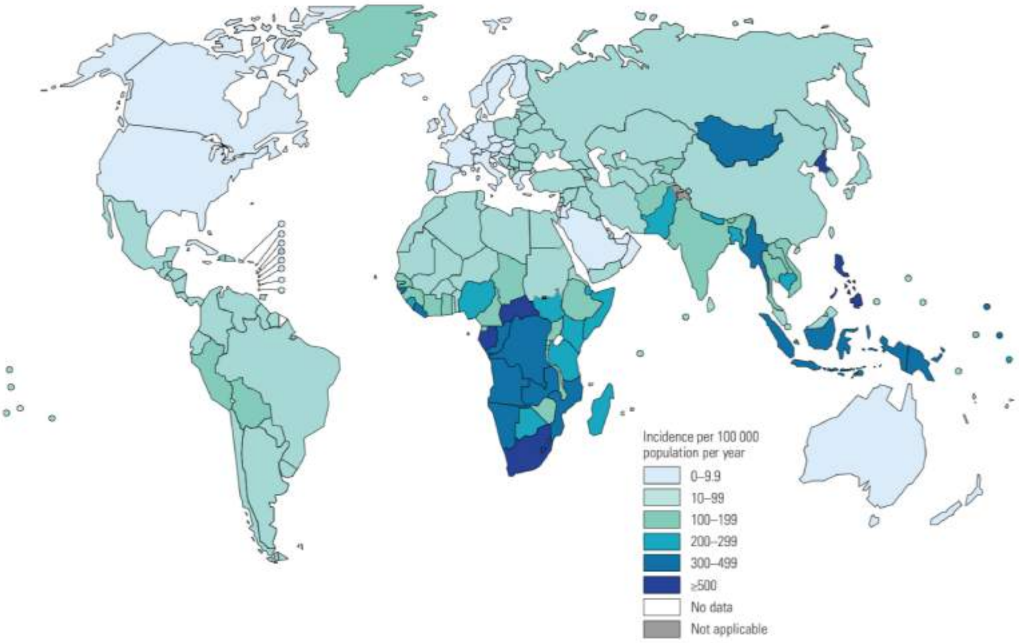

Figure 1.1: **Estimated tuberculosis incidence rates in 2019** [7]

dence rates of tuberculosis are predominantly in Africa (South Africa, Gabon, Central African Republic) where TB epidemics are fuelled by coinfection with HIV/AIDS and multidrug resistance [23] [24].

1.1.2 MDR-TB burden

MDR-TB diagnosis and treatment remains a significant obstacle and is far from being wholly solved [25]. Rifampicin resistance was observed in 61% of people with pulmonary tuberculosis worldwide in 2019, up from 51% in 2017 and 7% in 2012. A total of 206,030 individuals with MDR/RR-TB were identified and registered, a 10% rise from 2018[7]. Extensively drug resistance (XDR) strains, probably originated from MDR-TB strains. The XDR-TB show resistant to the second line drugs fluoroquinolones (FLQ) and aminoglycosides (AMI), have been registered in 92 countries. 6% of MDR-TB cases are estimated to be XDR-TB [26].

1.1.3 Burdden of Tubrculosis in Morocco

In Morocco, despite the Ministry of Health (MH) efforts to mitigate the disease, TB remains a significant public health issue [27]. According to the WHO report, in 2019, 31,536 cases were reported to have active TB causing 656 death [28]. Morocco ranked 66th among countries with TB incidence, with an estimated MDR number of 265 and two confirmed XDR-TB [7]. Furthermore, the studies conducted in some parts of Morocco have also revealed a high prevalence of MDR-TB, 26.4% in the Northern Region [29] and 12.8% in Casablanca [30]; A study by Chaoui et al. Showed that 30% of TB are resistant to fluoroquinolones and harbored the drug resistance-associated mutation in *gyrA* [31], while an other research identified 4 (2.6%) XDR and 18 (11.8%) pre-XDR isolates [32].

At the end of the seventies, a national TB program was set up to avoid, regulate, and gradually eradicate TB in Morocco. Free structured therapy regimens are given for tuberculosis patients [28]. Two reference national laboratories provide testing for TB infection. In 2004, Morocco managed to reach the WHO objectives related to TB diagnosis and treatment [27]. Thus, in 2015, 83% of the cases were detected, 85% were treated for TB [27]. However, TB incidence did not seem to decrease in Morocco. The recent statistics showed that TB incidence in Morocco was as high as in 2015 [27].

Poverty, gender, smoking, drinking, and HIV infection are significant risk factors for TB. In a two year retrospective study by Hanja et al., showed that more than two-thirds of the studied population were men (69%), three-quarters of these cases (71%) were smokers, 21% were cannabis addicts, and 7% were alcoholics [33]. On the other hand, TB is believed to prevail in prefectures rather than in provinces, and that population density in Morocco is a major risk factor. Such claims require further research.

Association between TB incidence and meteorological factors has also been explored and showed that annual rainfall in the east of Morocco has a significant effect on the TB rate [34]. More may need to be explored about TB in Morocco. Such as risk factors and Studies on spatial clusters of TB incidence that would have given a better understanding of where interventions are most required are lacking in Morocco.

1.1.4 Taxonomic Hierarchy and Microbiology of *M. tuberculosis*

TAXONOMY AND CLASSIFICATION:
KINGDOM: Bacteria PHYLUM: Actinobacteria
ORDER: Actinomycetales
SUBORDER: Corynebacterineae
FAMILY: Mycobacteriaceae
GENUS: Mycobacterium
SPECIES: *Mycobacterium tuberculosis*

There are over 170 species in the genus of Mycobacterium that can be divided into one of three major groups [36]: *Mycobacterium tuberculosis* complex species, *Mycobacterium tuberculosis* mycobacterial species, and *Mycobacterium leprae* [37]. Depending on their growth rate, the Mycobacterium genus is usually divided into two major groups: Slow-growing species, including *M.tuberculosis*, *M.bovis* and *M.leprae* and fast-growing species, such as *M.smegmatis.*

Human tuberculosis is a pulmonary condition mainly (but not exclusively) caused by MTBC [37]. The complex contains three human-specific members, *M.tuberculosis, M.canetti* and *M.africanum.* The rest primarily infect animal hosts, but share the potential to infect humans. At the genome level, *M. tuberculosis, M.canettii, M.africanum, M.bovis, M.microti, M.pinnipedii,* and *M.caprae* all share an almost 99% homology throughout their ribosomal RNA gene sequences [38],[39].

Neither Gram-positive nor Gram-negative; acid-fast bacterium identifiable by Ziehl–Neelsen

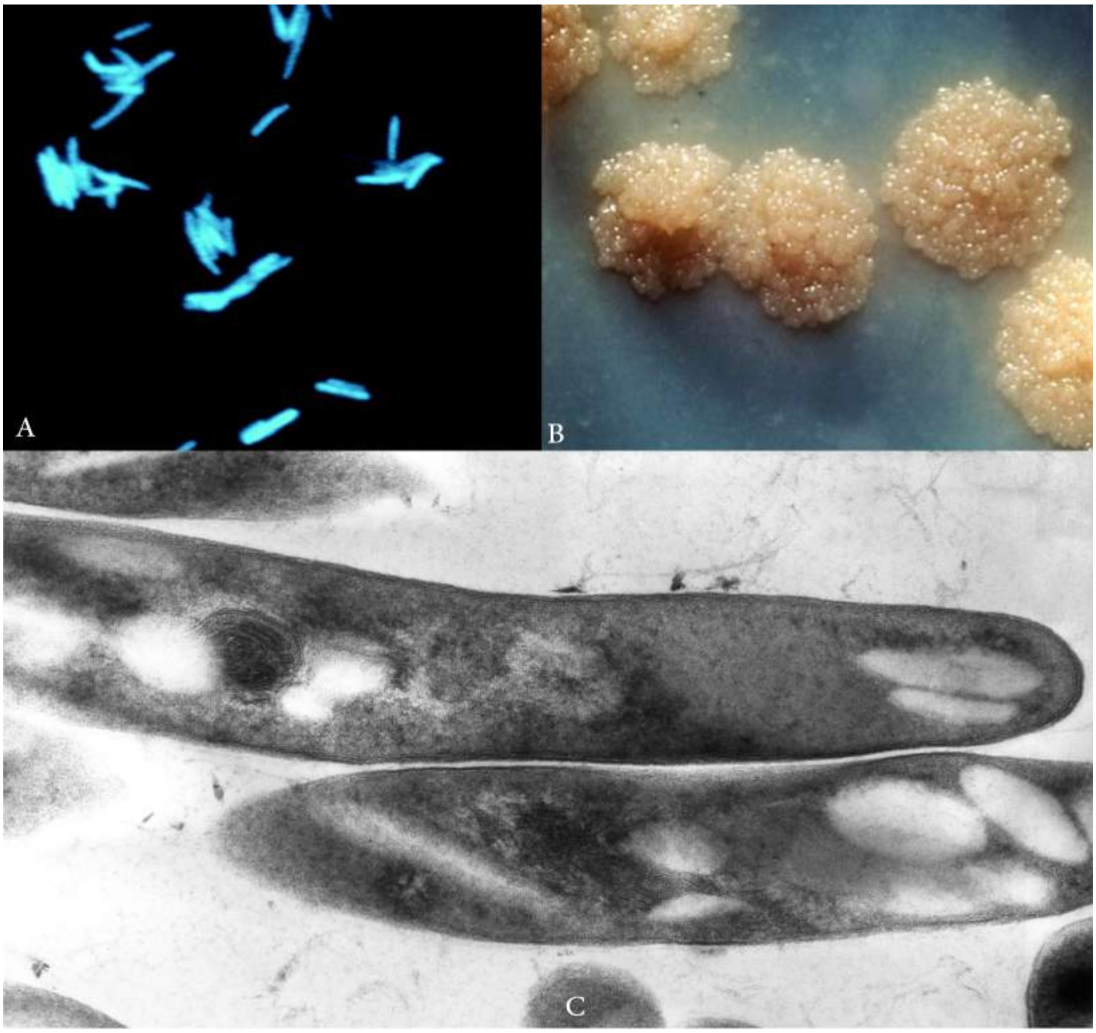

Figure 1.2: ***Mycobacterium tuberculosis* aspects**
.(**A**) *TB coloration*, (**B**) *TB Culture*, (**C**) TB Foluorecence microscopy [35]

staining (**Figure 1.2**). *Mycobacterium tuberculosis* is not identified in Grams' stain preparation due to the abundant lipid content in their cell envelope (40% cellular dry mass) [40] mainly composed of thick layer of peptidoglycan, lipids, glycolipids and polysaccharide. Furthermore, the tubercle bacilli aerobic strics, non-spore-forming, without flagella and often form beads in the culture media [39]. While *M. tuberculosis* can synthesize all the amino acids and enzymatic reaction cofactors, it is a slow grower and needs 22 hours in a growth medium compared with 20 minutes in *Escherichia coli* to double its number [41].This was determined to be related to the slower polymerization rate of a protein found in *M.tuberculosis* which affects cell division and cell wall biosynthesis [42].

1.1.5 Pathology:

1.1.5.1 Transmission and infection

Current WHO estimation indicate that *Mycobacterium tuberculosis*, transmitted by aerosol, infects approximately one third of the global population. Human to human transmission of TB is airborne occurs through small aerosols (5 m) by coughing, talking, or sneezing, the patient produces infectious droplets that can remain in the air for several hours. Contamination occurs during the inhalation of infectious droplets (**Figure 1.3**).

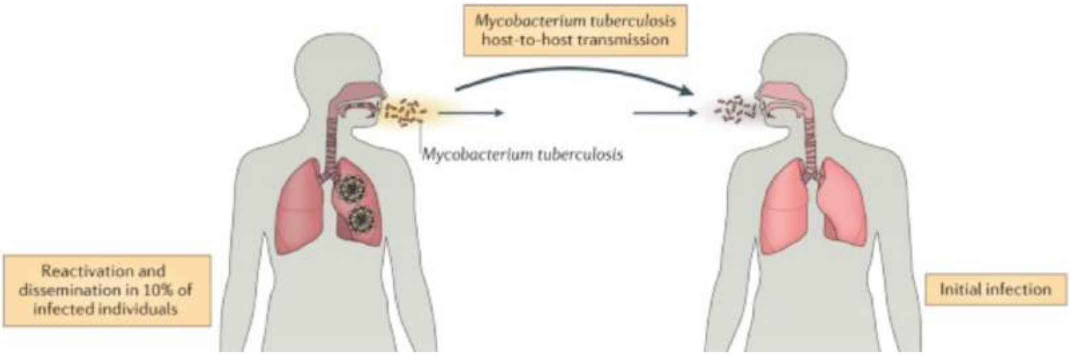
Mycobacterium tuberculosis host-to-host transmission
Mycobacterium tuberculosis
Reactivation and dissemination in 10% of infected individuals
Initial infection

Figure 1.3: **Infection initiation by the inhalation of aerosol droplets that contain bacteria.**

[8]

A person exposed to a contagious patient is not necessarily infected. The probability of contamination by *M. tuberculosis* depends on three factors:

1. **The degree of contagiousness of the infected person**
    - Positive bacteriological status
    - Virulence of bacteria (some strains are highly transmissible).
2. **Exhibition environment:**
    - Small unventilated rooms are favorable conditions for transmission.
    - The proximity of the source patient.
3. **Duration of exposure:**
    - People in close contact with patients with TB are most at risk of infection.

1.1.5.2 Pathogenesis & Immunity

Following transmission to a new host via aerosol, *M. tuberculosis* is believed to first undergo phagocytosis by alveolar macrophages (AMs) and subsequently interstitial macrophages [43]. In other pathogens, these macrophages are microbicidal and recruited through Toll-like receptor (TLR)-mediated signaling activated by the so-called pathogen activated molecular patterns (PAMPs) present on bacterial surfaces. However, Tb has developed strategies to avoid the microbicidal macrophages by expressing a surface lipid phthiocerol dimycoceroserate (PDIM) that masks the PAMPs. Once the PAMPS are hided the host's innate immune system does not recognize them (**Figure 1.4 - B**).

Simultaneously, a related surface lipid, phenolic glycolipid (PGL), is used to induce the macrophage chemokine CCL2 to recruit and infect growth-permitting macrophages.

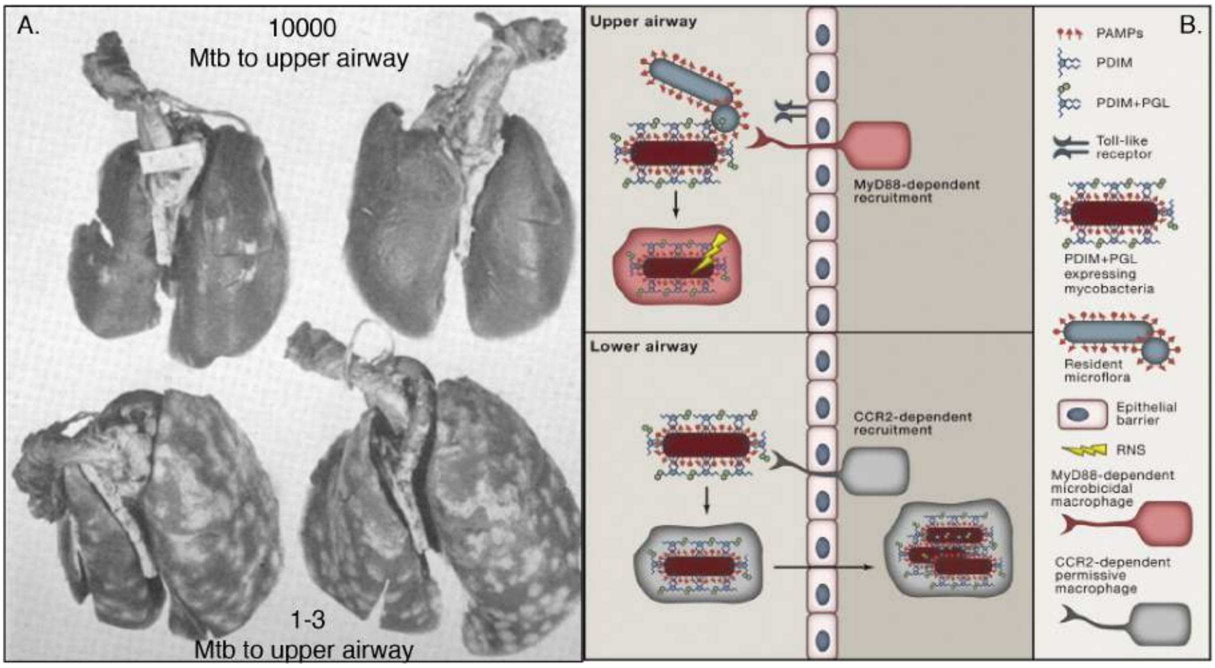
A.
10000
Mtb to upper airway
1-3
Mtb to upper airway
Upper airway
MyD88-dependent recruitment
Lower airway
CCR2-dependent recruitment
B.
PAMPs
PDIM
PDIM+PGL
Toll-like receptor
PDIM+PGL expressing mycobacteria
Resident microflora
Epithelial barrier
RNS
MyD88-dependent microbicidal macrophage
CCR2-dependent permissive macrophage

Figure 1.4: **Mechanism of priminary infection of *Mtb*.** **(A)** Lungs from animals breathing different numbers of tubercle bacilli. **(B)** *Mtb* avoids the recruitment of microbicidal macrophages to the site of infection by masking its PAMPs with the PDIM lipid.[8][9]

However, this strategy is ineffective in the upper airway, which is full of TLR-stimulating commensal bacteria [8]. Thus tb infection must be initiated through uses small goutlets that directly settle into the lower lung spaces since they contain few if any, commensals [9] as the third strategy of infection (**Figure 1.4 - A**).

Once internalized, ESX-1 activate the secretion system to blocks the fusion between the phagosome and lysosome. To start cytosolic surveillance, *Mtb* releases bacterial products into the macrophage cytosol, which induces type I interferon response, resulting in bacteria's survival and rapid growth [8]. After phagocytosis, infected cells migrate to local draining lymph nodes. T cells recognize antigens on *Mycobacterium tuberculosis* and differentiate into specific T cells. This differentiation results in the release of lymphokines and macrophages' activation, which inhibit the growth of phagocytosed bacteria. Macrophages initiate a signalling cascade leading to the recruitment of additional monocytes and lymphocytes. This aggregate of immune cells surrounds the infection site and forms an organized cellular structure known as a granuloma (**Figure 1.5**)[8].

Following infection, *M. Tuberculosis* often enters a prolonged state of latent, asymptomatic infection. This phase can last for decades within granulomas, if not for the rest of the host's life. There is a subgroup of hosts in which the latent infection re-activates, causing active disease. The immune response in the centre of the granuloma contains pro-inflammatory molecules[44].

In contrast, the surrounding area has anti-inflammatory components. The balance of *Mtb* and macrophage interactions can influence the granuloma outcome, which may either constrain the infection or promote its systemic dissemination. If the bacterial load is too high, granuloma may fail to contain the infection, and bacteria will spread to other organs, including the brain. The bacterium can enter the bloodstream or respiratory tract to be released, and this is said to be the first phase of active TB disease[8].

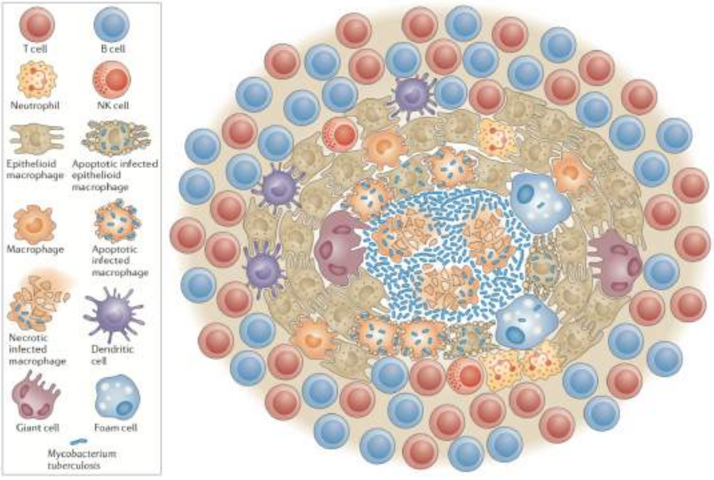
T cell
B cell
Neutrophil
NK cell
Epithelioid macrophage
Apoptotic infected epithelioid macrophage
Macrophage
Apoptotic infected macrophage
Necrotic infected macrophage
Dendritic cell
Giant cell
Foam cell
Mycobacterium tuberculosis

Figure 1.5: **Structure and cellular constituents of the tuberculous granuloma**.[10]

1.1.6 *Mycobacterium tubrculosis* diversity and lineagess

Evolutionary classification of MTB strains is mainly based on large sequence polymorphism (LSP) and single nucleotide polymorphism (SNP) due to the hight genome stability and low incidence of horizontal gene transfer, reacquisation of deleted region sits. SNP and LSP classification have categorized Mycobacterium tuberculosis strains into 7 phylogenetic major linages. Thus, recent studies has showen the emergence of a 2 new lineages mainly in Africa named L8 and L9 (**Figure 1.6**).

a) Lineage 2 - East-Asian:

One of the most virulent and successful MTBC variants; This lineage is responsible for more than 25% of the global tuberculosis epidemic [46], [47]. It is mostly known by the so-called Beijing family more predominant in East and South East Asia, accounting for over 50% - 85% of total cases [48], [42]. A significant increase of L2–Beijing prevalence was also reported in Africa over time [49][50][51]. L2 was associated with the emergence of MDR strains [52],[53][54][42]. Beijing strains may be more pathogenic or virulent compared to the *M.tuberculosis* strains, experimental and clinical evidence suggests a hyper-virulent phenotype of Beijing strains [55] [56]) and a higher mutation rate compared with other strains [57]. Several hypotheses have been proposed to explain the widespread dissemination of these strains:

- The extensive dissemination may be related to the global migrations of Asian in the twentieth century [55]
- A positive selection of this genotype after the use of the bacille Calmette-Guérin (BCG) by having less protective efficacy against this lineage [58], [59]

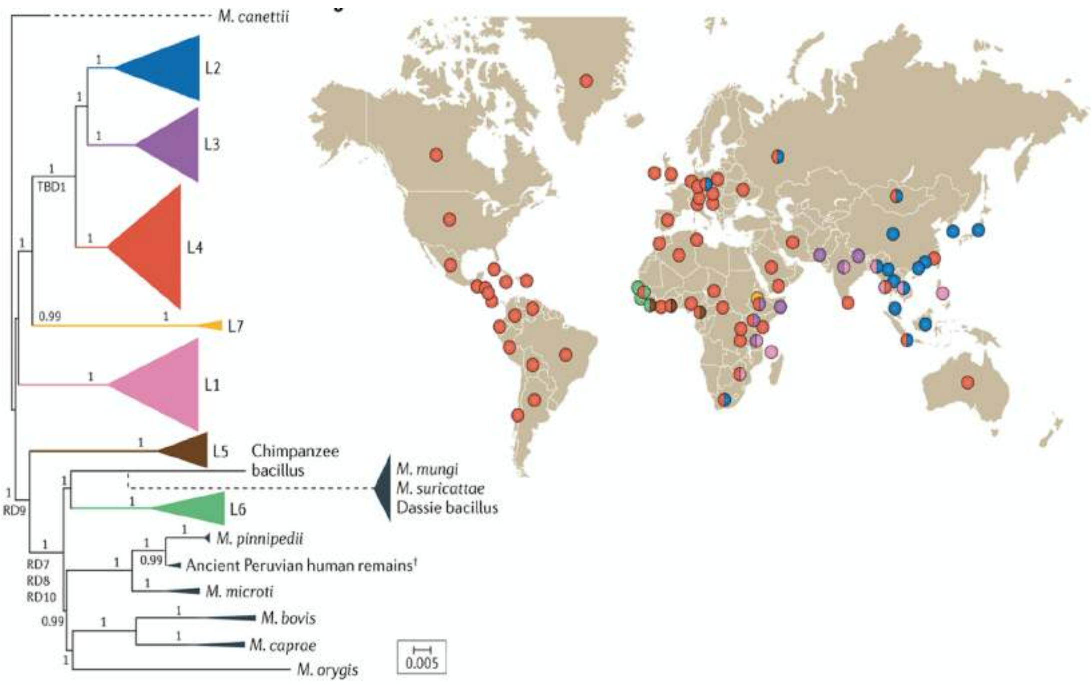

Figure 1.6: Distribution of *M. tuberculosis* lineage worldwid
[45]

b) Lineage 3 -East African-Indian

One of the most ancient *Mtb* lineages, Bayesian statistics study by Wirth et al. l estimated that EAI arose and spread out of Mesopotamia 10000 years ago. Currently the EAI is essentially localized Asian countries particularly Cambodia [60] (60%), Myanmar (48.4%) [61], Northern Vietnam (38.5%)[62] the southern region of Taiwan (32.1%)[63], Singapore (25.6%)[56], and Saudi Arabia (23.8%) [57]. The reasons for the significant EAI prevalence in Asian countries may be related to different selection pressures caused by the BCG vaccination, environmental factors such as community hygiene, tropical climates, population density, and biological factors host-bacterial interaction [64]. Chen et al. recently supported findings demonstrating that EAI isolates in Taiwan induced very high levels of proinflammatory cytokines in modern Beijing lineages and explain the fast dissemination of this lineage in Taiwan [65] [66][67].

c) Lineage 4 - Euro-American

L4 is the most widespread, affecting humans in all countries. Population genomics and phylogeographic analyses of MTB lineage 4 found that historical migrations out of Europe were the leading cause of this lineage4 dispersal [68]. A recent study also reported that the main driving force behind L4's global dispersal in Africa and the Americas is the European colonization efforts [69]. The same study also reported that the spread of L4 was also associated with MDR Strains' emergence in these regions. L4 was also dominated in the Middle East and Oceania [63].

L4 consists of 10 distinct families and other unclassified families, from which LAM, T, X, H, and S families are widespread worldwide. The distribution of these specific families varies according to the regions. Dodo et al. demonstrated that L4 have a higher replication rate and produce a significant amount of proinflammatory cytokines, therefore, contribute to pathogenicity increase.

d) Lineage 5 - West Africa 1, Lineage 6 - West Africa 2

In 1968, Castets and colleagues first described africanum as a specific subspecies of the *Mycobacterium tuberculosis* complex (MTBC) ([70]. It was further divided into two phylogenetically distinct lineages: *M.africanum* West African 1 (Gulf of Guinea) and *M.africanum* West African 2 (western West Africa MAF) [70]. These lineages are restricted to West Africa, where they cause up to half of the human pulmonary tuberculosis. However, *M.africanum* is rarely reported in other continents. it was assumed that *M.africanum* dissemination in these countries is mainly due to human migration from disease-endemic West African regions.

Some comparative studies suggest the MAF lineages are different and occupy different ecological niches. Thus, Three hypotheses attempt to explain the restriction of MAF to West Africa:

- The MAF may have emigrated outside Africa but was subsequently outcompeted by MTB, which in animal models is more virulent than MAF. This hypothesis was supported by the low production of proinflammatory cytokines at the early stage of the infection compared to modern MTBC, thus allowing slower transmission [71]. The virulence mechanism was also shown to affect MAF in many essential genes, such as S DosR regulon [72], ESAT6 [73], and PhoP/R [74].
- Specific host-pathogen interactions and adaptation to West African human populations. The statistical association of L5 with the native West African ethnic group known as “Ewe” was reported by two independent studies in Ghana [75][76].
- MAF might be zoonotic with an animal reservoir limited to West Africa. Possibility relies on the phylogenetic placement of MAF in the middle of an animal-adapted MTBC cluster [77].

e) Lineage 7,8 and 9

Lineage 7 is predominantly restricted to Ethiopia in the horn of Africa [78][79]. This lineage is characterized by smaller colony diameter and weight compared to lineages 3 and 4 combined. In vitro, Lineage 7 grow slower than non-lineage 7 *Mtb* strains, explaining the delay in seeking treatment observed in patients infected with this lineage [80]. On the other hand, lineage 8 and 9 have been recently proposed and restricted to Ethiopia and east of Africa, respectively. Litlle is known for there virulence and transmission mechanism.mechanism [81][82].

1.1.7 Diagnostic Tests

The diagnosis of tuberculosis should be made in a favorable epidemic situation or the presence of a person presenting general signs suggestive of the disease. Nevertheless, whatever the epidemiological context, the deterioration of a person’s general condition with weight loss, asthenia, and moderate fever should orient the diagnosis towards tuberculosis, especially if all of these signs persist for more than three weeks. These general signs can also be accompanied by respiratory manifestations such as a persistent and increasingly frequent cough, hemoptysis or even dyspnea [83]. In latent tuberculosis, it is usually not accompanied by symptoms, clinical signs or radiological signs.

1.1.7.1 Direct examination by microscopie

In cases of suspected pulmonary *Mtb*, the bacteriological examination of the patient's sputum is performed. To demonstrate the property of specific acid-alcohol resistance of mycobacteria, a smear or spread of the biological sample on a thin slide is carried out, then stained. There are two stains: Ziehl-Nelsen stain, or auramine stain. It can only detect bacteria in a sample if there are at least 0.5 to 1.1 bacteria per micro-liter of the sample [84].

1.1.7.2 Chest x-rays

Chest x-ray provides an initial assessment of chest lesions which may vary in morphology and extent. There are three types of primary lesions: the nodule, the infiltrate and the cavern. The appearance of the lesions is unrelated to the severity of the disease. In its pulmonary form, tuberculosis is manifested by the presence of infiltrates and nodules mainly located in the tops of the lungs and sometimes associated with caverns.

1.1.7.3 Tuberculin Test

Tuberculin intra-dermal reaction (IDR) is a delayed hypersensitivity skin reaction caused by the influx of T cells and macrophages to the site of intra-dermal injection of tuberculin (mycobacterial antigens). This reaction demonstrates the existence of cell-mediated immunity to mycobacteria in the patient, induced either by prior vaccination with BCG or by prior contact with *Mtb* or certain atypical mycobacteria. The IDR involves injecting 0.10 ml of liquid tuberculin solution into the dermis of the anterior aspect of the forearm and measuring the diameter of the induced induration after 72 hours. The test will be positive if this diameter is greater than 5 mm and very positive if it is greater than 10 mm. In this second case, there will be a presumption of tuberculosis disease.

1.1.7.4 Interferon Gamma Release Assays

Immunological tests such as the QuantiFERON-TB-Gold® test and the TSPOT-TB® test have recently been developed for the diagnosis of latent tuberculosis infection [85][86]. These tests have the advantage of not being disturbed by a previous vaccination with BCG or by an infection caused by another mycobacteria [87] if the blood is positive for interferon gamma is considered potentially TB positive.

1.1.7.5 Xpert MTB/RIF assay

These tests are based on gene amplification by real-time PCR on specific *M. tuberculosis* sequences [88], allowing the detection of nucleic acid sequences specific to tuberculosis bacilli. These techniques exhibit excellent sensitivity and specificity when used from culture extracts. The sample is loaded into the GeneXpert device, where the gene of interest is amplified via RT-PCR, after which five probes are used to detect different segments of this gene. This test targets the 81-bp rifampicin-resistance determining region (RDRR) of the rpoB gene (5 probes: A, B, C, D, E) while simultaneously testing for drug resistance.

1.1.7.6 Direct examination

A smear, or spread of the biological sample on a thin slide, is carried out, then stained to demonstrate the specific property of mycobacteria: their acid-alcohol resistance. Two methods can be used, the Ziehl-Neelsen method or the auramine (fluorochrome) method. Ziehl-Neelsen stain shows Acid-Alcohol-Resistant Bacilli (AFB) as small red sticks isolated or in small clusters on a blue background, while they will appear as small shiny yellow-green rods on a dark background with auramine staining.

1.1.7.7 Culture

Culture is the gold standard for diagnosing TB. Only a positive culture for *M. tuberculosis* is conclusive proof of the diagnosis of TB. It is more sensitive than microscopy. Its sensitivity varies from 80% to 85% while the sensitivity varies from 50% to 80% for microscopy. It is carried out on enriched, solid culture medium (Löwenstein-Jensen's medium or Colestos medium) where results are obtained between 4 to 8 weeks; or on liquid medium (Middlebrook 7H11, 7H10, 7H9, MGIT / Mycobacteria Growth Indicator Tube or Dubos) where results are obtained more quickly (approximately 15 days).

1.2 SARS-CoV-2

1.2.1 COVID-19 Burden

1.2.1.1 COVID-19 burden worldwide

In December of 2019, a novel coronavirus (2019-nCoV) was identified in Wuhan in China's Hubei region. The outbreak quickly spread to neighboring countries. Due to quick communication and rapid events, quarantines and screening for travelers were put in place to prevent the disease's spread. The initial infection affected China, but a second cluster was found on a cruise ship called the Diamond Princess docked in Japan [89].

SARS-Cov-2 has been identified to have zoonotic origin (Pangolin) [90][91]. The ability to mutate rapidly, fast transmission, and adapt to a new host is a unique feature of these viruses. In January 2020, China had seen an exponential increase in the number of cases. The problem has become global today and declared on the on 11 March 2020 by World Health Organization (WHO) as a pandemic, and a real threat with the greatest risk effect on all levels. The latest WHO public health surveillance guidance asks countries to report confirmed cases [92]. The pandemic swept through two waves over the ten months. In May and June 2020, the number of new cases decreased drastically. Thus, revealed the typical phases: first, the initial outbreak period of exponentially increasing infections; second, the phase characterized by high social distance measures; and finally, the phase involving disease spread reduction correlated with weakened containment measures [93].

Since October, a second severe wave of COVID-19 started. While no clear pattern exists between countries. An increase of number of cases started from mid-October, to decline afterward during November 28th. Governments have progressively re-enforced new restrictive measures from 13 October on-wards. At the end of November the second wave was in a very weak period, with the reproduction number decreasing to values below the threshold of 1 nearly everywhere. Globally, as of 5:52 pm CET, 16 January 2021, 92,506,811 confirmed cases of COVID-19, including 2,001,773 deaths, were reported to WHO (**Figure 1.7**).

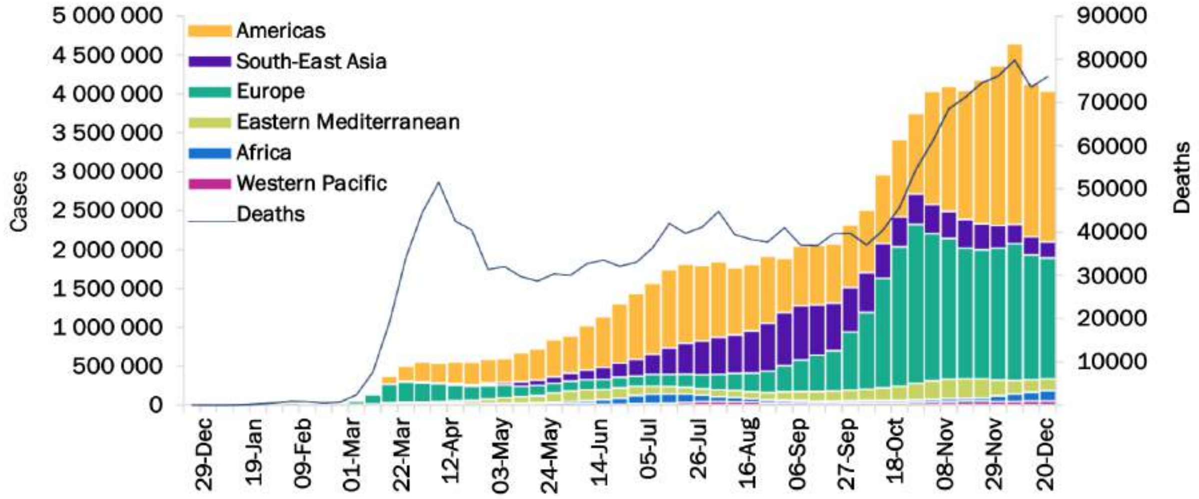

Figure 1.7: Covid-19 evolution over time [11].

In the last few months, two new variants of SARS-CoV-2 have been reported to the WHO as unusual public health events. Recent reports have again raised interest and concern about the impact of viral changes on various variants of SARS-CoV-2 [94][95][96]. The first variant was referred to as Voc 202012/01 from the United Kingdom of Great Britain and Northern Ireland, while the second from South Africa was designated as 501y. Preliminary epidemiologic, modeling, and clinical studies of the new variant strain of SARS-CoV-2 (reported as SARS-CoV-2 VOC 202012/01) suggest an increase in transmissibility [97][98] and reinfection cases [97]. Simultaneously, no change in disease severity (as measured by the length of hospitalization and 28-day case fatality).

So far, 40 countries outside of the United Kingdom, and six countries outside of South Africa have reported cases of the new Variants.Further studies are currently conducted to determine the transmissibility, severity, risk of reinfection and antibody response to the new variants, as well as potential impact on therapies, Vaccines and diagnostics methods [99].

1.2.1.2 COVID-19 burden in Morocco

The first confirmed case was detected on Monday 02 March 2020 and few days later, other new cases was registered. Just after the first death notification, the Moroccan government decided to close territorial borders in order to prevent the virus transmission and quarantine has been imposed on Friday 20 March 2020. Morocco adopts the therapy with chloroquine publishes by Gautret et al[100].

A total of 78 cases are being taken care of in the past 24 hours by intensive care and resuscitation services, reporting a drop to around 3% in the number of cases in serious or critical condition, while cases with little or no symptoms represent around 92%.3 In the lead is Casablanca-Settat region with 1203 infected people. The Marrakech Safi region is in second position, totaling 1059 cases, followed by that of Tangier Tetouan Al Hoceima, which has had 657 cases. The region Meknes-Fez is in fourth position with 584 cases and 546 cases in the Daraa-Tafilalet in fifth place and the Rabat Sale Kenitra region comes in sixth position with 329 cases and the Orientale region in seventh position with 175 cases. The regions with less than 100 infected people are Beni Mellal Khenifra (83), Souss Massa (51), Laayoune Sakia Al Hamra (4), Dakhla Oued Deheb (2) and Guelmin-Oued-Noun (36). according to Circular No.22 relating to the prescription and dispensation of Chloroquine and hydroxychloroquine at the level of healthcare establishments. However, no serious epidemiological study has been developed to assess and analyse the health situation in Morocco as a whole, according to each principal measure of emergency.

1.2.2 SARS-Cov-2 propreties

1.2.2.1 Structure

Spherical, enveloped virus of 60-220 nm, comprises from the outside to the inside, the glycoprotein Spike (S) (gives the crowned appearance to the virus under electron microscopy), the envelope, the membrane and the nucleocapsid it even, icosahedral with cubic symmetry. The latter contains a viral genome molecule: single-stranded, unsegmented, positive ribonucleic acid (RNA) (29,881 base pairs) (**Figure 1.8**) .

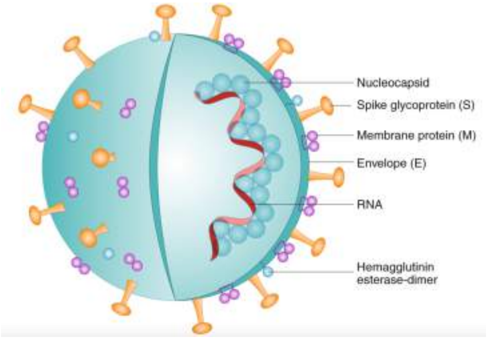

Figure 1.8: SARS-CoV-2 structre
[101]

1.2.2.2 Genome

The CoV genome has a varying number of open reading frames (ORFs). Two-thirds of the viral RNA is located primarily in the first ORF (ORF1a / b), translates two polyproteins, pp1a and pp1b, and encodes 16 non-structural proteins (NSPs), while the remaining ORFs encode proteins from structure and accessory proteins. The rest of the virus genome encodes four essential structural proteins, including glycoprotein (S), envelope protein (E), matrix protein (M) and core protein (N), as well as several accessory proteins , which interfere with the host's immune response [102]. The study by Tang et al. analyzed 103 genomes of patients infected with Covid-19 and identified two strains of Sar-CoV-2: strain L and strain S. The strain L is more aggressive and contagious [103].

1.2.2.3 Pathogenesis

The Sars-CoV-2 multiplication cycle in the cell comprises attachment, penetration and decapsidation, followed by the synthesis of macromolecules (nucleic acids and proteins) in three phases: early-immediate, immediate and late. These syntheses will allow the assembly of the nucleocapsids. Then envelopment and release of the infectious virions simultaneously as lysis of the infected cell—this lytic cycle exists in respiratory cells infected with the virus. The virus attaches specifically to the susceptible cell receptor through a high-affinity interaction between the viral S protein and ACE2 (Angiotensin-converting enzyme), the host's cellular receptor. The S protein is made up of two functional subunits: the S1 subunit allows the virus to bind to the host cell's receptor and the S2 subunit ensures the fusion of the viral envelope and the cell membrane. Cleavage of the S protein by proteases from the host cell activates fusion at two tandem sites, heptad repeat 1 (HR1) [7] and HR2 [8]. Thus, the viral RNA is released into the cytoplasm. The replication-transcription complex(RTC) ensures the replication of the genome, the synthesis of proteins.

The pathogenesis of SARS-CoV-2 infection in humans can vary from mild symptoms to severe respiratory failure. after binding to epithelial cells in the respiratory tract, SARS-CoV-2 begins to replicate and migrate to the airways and enters alveolar epithelial cells in the lungs. The rapid replication of SARS-CoV-2 replication triggers the immune response. In general, the leading cause of death in patients with COVID-19 is the severe Cytokine storm which may cause respiratory failure[104][105].

1.2.3 SARS-CoV-2 Diagnosis Tests

1.2.3.1 Nucleic acid amplification tests (NAATs)

WHO guidelines recommend the use of RdRp, E, N and S genes in different combinations diagnosis is based on targeting the RdRp/Helicase (Hel), Nucleocapside (N) and Envelop (E), genes spike (S) and E rdrp gene of the virus with a good specificity (differentiating SARS-CoV-2 from SARS-Cov-1) and sensitivity. A comparison between all targeted genes revealed that the best results were obtained with RdRp/Hel genes , RNA extraction methods can generally be classified into (a) one step (with the RT step and the PCR reaction in the same tube) and (b) two-step RT-PCR (initial creation of DNA copies with RT reaction followed by their addiction to the PCR reaction).

1.2.3.2 Lamp Tests

Loop-mediated isothermal amplification (Lamp). It is a method of rapid, sensitive and efficient visual amplification of nucleic acids. Lately, this method has been widely used for the isolation of influenza virus, Middle East respiratory syndrome-CoV, West Nile virus, Ebola virus, Zika virus, yellow fever virus and 'a variety of other pathogens.

1.2.3.3 Serological testing

Serology is a major challenge for evaluating the immune protection of populations against the virus and for better understanding the epidemiology. Serological tests target the following antigens in particular: the nuclear capsid protein which is highly expressed and provides high sensitivity, and the spike protein which provides high specificity.

One of the downsides of serological assays is the limited sensitivity at an early stage, when the host has not yet developed specific antibodies. In the specific case of SARS-CoV-2, data from literature showed production of IgM and IgG starting after the first week from infection and generally detectable from the second, leaving some space for delayed antibody responses, previously associated (for MERS-CoV) with more severe disease.

1.3 Aims

The general aim of this thesis is to study the transmission, dynamics and evolutionary diversity of two pulmonary tract infection diseases using whole-genome sequencing. Specifically to:

1. Understand the population structure of *Mtb* strains circulating in Morocco and assess the evolutionary history of the dominant strains
2. Understand the diversity of the PE and PPE genes in *Mtb*
3. Characterize drug resistance-associated mutations
4. Identify new mutations and predict their effect on resistance using protein modelling.
5. Assess the utility of Oxford Nanopore Technologies in the sequencing of viral genomes.
6. Identify SARS-CoV-2 Moroccan lineages and variations.
7. Predict the transmission dynamic and origin of SARS-CoV-2 strains in Morocco.
8. Characterize the genomics variations of SARS-CoV-2 worldwide.
9. Identify the geographic location and origin of hostpot mutations in SARS-CoV-2.
10. Identity variants circulating in the world and estimate divergence and mutational rate of SARS-CoV-2.
11. Predict potential therapeutic targets in SARS-CoV-2 genomes.
12. Evaluate the possibility of universal vaccine efficiency.

Chapter 2

Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculous drug resistance and population structure

2.1 Preface

As noted in the previous chapter, tuberculosis disease caused more than 645 death in Morocco, 2019. Despite considerable efforts deployed and the free tuberculosis treatment, the annual incidence remains high.

Several studies have been conducted on Mtb to evaluate its prevalence in which the use of whole-genome sequencing was lacking. Therefore, we used WGS to analyze these genomes, looking for patterns in the sequences that could provide clues in understanding resistance, transmission and origin.

We released a sequencing of 2 *Mtbs* strains and identified two main sublineages (H1 and LAM9). The overall Moroccan genome sequences analysis confirmed that the MDR strains belonged to the LAM9 while susceptible strains were Harlem. Additionally, resistance strains showed a higher mutational rate than those belonging to LAM9, with the majority of mutations identified in PPE-PE genes. A flow up study was conducted using seven additional *Mtb* genomes. The primary aim of this study was to evaluate the prevalence of drug resistance-associated SNVs and their relationship to samples phenotypic pattern (**Figure 2.3**).

These types of analyses led to a more detailed understanding of *Mtb* resistance and allowed identifying new drug resistance-associated mutations. The effect of the newly identified mutation was explored with protein modelling analysis and highlighted an increase of affinity between GyrA/GyrB and levofloxacin.

The study conducted in this chapter could be of great help in highlights the significance of employing WGS in diagnosis and monitoring *MDR-Mtb* strains.

2.2 Whole-Genome Sequencing of Two Moroccan *Mycobacterium tuberculosis* Strains

M.Laamarti[a], N. El Mrimar[b], T. Alouane[a], S. Kartti[a], E. Belouad[b], F. Bssaibis[b], A. Zegmout[c], R. El Jaoudi[a], A. Maleb[e], A. Abid, N. El Hajjami[f], A. Lemnouer[b], S. Siah[d], L. Belyamani[f], M. Elouennass[b], A. Ibrahimi [a]

[a] Biotechnology Lab (MedBiotech Center), Rabat Medical and Pharmacy School, University Mohammed V, Rabat, Morocco
[b] Department of Bacteriology, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco
[c] Pneumology Department, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco
[d]Department of Burns, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco
[e] Laboratory of Microbiology, Mohammed VI University Hospital/Faculty of Medicine and Pharmacy, University Mohammed the First, Oujda, Morocco
[f]Emergency Department, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco

2.2.1 Abstract

Mycobacterium tuberculosis is known to cause pulmonary and extrapulmonary tuberculosis. In Morocco, the spread of multidrug-resistant (MDR) tuberculosis (TB) has become a major challenge. Here, we announce the draft genome sequences of two *Mycobacterium tuberculosis* strains, MTB1 and MTB2, isolated from patients with pulmonary tuberculosis in Morocco, to describe variants associated with drug resistance.

2.2.2 Announcement

Tuberculosis is an urgent public health problem in Morocco caused by *Mycobacterium tuberculosis* bacteria. Control of the bacteria has been recently complicated by the emergence of multidrug-resistant (MDR) strains showing resistance to the second-line treatment (rifampin, isoniazid) due to probable excessive antibiotic use (1–3). The identification of differences between MDR and sensitive *Mycobacterium tuberculosis* would improve the identification of the drug resistance site. Whole-genome sequencing using next-generation sequencing technologies is emerging as a rapid method for genetic characterization (4-5). Thus, we provide the whole-genome sequencing data of two *Mycobacterium tuberculosis* strains, MTB1 and MTB2, with different resistance profiles.

Two *Mycobacterium tuberculosis* strains recovered from patients with pulmonary tuberculosis at the military hospital in Rabat, Morocco, and grown in Middlebrook 7H9 medium using the Bactec MGIT 320 system (Becton, Dickinson) were received for sequencing. DNA was purified using the Qiagen DNA extraction kit (QIAamp DNA minikit) following the kit protocol, and DNA libraries were prepared using the Nextera XT library preparation kit V3. Genomic DNA was sequenced using the Illumina MiSeq platform (San Diego, CA, USA) in paired-end (2x300-bp) format. Yields of 1,616,965 and 620,966 reads were preprocessed for quality checking using FastQC, further assembled using A5-miseq with default parameters (6) (the default settings include running Trimmomatic for read preprocessing) into 4,341,655-bp and 4,291,403-bp draft genome sequences, and divided into 242 (N50, 43,730 bp) and 238 (N50, 35,128 bp) contigs for MTB1 and MTB2, respectively. The assemblies shared a GC content of 60% and 3 arnT operons. The genome annotation was performed using the NCBI annotation pipeline (7) and identified 4,265 and 4,422 coding DNA sequences (CDS) in MTB1 and MTB2, respectively.

Furthermore, reads were mapped to the h37rv reference genome using BWA (8). Variants were called using SAMtools (9) and annotated by SnpEff (10). The genome sequence displayed hot spot mutations in genes associated with resistance. MTB1 harbored mutations in katG (Ser315Thr) and rpoB (Ser450Leu) responsible for resistance to isoniazid and rifampin, respectively (11). No resistance-associated mutations were identified in the genes (gyrA, gyrB, and rrs) linked to second-line drugs (12). However, pyrazinamide resistance was due to a mutation in pncA (Gly97Asp) (13), which classifies MTB1 as pre-extensively drug resistant (pre-XDR), while MTB2 did not show any mutations associated with first- or second-line drug resistance in its genotype. This study represents an initial analysis of a *Mycobacterium tuberculosis* collection to highlight the resistance profile in Morocco.

2.2.3 Data availability.

The genome sequences of *Mycobacterium tuberculosis* MTB1 and MTB2 have been deposited in DDBJ/ENA/GenBank under the accession numbers **NARM00000000.1** and **NARL00000000.1**, respectively. The SRA accession numbers for MTB1 and MTB2 are **SRR12031346** and **SRR12031345**, respectively.

2.2.4 References

1. Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song,Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirusfrom patients with pneumonia in china, 2019.New England journal of medicine,2020.

2. Tung Phan. Genetic diversity and evolution of sars-cov-2.Infection, genetics andevolution, 81:104260, 2020.

3. Josh Quick. ncov-2019 sequencing protocol.Protocols. io, 2020.

4. Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform.bioinformatics, 25(14):1754–1760, 2009.

5. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Ga-bor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/mapformat and samtools.Bioinformatics, 25(16):2078–2079, 2009.

6. Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, LuanWang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotatingand predicting the effects of single nucleotide polymorphisms, snpeff: Snps in thegenome of drosophila melanogaster strain w1118; iso-2; iso-3.Fly, 6(2):80–92, 2012.

7. Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. Mafft version5: improvement in accuracy of multiple sequence alignment.Nucleic acids research,33(2):511–518, 2005.

8. Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh.Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihoodphylogenies.Molec biology and evolution, 32(1):268–274, 2015.

9. James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Pot-ter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher.Nextstrain: real-time tracking of pathogen evolution.Bioinformatics, 34(23):4121–4123, 2018.6

10. Lizhou Zhang, Cody B Jackson, Huihui Mou, Amrita Ojha, Erumbi S Rangarajan,Tina Izard, Michael Farzan, and Hyeryunc Choe. The d614g mutation in the sars-cov-2 spike protein reduces s1 shedding and increases infectivity.BioRxiv, 2020.

11. Tarek Alouane, Meriem Laamarti, Abdelomunim Essabbar, Mohammed Hakmi,El Mehdi Bouricha, MW Chemao-Elfihri, Souad Kartti, Nasma Boumajdi, HoudaBendani, Rokia Laamarti, et al. Genomic diversity and hotspot mutations in 30,983sars-cov-2 genomes: moving toward a universal vaccine for the “confined virus”?Pathogens, 9(10):829, 2020.

12. Meriem Laamarti, Tarek Alouane, Souad Kartti, MW Chemao-Elfihri, MohammedHakmi, Abdelomunim Essabbar, Mohamed Laamarti, Haitam Hlali, Houda Bendani, Nassma Boumajdi, et al. Large scale genomic analysis of 3067 sars-cov-2genomes reveals a clonal geo-distribution and a rich genetic variations of hotspotsmutations.Plos one, 15(11):e0240345, 2020.

2.3 Genomic analysis of *Mycobacterium tuberculosis* strains from Morocco

M.Laamarti[a], N. El Mrimar[b], T. Alouane[a], S. Kartti[a], E. Belouad[b], F. Bssaibis[b], A. Zegmout[c], R. El Jaoudi[a], A. Maleb[e], A. Abid, N. El Hajjami[f], A. Lemnouer[b], S. Siah[d], L. Belyamani[f], M. Elouennass[b], A. Ibrahimi [a]

[a] Biotechnology Lab (MedBiotech Center), Rabat Medical and Pharmacy School, University Mohammed V, Rabat, Morocco
[b] Department of Bacteriology, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco
[c] Pneumology Department, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco
[d]Department of Burns, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco
[e] Laboratory of Microbiology, Mohammed VI University Hospital/Faculty of Medicine and Pharmacy, University Mohammed the First, Oujda, Morocco
[f]Emergency Department, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, University Mohammed V, Rabat, Morocco

2.3.1 Abstract

Many sequencing efforts from across the globe have revealed genetic diversity among clinical isolates. Here we present, WGS results of 9 clinical isolates of *M. tuberculosis* from Morocco.
We observed that the 9 *M. tuberculosis* clinical isolates have different levels of drug resistance and harboured different numbers of SNPs. The numbers of SNP was higher in the MDR strains compared to susceptible strains. All the analyzed strains belonged to the lineage L4. However, MDR strains clustered with LAM9 sub-lineage while susbtible strains are considered Harlem. A total of 39 mutations were identified in drug resistance-associated loci, including 14 previously reported ones and 25 newly identified. 7 novel mutations were identified in the proteins' active site. The docking analysis confirmed the phenotypic resistance effect of gyrA/gyrA mutations identified in MTB1.
This study highlights the significance of employing WGS in diagnosis and for monitoring MDR-TB strains.

2.3.2 Introduction

Mycobacterium tuberculosis is one of the most harmful human pathogens, ranked seven as the cause of global mortality and morbidity worldwide, responsible for nearly 1,4 M

deaths and10 million new cases in 2019 [1]. Considerable effort has been made to combat the spread of the germ. Still, the emergency of MDR-TB strains has been recently complicated, showing resistance to the second line drugs treatment [1]. In 2019, 558000 new cases with resistance to rifampicin, the most effective second-line drug, were registered, 82% were MDR strains with 6,1;with at least one XDR case reported in 123 case [1].
Morocco is not among the 20 counties with a severe MDR-TB burden[1]). However, the tab incidence still constant and slightly increased in the last three years, with one contaminated person in 800 people representing 31618 new cases per year [1] [2]. In 2017, WHO estimates that 1% of new cases and 11% of previously treated patients are MDR-TB [2]. New drugs must be developed to control antibiotic resistance,but most importantly, the existing drug must be prescribed wisely to prevent the extending of resistance associated with genomic mutations [3]. The conventional method for drug resistance identification is time-consuming and requires several weeks to detect bacterial growth. To overcomes the limitation [4], the WHO has recommended the implementation and use of molecular method for mutation detection. However, this technique doesn't cover all genes associated with drug resistance [5] [6].

Sequencing technologies provide a breakthrough in tuberculosis research [7] [8], opening the way to understand not only the resistance profile [9] but also the virulence [10], the transmission [11], and diversity of *Mycobacterium tuberculosis*. Only a few studies have investigated *Mycobacterium tuberculosis* diversity in Morocco, using 12-locus MIRU-VNTR typing only [12][13][14].
The development of sequencing technology, decreasing cost and time leads to an increase in the amount of pathogen genome sequences available in the public databases. Therefore, we used WGS to determine the drug resistance profile and genetic diversity of *Mycobacterium tuberculosis* strains from different parts of Morocco with the ultimate goal of improving the management and control of tuberculosis infection.

2.3.3 Methods

2.3.3.1 Sample culture

In this study, nine clinical strains were randomly isolated from patients with pulmonary tuberculosis at the military hospital in Rabat, Morocco. The septum sample was decontaminated by the standard N-acetyl L-cysteine (NALC)-NaOH method [15]. The pellet was reconstituted to 2,5 ml with phosphate buffer and inoculated on liquid and solid Lowenstein-Jensen (LJ) media and incubated at 37°C. Cultures were considered negative when no colonies were seen after eight weeks.

2.3.3.2 drug susceptibility testing

The clinical samples were cultivated in the Bactec system on Lowenstein-Jensen (L/J) culturemedia and on Growth Indicator Tubes (MGIT) 960 culture tubes. The first-line drug susceptibil-ity test was conducted using the 1% isoniazid (INH), rifampicin (RMP), streptomycin (SM) and ethambutol (EMB) proportion method on the L/J medium [16] and BACTEC MGIT 960 SIRE kit Becton Dickinson, CA, USA. Tow loopful of colonies from the solid media were transferred into 2 ml Eppendorf tubes containing 250 l of 1xTE buffer (10mM Tris-HCl pH: 8.0, 1mM EDTA pH:8.0) and heat-killed at 80°C for 50°C, then centrifuge. DNA purification was done using DNAeasy blood and tissue DNA extraction kit (Invitrogen per the manufacturer's instructions).

2.3.3.3 Whole-genome sequencing, assembly and annotation

Genomic libraries were prepared from 1ng/ul DNA according to Nextera XT kit (Illumina Inc,San Diego, CA, USA) and sequenced on a Miseq platform (at Medbiotech laboratory).

Theoverall quality of the sequence was checked using FastQC [17], Reads were assembled using A5pipeline [18] and SPAdes [19]. The protein-coding genes were predicted using Glimmer 3.02 [20].In contrast, tRNAs can-SE [21] and RNAmmer [22] were used to identify tRNA and rRNA,respectively.

The genome sequence was also uploaded into Rapid Annotation using Subsystem Technology (RAST) [23] for sequence annotation. The functions of predicted protein-coding genes were then annotated through COG databases [24] and KEGG [25].

2.3.3.4 Whole gneome alingment

The nine strains and 78 genomes downloaded from NCBI and ENA databases were mapped to *Mycobacterium tuberculosis H37rv* reference genome using BWA MEM (24). Picard V 2.18 was used for SAM-BAM conversion, sorting, and marking duplicated; local realignment was per-formed using GATK (25). Variant (SNP, Indels) were called using GATK (25) with customizing filtering options; only variants with QD≥ 2.0, $FS \geq 600forSNPandFS \geq 200forindels, MQ \geq 400, Sor \geq 40passedthefiltered.Theywere$

annotated using SNpEff (26). An accurate list of genomic polymorphisms that confer drug resistance to the eight drugs was made based mainly on Tb-Dream (27), CARD database (28), and literature review.

2.3.3.5 Phylogenetic analysis and Spoligotyping

Spoligotyping analysis of the nine strains was determined using PhyTB [26] and SpoTyping tools [27]. The phylogenetic classification was verified using previously described strains. The Spoligotype family determination was based on the international database SITVIT.

2.3.3.6 Mutation effect prediction

Mutant models of embC (PDB id: 3pty), embR (PDB id: 2fez), gyrA (PDB id: 5bs8), gyrB (PDBid: 5bs8), katG (PDB id: 1sj2), and rpoB (PDB id: 5uhb) were generated in the crystallographicobject-oriented toolkit software (COOT, version 0.8.9.2 EL for Windows) using the "mutate and autofit" function. Drugs and mutant structures were prepared in Autodock tools (ADT,version 1.5.7rc for Windows) [28] and exported in Autodock's PDBQT format for further usein the docking simulations.

Table 2.1: **Binding sites and boxes**

Structure	Binding site reference	Box parameters
GyrA, GyrB	Bound ligand (MFX)	center(36,20,8) ; size(20,20,20)
RpoB	Bound ligand(RFP)	center(164,162,19) ; size(20,20,20)
EmbR	Bibliography [106]	center(12,39,28) ; size(25,20,20)
EmbC	Bibliography [107]	center(95,10,06) ; size(20,20,20)
KatG	Bibliography [108]	center(37,-10,27) ; size(20,20,20)

Ligand-receptor docking was performed in Autodock vina (version1.1.2) using default options. Binding sites determination and box parameters areshown in **table 3.1**.

2.3.4 Results

The assemblies shared the same GC content of 60%, RNAt, CDS, and 3 RNAr operons. Other genomic characteristics like genome assembly, contigs, N50, CDS, number of genes are in **Table 3.2**. All genomes are matching the H37rv genome. The sequences were annotated with COG database to identify orthologues. A slight difference in COG proteins was observed in the nine strains compared to *Mycobacterium tuberculosis H37rv*, including genes associated with defense, transport, virulence, and cell signaling.

Furthermore, the comparative genomic analysis identified 4044 non-repetitive SNP among a shared pool of 396 mutations shared by all strains. All the information about SNP distribution is summarized in **Figure 3.1**. Drug-resistant strains shared 350 mutations, from which a total of 26 mutations were in drug resistance-associated loci. Including 15 already reported (Table S1 in Supplementary Material).

Table 2.2: **Genomic Features of 4 sequenced *Mycobacterium tuberculosis* strains**

Strains	MTB1	MTB2	MTB3	MTB8
Assembly	A5	A5	A5	A5
Annotation	NCBI	NCBI	NCBI	NCBI
Size	4341655	4291403	4291580	4315169
Contigs	242	238	264	223
N50	43730	35128	33786	57279
CDS	4265	4422	4390	4383

Seven out of the twelve nonsynonymous mutations were previously reported in the literature. The analysis of SNP clustering density showed a nonrandom distribution of SNP's. We further analyzed SNP according to the COG category that shows that the majority of SNP are distributed in genes with unknown function, flowed by genes belonging to Cell wall/membrane/envelope biogenesis (M), General function ®, Defense mechanisms (V). In contrast, genes in secondary metabolism and transport show a very low SNP density.more that half of insertion, deletion and mutation were presente in PPE genes representing a major source of variability (Table S2 in Supplementary Material).

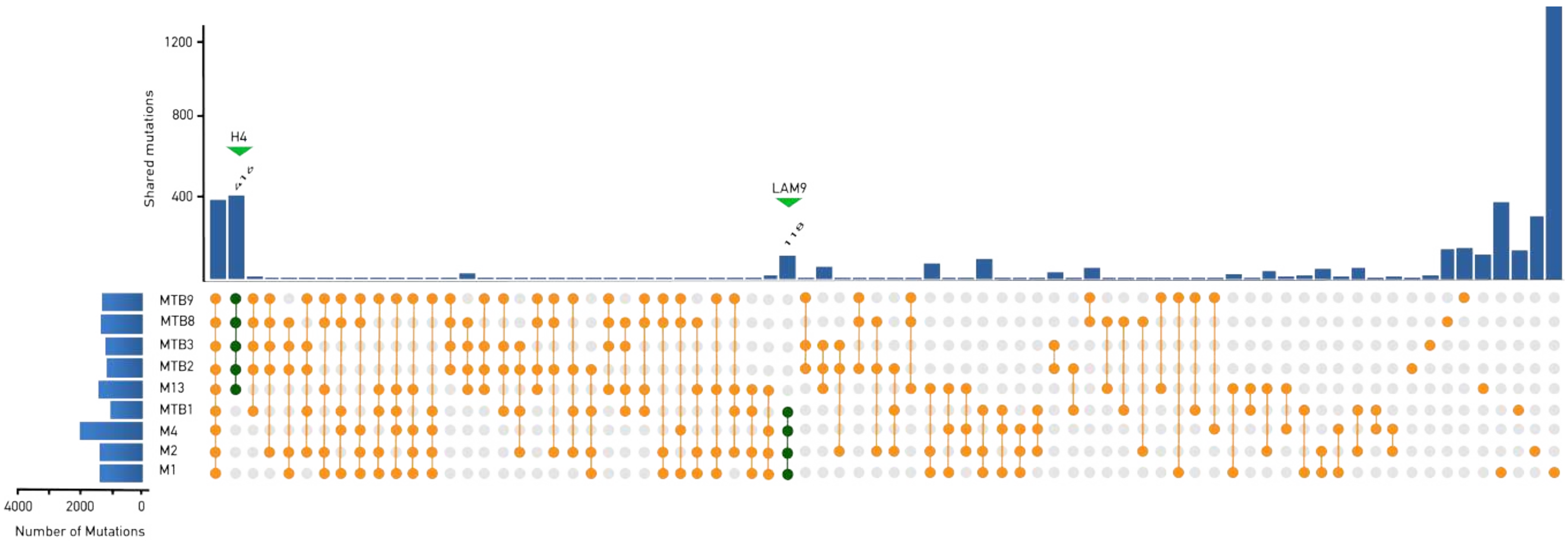

Figure 2.1: **Distribution of single nucleotide polymorphisms in the 9 *M. tuberculosis.*** **(A)** correspond to the number of mutation in each strain. **(B)** corresponds to the number of mutation shared between the compared genomes. The Green color corresponds to different mutation shared by the different sub-lineages. M1 and M2 have the highest number of unique SNP's 500 and 360 unique SNPs respectively. Strain belonging to LAM group have 119 SNPs in common, Haarlem have 416 SNPs in common, while M4 have more than 1200 unique SNPs which may be due to the bad quality of the sequencing. 387 SNPs are common to all isolates

2.3.4.1 Whole-genome SNV phylogenetic tree and spolygotype

The nine strains were incorporated into a phylogeny analysis that included the 67 lineage-defined isolates. The resulting tree shows that the nine strains and lineage-defined isolates clustered into lineages 2, 3, 4, and 5 **(Figure 3.2)**. Lineages 1, 6,7, and 8 are not shown.

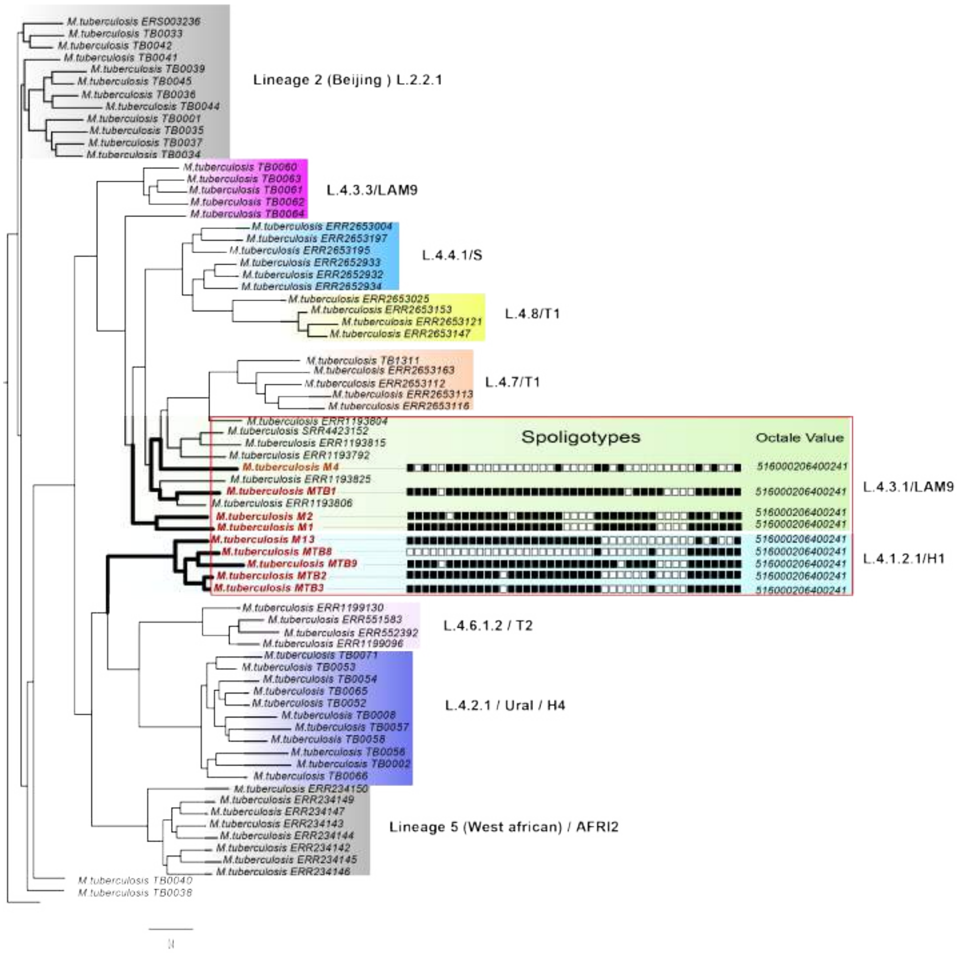

Figure 2.2: **Phylogenetic relationships of *M. tuberculosis* isolates based on SNPs from whole genome sequences** Phylogenetic tree of *Mycobacterium tuberculosis* lineages strains comprising the 9 isolates from Morocco and 65 from the public database. A phylogenetic tree based on concatenated SNVs from all the strains (the 9 isolates from Morocco and 65 from the public database) was constructed using RAxML. *M. canettii* was used as an out-group. The sub lineages are labelled on the right, were named sequentially according to their positions in the tree. Octal code (in red: defining octal rule). No bootstrap support is shown for the MTBC phylogeny as all the case were higher than 95%.

All of the nine strains clustered with the lineage 4 isolates with a low lineage diversity. Our analysis showed that lineage 4 is the most prevalent in Morocco since all the isolates belonged to this lineage. The sublineage distribution of lineage 4 included: 4 LAM strains (MA, M2, MTB1, M4) in green , 5 Haarlem (M13, MTB3, MTB8, MTB9, MTB2) in red. According to the SIVITI database, all the LAM strains belonged to the LAM9 spoligotype while the Harlem one belongs to H1. Only two strains in H1 has the same spoligotype profile, MTB2 and MTB3. MTB1 was most closely related to 5 strains isolated in Switzerland and TB SSR4423152 from Bangladesh. Surprisingly, all the 4 MDR strains belong to the Euro-American LAM9 sublineage. In contrast, all the susceptible strains to all drug resistance belong to the Haarlem clone H1.

2.3.4.2 Dug resistance associated mutations

In this study, we classified the isolates according to their resistance profile. From the nine strains, we identified five susceptible strains that didn't show resistance to any drug in the phenotypic tests (MTB2, MTB3, M13, MTB8 and M20), three MDR strains resistant to Rifampicin and Iszodiazid (M1, M2, M4), and one pres-XDR strain resistant to Rifampicin, Iszodiazid, and Streptomycin (MTB1). These mutations were already described in the literature as resistance-conferring mutations. A total of 39 types of mutations were identified in drug resistance-associated loci, including 14 previously reported and 25 newly identified **(Figure 3.3)**. The phenotypic results were further verified by genotypic test for the second-line drugs using the HAIN genotype; the tests showed mutations in the hotspot genes tested. MTB1, M2 and M1 exhibited a high level of resistance to isoniazid, mainly due to mutations in katG (Ser315Thr), while M4 harboured the mutations -15 at the *inh* promoter, which confers a moderated resistance $\geq 0; 25. Allisoniazid - resistantisolateshadasinglegenemutationexceptforM4, whichharbouredTowmutations.$

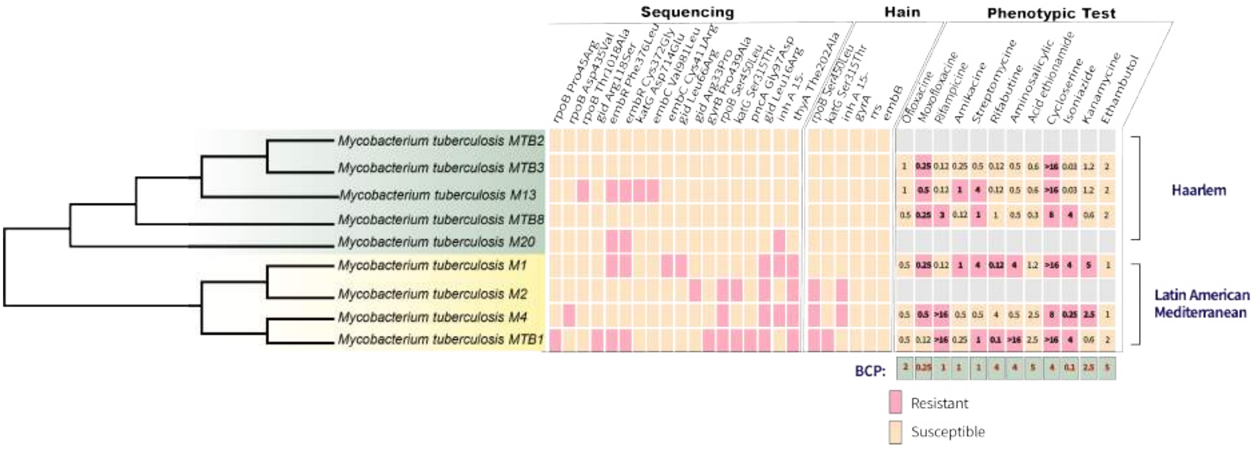

Figure 2.3: **Phylogenetic clustering of the 9 strains with a comparaison of drug resitance using sequencing and hain genotyping and phenotypic tests**. Phylogenetic tree was constructed using Maximum likehood method with MEGA 7.0. Bootstrap values over 95%. the tabe at the right represent the resistance profile for each strain using the different techniques, the blocks on pink and yellow represent the presence or absence of Resistance respectively. The Haarlem and LAM show different resistance profile.

Some results of rifampicin resistance differed between phenotype and genotype tests. In contrast, all the rifampicin resistance isolates harboured the high-level resistance mutations rpoB Ser450Leu ($\geq 16mg)detectedbyHain.$

MTB 8 showed a moderated resistance level to Rifampicin phenotypically with no known mutations detected in the associated genes. Predictions for resistance to pyrazinamide was associated with a mutation in pncA (Gly97Asp) found only in MTB1. Phenotypic resistance to streptomycin was associated with mutations in *gid* gene harboured by M1, MTB1 and MTB8. All the analysed strains lack the common mutation associated with resistance to fluoquinolone. However, M4 and MTB8 showed moderate resistance in the phenotypic test. They are suggesting either the presence of other mutations in *gyrA* and *gyrB* genes or another mechanism of resistance.

2.3.4.3 In silico docking analysis.

The impact of the selected mutations on the structure of RpoB, GyrA and GyrB, EmbC, and EmbC the binding energy and affinity of ligands for the proteins variants were calculated and are shown in Table 3. All of the mutations were in or close to the active site of the protein. overall, Val981Leu (EmbC), "Phe376Leu Cys372Gly" (EmbR), Asp714Glu (KatG) and Thr1018Ala (RpoB) mutations showed a limited effect on the proteins. The WT had binding energy of -4,2 Kcal/mol (EmbC), -4,0 Kcal/mol (EmbR) -5,9 Kcal/-mol(KatG) -10,1 Kcal/mol(RpoB) while the mutatants showed a similar or a slightly higher biding affinity, which may increase the protein-ligand binding and make the strain more suscptible **(Table 3.3)**.

Table 2.3: Effect of the novel mutation on the binding affinity with the ligands

Protein	**Mutation**	**Antibioteics**	**W.T**	**M.T**	**DDG**	**Effect**
EmbC	Val 981 Leu	Ethambutol	-4,2	-4.3	-0,1	Increase
EmbR	Phe 376 Leu Cys 372 Gly	Ethambutol	-4	-4.2	-0,2	Increase
GyrAB	Pro 439 Ala	Ciprofloxacin	-17,5	-18	-0,5	Increase
GyrAB	Pro 439 Ala	Gatifloxacin	-18,2	-17,8	0,6	Decrease
GyrAB	Pro 439 Ala	Levofloxacin	-19,1	-15,6	3,5	Decrease
GyrAB	Pro 439 Ala	Moxifloxacin	-18	-14,2	3,8	Decrease
KatG	Asp 714 Glu	Isoniazid	-5,9	-5,7	0,2	Decrease
RpoB	Thr 1018 Ala	Rifampicin	-10,1	-10,1	0	
RpoB	Asp 435 Val	Rifampicin	-10,1	-8,9	1,2	Decrease

Mutations Asp435Val in RpoB displayed a lower rifampicin affinity than the wt protein -8.9 Kcal/mol and -10,1 Kcal/mol respectively. Similarly, mutation in Pro439Ala (GyrB) shown to be effective in lowering the affinity of gyrA/B for levofloxacin and Moxifloxacin with a $\triangle\triangle Gdifferenceof-3,5Kcal/moland-3,8Kcal/molforLevofloxacinandMoxifloxacinrespec$

Meanwhile, Pro439Ala (GyrB) mutations increased the binding affinity with ciprofloxacine

.

2.3.5 Dicussion

To study the genomic variability and defined its classification among the TB sublineages. Genetic mutations have an important impact on MTC virulence and pathogenicity. In this study, the genomic variation was not dependent on the resistance profile since all strains shared more than 360 SNP's. However, uniq SNP's in MDR-TB and susceptible strains were also identified. The difference in SNP number between MDR and susceptible strains is indenaiable. MDR strains showed higher genomic diersity. The relatively high difference of SNP number between both groups could be partially caused by their belonging to different sub-lineages or natural variation, as we included patients diversely located.

The SNP distribution was not random since the majority of SNP's were clustered in regions with unknown functions (32), suggesting that those genes tend to accumulate SNP's as it was previously demonstrated (33). PE PPE multigene family harbord a significantly higher number of SNP than the whole genome. These genes evolve by specific duplication events usually encoded in bicistronic operons with conserved structure and repeat playing a role in virulence and evasion to the host immune system (34).

Sixty tow SNP variation were found in all strains and belonged to the PE PPE gens family. The majority of SNP represented silent mutation and clustered in CPGRS 4 , 3 , 27 . Mutations in PPE-62was observed exclusively in the Haarlem group and were recently described to play a role on bacterial binding during the invasion and may contribute to the antigenic variation. mutations were also identified in PE-PGRS 28 (MTB1). This gene has already been described to be implicated in the modulation of vacuole acidification in *Mycobacteirum bovis*(35). More interestingly, MTB1 Pre-XDR harboured a high proportion of polymorphism in this genes family, which could be considered as a fitness mechanism.

Multiple deletions and insertions were found in the nine strains, most in genes playing roles in transport, as sdhA a potassium transfering protein, mull9 gene which plays an essential role in oxidative stress response (37). Deletion also accrued in ppsA and has already been demonstrated to affect the susceptibility to antibiotics (38).
The indel distribution was independent of the resistance profile of the nine strains. Many indels associated with virulence mechanism were found in both groups, suggesting that drug resistance is not necessarily an indicator of increased virulence. Whole-genome SNV analysis was conducted to construct the phylogenetic tree. Our results report the lineages circulation in Morocco and confirm previous studies realized by Chaoui et al. (6).

In our collection of isolates, we identified sublineages circulating in Morocco. Our finding reveals that all the isolates belonged to the L4, called Euro-American lineage. The strains were grouped in a single cluster divided into two different clades, Harleem and LAM9. This lineage has a significant presence in North Africa and was previously described in Tunisia and Algeria (11). A possible explanation for its presence is the significant immigration from Europe to Morocco in the last 30 years.

We have combined the power of whole-genome sequencing in association with genotyping and phenotyping to get insight into tb drug resistance in Morocco. A large number of mutations related to drug resistance were already identified worldwide. In this study, the whole genome sequencing identified well establish resistance-conferring loci. Isoniazid resistance in MDR TB was mostly conferred by a mutation in katG gene that encodes for catalase peroxidase. This mutation is considered the most frequent isoniazid resistance-associated mutation worldwide (40), consistent with recent studies that highlight their pivotal role in the emergence of INH resistance (41). However, in the absence of katG mutation in M4, isoniazid resistance was conferred by a mi sense mutation in fabG-inhA operon promoter, as already demonstrated (42). These two mutations could emerge before MDR-Tb, as it was the case in Lisbon strains and shown in our results of MTB13 and M1. A portion of samples exhibited discrepancy in drug resistance determination between Phenotipic vs WGS results, most having "susceptible" results by WGS despite having been determined "resistant" by MGIT. This discrepancy could be due to technical or biological causes. For example, several pDST-determined samples do not carry SNPs at the gid locus, so resistance may be due to secondary loci which have not been identified as having a strongly significant effect of streptomycin drug resistance such as the case of M13 and in intermediated resistance in MTB8.

Fluoroquinolones are a class of antibiotics, which inhibit DNA gyrase and thus prevent bacterial DNA synthesis. They play an important role in MDR-TB treatment, and several bacterial infections. In this analysis no XDR strains were identified. However, priliminary resistance to Fluoroquinolones were observed in phenotypic test with no known mutations in second line associated loci. the Increasing fluoroquinolone prescriptions for many infections has resulted in emergence of fluoroquinolone resistance. Global surveillance studies demonstrate that fluoroquinolone resistance rates increased in the past years in almost all bacterial species.

Our study identified several novel genetic variations that were not previously reported in *M. tuberculosis*. Among these were several novel genetic variations in genes that are known to confer drug resistance. Novel mutations were also identified in the five susceptible strains on resistance-associated genes, suggesting that these mutations do not affect the resistance pattern.
Furthermore, we found the mutation in Rpob at position 45 in association with mutation in 450 showed a slight decrease of the binding affinity to Rifampicin, suggesting a potential resistance effect of this mutation on the bacteria.
Additional mutations in *rpoB* may trigger compensatory transcriptional changes in secondary metabolism genes analogous to those observed in related actinobacteria.
Similarly, new mutation in *gyrA/gyrB* increase the resistance to levofloxacin and ofloxacin in phenotypic tests and was proven to increase the binding affinity. As novel SNVs may have direct implications on drug resistance; country specific probes would be needed for rapid and effective diagnosis and treatment of drug resistant TB.

2.3.6 Conclusion

We examined the phylogenetic and drug-resistance properties of *M.tuberculosis* isolates collected from 9 Moroccan patients. Our analyses confirmed the phylogenetic separation of pathogenic *M. tuberculosis* strains and support the prevalence of Lineage 4, showing high MDR levels among LAM9. We identified known and novel genetic determinants that could impact bacterial virulence, pathogenicity, and drug resistance. However, the sensitivity of direct whole-genome sequencing remains low compared with culturing.

2.3.7 Reference

1. Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, TerrenceDisz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth MGlass, Michael Kubal, et al. The rast server: rapid annotations usingsubsystems technology.BMC genomics, 9(1):1–15, 2008.

2. Ernest D Benavente, Francesc Coll, Nick Furnham, Ruth McNerney, Ju-dith R Glynn, Susana Campino, Arnab Pain, Fady R Mohareb, andTaane G Clark. Phytb: Phylogenetic tree visualisation and sample po-sitioning for m. tuberculosis.BMC bioinformatics, 16(1):1–5, 2015.

3. Nada Bouklata, Philip Supply, Sanae Jaouhari, Reda Charof, FouadSeghrouchni, Khalid Sadki, Youness El Achhab, Chakib Nejjari, Abdelka-rim Filali-Maltouf, Ouafae Lahlou, et al. Molecular typing of *mycobacterium tuberculosis* complex by 24-locus based miru-vntr typing in con-junction with spoligotyping to assess genetic diversity of strains circulatingin morocco.PloS one, 10(8):e0135695, 2015.

4. Amanda C Brown, Josephine M Bryant, Katja Einer-Jensen, Jolyon Hold-stock, Darren T Houniet, Jacqueline ZM Chan, Daniel P Depledge, Vla-dyslav Nikolayevskyy, Agnieszka Broda, Madeline J Stone, et al. Rapidwhole-genome sequencing of *mycobacterium tuberculosis* isolates directlyfrom clinical samples.Journal of clinical microbiology, 53(7):2230–2237,2015.

5. Haley A Brown, Evgeny Vinogradov, Michel Gilbert, and Hazel M Holden.The *mycobacterium tuberculosis* complex has a pathway for the biosyn-thesis of 4-formamido-4, 6-dideoxy-d-glucose.Protein Science, 27(8):1491–1497, 2018.

6. Imane Chaoui, Thierry Zozio, Ouafae Lahlou, Radia Sabouni, MohammedAbid, Rajae El Aouad, Mohammed Akrim, Said Amzazi, Nalin Rastogi,and Mohammed El Mzibri. Contribution of spoligotyping and miru-vntrs to7 characterize prevalent *mycobacterium tuberculosis* genotypes infecting tu-berculosis patients in morocco.Infection, Genetics and Evolution, 21:463–471, 2014.

7. David Coil, Guillaume Jospin, and Aaron E Darling. A5-miseq: an up-dated pipeline to assemble microbial genomes from illumina miseq data.Bioinformatics, 31(4):587–589, 2015.

8. Mireia Coscolla and Sebastien Gagneux. Consequences of genomic diversityin *mycobacterium tuberculosis.* InSeminars in immunology, volume 26,pages 431–444. Elsevier, 2014.

9. Christophe Demay, Benjamin Liens, Thomas Burguiere, Veronique Hill,David Couvin, Julie Millet, Igor Mokrousov, Christophe Sola, ThierryZozio, and Nalin Rastogi. Sitvitweb–a publicly available internationalmultimarker database for studying

mycobacterium tuberculosis geneticdiversity and molecular epidemiology.Infection, genetics and evolution,12(4):755–766, 2012.

10. Wifak Ennassiri, Sanae Jaouhari, Radia Sabouni, Wafa Cherki, RedaCharof, Abdelkarim Filali-Maltouf, and Ouafae Lahlou. Analysis of iso-niazid and rifampicin resistance in *mycobacterium tuberculosis* isolates inmorocco using genotype mtbdrplus assay.Journal of global antimicrobialresistance, 12:197–201, 2018.

11. Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genesand genomes.Nucleic acids research, 28(1):27–30, 2000.

12. Benson R Kidenya, Stephen E Mshana, Daniel W Fitzgerald, and OksanaOcheretina. Genotypic drug resistance using whole-genome sequencing of *mycobacterium tuberculosis* clinical isolates from north-western tanzania.Tuberculosis, 109:97–101, 2018.

13. Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stær-feldt, Torbjørn Rognes, and David W Ussery.Rnammer: consistentand rapid annotation of ribosomal rna genes.Nucleic acids research,35(9):3100–3108, 2007.

14. Ouafae Lahlou, Julie Millet, Imane Chaoui, Radia Sabouni, AbdelkarimFilali-Maltouf, Mohammed Akrim, Mohammed El Mzibri, Nalin Rastogi,and Rajae El Aouad. The genotypic population structure of *mycobacterium tuberculosis* complex from moroccan patients reveals a predominance ofeuro-american lineages.PLoS One, 7(10):e47113, 2012.

15. Robyn S Lee and Madhukar Pai.Real-time sequencing of *mycobacterium tuberculosis*: are we there yet?Journal of clinical microbiology,55(5):1249–1254, 2017.8

16. Qiao Liu, Dandan Yang, Weiguo Xu, Jianming Wang, LV Bing, Yan Shao,Honghuan Song, Guoli Li, Haiyan Dong, Kanglin Wan, et al. Molecular typ-ing of *mycobacterium tuberculosis* isolates circulating in jiangsu province,china.BMC infectious diseases, 11(1):1–10, 2011.

17. Todd M Lowe and Patricia P Chan. trnascan-se on-line: integrating searchand context for analysis of transfer rna genes.Nucleic acids research,44(W1):W54–W57, 2016.

18. Sergey Nurk, Anton Bankevich, Dmitry Antipov, Alexey Gurevich, An-ton Korobeynikov, Alla Lapidus, Andrey Prjibelsky, Alexey Pyshkin,Alexander Sirotkin, Yakov Sirotkin, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In Annual International Con-ference on Research in Computational Molecular Biology, pages 158–170.Springer, 2013.

19. World Health Organization et al. Global tuberculosis report 2020: execu-tive summary. 2020.

20. Lee W Riley. Detection of drug resistance in *mycobacterium tuberculosis*.Germs, 6(1):7, 2016.

21. S Bittencourt a S. A. Fastqc: a quality control tool for high throughputsequence data. 2010.

22. Steven L Salzberg, Arthur L Delcher, Simon Kasif, and Owen White. Mi-crobial gene identification using interpolated markov models.Nucleic acidsresearch, 26(2):544–548, 1998.

23. D.M. Shlaes, S.J. Projan, and J.E. Jr. Antibiotic discovery: State of thestate.ASM News, 70:275–281, 06 2004.

24. SO Simons, T Van der Laan, R De Zwaan, M Kamst, J Van Ingen, PNRDekhuijzen, MJ Boeree, and D van Soolingen. Molecular drug suscepti-bility testing in the netherlands: performance of the mtbdrplus and mtb-drsl assays.The international journal of tuberculosis and lung disease,19(7):828–833, 2015.

25. Karen R Steingart, Hojoon Sohn, Ian Schiller, Lorie A Kloda, Catharina CBoehme, Madhukar Pai, and Nandini Dendukuri. Xpert®mtb/rif assayfor pulmonary tuberculosis and rifampicin resistance in adults.Cochranedatabase of systematic reviews, (1), 2013.

26. Roman L Tatusov, Michael Y Galperin, Darren A Natale, and Eugene VKoonin. The cog database: a tool for genome-scale analysis of proteinfunctions and evolution.Nucleic acids research, 28(1):33–36, 2000.

27. Eryu Xia, Yik-Ying Teo, and Rick Twee-Hee Ong. Spotyping: fast and ac-curate in silico mycobacterium spoligotyping from sequence reads.Genomemedicine, 8(1):1–9, 2016.9

28. Matteo Zignol and Katherine Floyd. Resistance of *mycobacterium tuberculosis* isolates to pyrazinamide and fluoroquinolones–authors' reply.TheLancet Infectious Diseases, 17(1):25, 2017

2.4 Discussion

Direct whole-genome sequencing could be used to study transmission clusters of *M.tuberculosis* and conduct culture-independent surveillance. Compared with conventional approaches, direct whole-genome sequencing allows researchers to do real-time genomic epidemiology and drug resistance surveillance in settings where culture and drug susceptibility testing are not available.

This study highlights the significance of employing WGS in diagnosis and for monitoring MDR-TB. We examined 9 *M. tuberculosis* isolates for published variants associated with resistance to TB drugs population structure. Evidence presented here suggests the importance of lineage identification towards drug resistance in Mtb, showing a higher genomic diversity in MDR strain which all belonged to the LAM9.

A total of 39 mutations were identified in drug resistance-associated loci, including 14 previously reported ones and 25 newly identified ones. 7 novel mutations were identified in the proteins' active site. Sixty tow SNP variation were found in all strains and belonged to the PE PPE gens family. It has been suggested that the PE/PPE gene family encodes virulence factors and are a possible source of antigenic variation influencing immune evasion16.

Our analysis of WGS data identified known and novel genetic determinants that could influence bacterial virulence, pathogenicity, and drug resistance. It will be interesting to see how methodologies outlined here might apply to other pathogen species in an antimicrobial resistance context or indeed in relation to other phenotypes of interest, such as transmissibility.

Chapter 3

Characterization of the Genetic diversity of SARS-CoV-2 strains in Morocco

3.1 Preface

SARS-CoV-2 was first reported in China in 2019 and soon spread to the rest of the world. Due to the increasing number of death during the first wave worldwide, the SARS-CoV-2 virus soon became the focus of numerous sequencing studies and public health initiatives. Our lab's work on SARS-CoV-2 surveillance and investigation in Morocco. We released 15 SARS-CoV-2 genomes sequences of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains. Obtained from nasopharyngeal swabs from Moroccan patients with COVID-19.

We performed an in-depth analysis of forty- eight Moroccan strains of SARS-CoV-2 collected from mid-March to the end of May and identified three major haplotypes (H1, H2, H3) circulating within the country. Likewise, the phylogenetic analysis revealed that these Moroccan strains were closely related to those belonging to the five continents, indicating no specific strain dominating in Morocco.

These findings have the potential to lead to new comprehensive investigations combining genomics data, epidemiological information, and clinical characteristics of SARS-CoV-2 patients in Morocco and could indicate that the developed vaccines are likely to be effective against Moroccan strains.

2 Genome Sequences of Six SARS-CoV-2 Strains Isolated in Morocco, Obtained Using Oxford Nanopore MinION Technology

Meriem Laamarti,[a] M. W. Chemao-Elfihri,[a] Souad Kartti,[a] Rokia Laamarti,[b] Loubna Allam,[a] Mouna Ouadghiri,[a] Imane Smyej,[c] Jalila Rahoui,[c] Houda Benrahma,[c] Idrissa Diawara,[c] Tarek Alouane,[a] Abdelomunim Essabbar,[a] Samir Siah,[i] Mohammed Karra,[a] Naima El Hafidi,[a] Rachid El Jaoudi,[a] Laila Sbabou,[d] Chakib Nejjari,[e] Saaid Amzazi,[f] Rachid Mentag,[g] Lahcen Belyamani,[h] Azeddine Ibrahimi[a]

[a]Medical Biotechnology Laboratory (MedBiotech), Bioinova Research Center, Rabat Medical and Pharmacy School, Mohammed V University in Rabat, Rabat, Morocco
[b]Medical Biotechnology Center, Moroccan Foundation for Science, Innovation & Research (MAScIR), Rabat, Morocco
[c]National Reference Laboratory, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
[d]Research Center of Plants and Microbial Biotechnologies, Biodiversity and Environment, Microbiology and Molecular Biology Team, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco
[e]International School of Public Health, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
[f]Laboratory of Human Pathologies Biology, Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco
[g]Biotechnology Unit, Regional Center of Agricultural Research of Rabat, National Institute of Agricultural Research, Rabat, Morocco
[h]Emergency Department, Military Hospital Mohammed V, Rabat Medical and Pharmacy School, Mohammed V University in Rabat, Rabat, Morocco
[i]Department of Burns, Mohammed V Military Teaching Hospital/Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, Rabat, Morocco

2.1 ABSTRACT

Here, we report the draft genome sequences of six severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) strains. SARS-CoV-2 is responsible for the COVID-19 pandemic, which started at the end of 2019 in Wuhan, China. The isolates were obtained from nasopharyngeal swabs from Moroccan patients with COVID-19. Mutation analysis revealed the presence of the spike D614G mutation in all six genomes, which is widely present in several genomes around the world.

2.2 ANNOUNCEMENT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is classified within the subgenus Sarbecovirus and genus Betacoronavirus and was first identified in Wuhan, China (1), as the causative agent for COVID-19 disease. Since then, the number of COVID-19 cases has risen dramatically (2).

In Morocco, the first SARS-CoV-2 case was confirmed on 2 March 2020. As of 29 June 2020, the number of cases had reached more than 12,248. To understand SARS-CoV-2 genetic diversity and molecular epidemiology in Morocco, we performed complete genome sequencing using the Oxford Nanopore MinION technology. In this study, we announce the genome sequences of six SARS-CoV-2 strains isolated from patients in Morocco. The samples were obtained by taking nasopharyngeal swabs from six patients with COVID-19. The viral RNA was extracted directly from the swab using the QIAamp viral RNA minikit (Qiagen, Germany), and the Transcriptor first-strand cDNA synthesis kit (Roche) with random hexamers was used to synthetize the viral cDNA. The ARTIC v3 primers were used with the Q5 high-fidelity DNA polymerase (New England BioLabs [NEB], USA) for virus DNA enrichment. Amplicons of 400 bp were purified using sample purification beads (SPBs) (Illumina, USA) (3)and then quantified with a Qubit 3.0 fluorometer and used for library preparation.

Sequencing was performed on a MinION MK1C instrument with a ligation sequencing kit (catalog number SQK-LSK109) according to a standard protocol (Oxford Nanopore Technologies [ONT], UK), and the six samples were multiplexed in one run. The R9 flow cell was used and run for 2 h.
The sequence reads generated were between 70,565 and 185,364 (Table 1) of raw data per sample, with average lengths of 454 bp, 455 bp, 455 bp, 452 bp, 455 bp, and 454 bp for strains RMPS-01, RMPS-02, RMPS-03, RMPS-04, RMPS-05, and RMPS-06, respectively. Raw reads were mapped to a SARS-CoV-2 reference genome under GenBank accession number NC_045512.2 using BWA-MEM v. 0.7.17 for single-end reads with default settings (4), and SAM/BAM files were manipulated by SAMtools v. 1.9.11 (5). Variant calling was performed using BCFtools v. 1.9 with "mpileup" (5), and variants were further annotated using SnpEff v. 4.3T (6). The consensus sequences were generated by mapping the variants to the reference genomes using BCFtools (5) and then were submitted to the GISAID database and NCBI (accession numbers are listed in **Table 1**).

The phylogenetic analysis was realized using 250 genome sequences retrieved from the GISAID database. The alignment was performed using MAFFT (7) for fast alignment, and maximum-likelihood trees were inferred with IQ-TREE v. 1.5.5 under the generalized time-reversable (GTR) model (8), implemented via the pipeline provided by Augur (github.com/nextstrain/augur). The generated tree was visualized using FigTree 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree). Major clades were defined by amino acid and/or nucleotide substitutions and were matched to the Nextstrain nomenclature (9) (https://nextstrain.org/ncov).

The size of the consensus sequences was similar to that of the Wuhan-Hu-1 reference genome (GenBank accession number NC_045512.2) and was 29,903 bp with a mean coverage ranging from 843.5× to 2,213×. The strain details are found in **Table 1**.

Table 3.1: **Genome features of six strains of SARS-CoV-2**

Strains	Number of raw reads	Genomes size bp	GC %	Coverage	(*)Mapped read %
RMPS-01	71570	29,903	37.96	870.2x	99.93%
RMPS-02	185364	29,903	37.96	2213.9x	98.92%
RMPS-03	88452	29,903	37.96	1022.5x	96.51%
RMPS-04	113813	29,903	37.96	1291x	94.9%
RMPS-05	70565	29,903	37.96	834.5x	98.2%
RMPS-06	128615	29,903	37.96	1291x	94.9%

We detected 16 different variants in the 6 analyzed genomes. All the genomes shared four mutations, namely, two synonymous (F924F and L4715L), one nonsynonymous (D614G), and one intergenic (241C¿T). Only one nonsynonymous mutation was detected (D614G) in the spike protein, which is known as the most prevalent variant worldwide (10), and it is also associated with the emergence of clade A2, which includes all Moroccan strains sequenced in this study **(Figure 2.1)**.

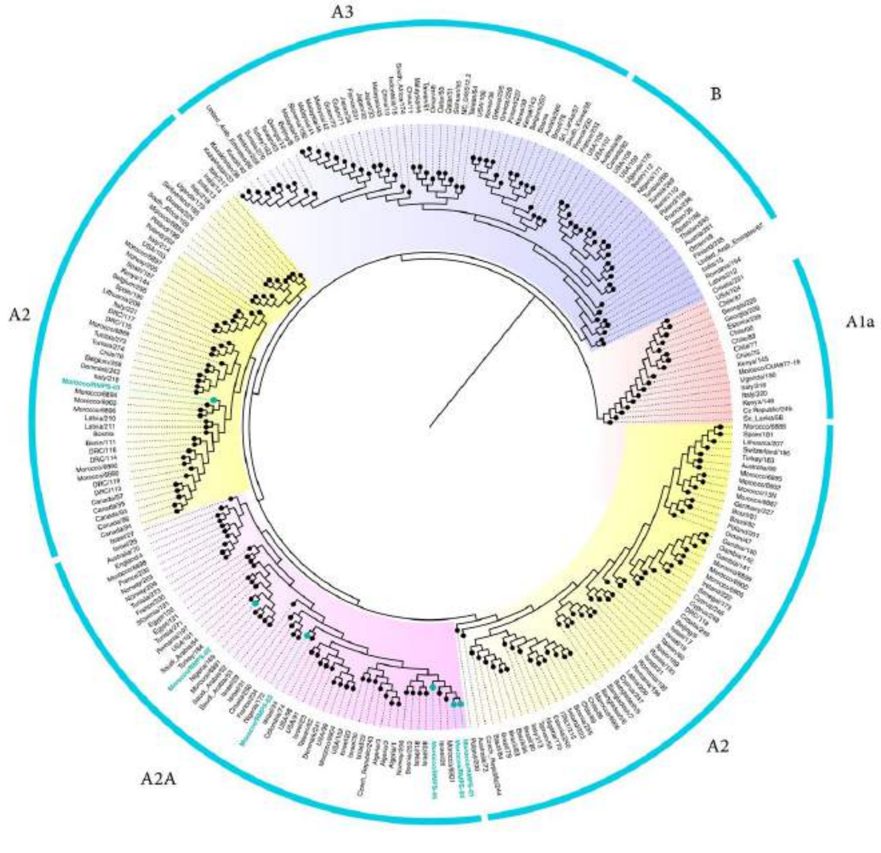

Figure 3.1: **Phylogenetic tree of six SARS-CoV-2 genomes from Morocco.** We are currently sequencing more genomes from Morocco to further investigate the spread of COVID-19 and to monitor the evolution of SARS-CoV-2 in Morocco.

This mutation was already associated with the observed transmission increase in the United States (10-12).

Data availability. The reads of the six SARS-CoV-2 strains were deposited in DDB-J/ENA/GenBank under the SRA accession numbers **SRR12109250, SRR12109251**, **SRR12109252, SRR12109253, SRR12109254**, and **SRR12109255**. The consensus sequences were also deposited in GenBank under the accession numbers **MT731285, MT731292, MT731673, MT731327, MT731468**, and **MT731764**.

2.3 Reference

1. Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song,Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirusfrom patients with pneumonia in china, 2019.New England journal of medicine,2020.

2. Tung Phan. Genetic diversity and evolution of sars-cov-2.Infection, genetics andevolution, 81:104260, 2020.

3. Josh Quick. ncov-2019 sequencing protocol.Protocols. io, 2020.

4. Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform.bioinformatics, 25(14):1754–1760, 2009.

5. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Ga-bor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/mapformat and samtools.Bioinformatics, 25(16):2078–2079, 2009.

6. Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, LuanWang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotatingand predicting the effects of single nucleotide polymorphisms, snpeff: Snps in thegenome of drosophila melanogaster strain w1118; iso-2; iso-3.Fly, 6(2):80–92, 2012.

7. Kazutaka Katoh, Kei-ichi Kuma, Hiroyuki Toh, and Takashi Miyata. Mafft version5: improvement in accuracy of multiple sequence alignment.Nucleic acids research,33(2):511–518, 2005.

8. Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh.Iqtree: a fast and effective stochastic algorithm for estimating maximum-likelihoodphylogenies.Molec biology and evolution, 32(1):268–274, 2015.

9. James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Pot-ter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher.Nextstrain: real-time tracking of pathogen evolution.Bioinformatics, 34(23):4121–4123, 2018.6

10. Lizhou Zhang, Cody B Jackson, Huihui Mou, Amrita Ojha, Erumbi S Rangarajan,Tina Izard, Michael Farzan, and Hyeryunc Choe. The d614g mutation in the sars-cov-2 spike protein reduces s1 shedding and increases infectivity.BioRxiv, 2020.

11. Tarek Alouane, Meriem Laamarti, Abdelomunim Essabbar, Mohammed Hakmi,El Mehdi Bouricha, MW Chemao-Elfihri, Souad Kartti, Nasma Boumajdi, HoudaBendani, Rokia Laamarti, et al. Genomic diversity and hotspot mutations in 30,983sars-cov-2 genomes: moving toward a universal vaccine for the "confined virus"?Pathogens, 9(10):829, 2020.

12. Meriem Laamarti, Tarek Alouane, Souad Kartti, MW Chemao-Elfihri, Mohammed-Hakmi, Abdelomunim Essabbar, Mohamed Laamarti, Haitam Hlali, Houda Bendani, Nassma Boumajdi, et al. Large scale genomic analysis of 3067 sars-cov-2genomes reveals a clonal geo-distribution and a rich genetic variations of hotspotsmutations.Plos one, 15(11):e0240345, 2020

3 Do the Moroccan SARS-CoV-2 genetic diversity hamper the use of the developed universal vaccines in Morocco?

M.Laamarti[a], A. ESSABBAR[b], T. Alouane[a], S. Kartti[a], N. BOUMAJDI[a], H. BENDANI[a], R. LAAMARTI[a], L. ALLAM[a], F. GHRIFI[a], I. SMYEJ[b], J. RAHOUI[b], H. BENRAHMA[b], L. DIAWARA[b], T. AANNIZ[a],N. EL HAFIDI[a],R. EL JAOUDI[a],C. NEJJARI[c],S. AMZAZI[d],R. MENTAG[e],L. BELYAMANI[f], A. Ibrahimi [a]

[a] Biotechnology Lab (MedBiotech Center), Rabat Medical and Pharmacy School, University Mohammed V, Rabat, Morocco
[b] National Reference Laboratory, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
[c] nternational School of Public Health, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
[d]Laboratory of Human Pathologies Biology, Faculty of Sciences, Mohammed V University in Rabat, Morocco
[e] Biotechnology Unit, Regional Center of Agricultural Research of Rabat, National Institute of Agricultural Research, Rabat, Morocco
[f]Emergency Department, Military Hospital Mohammed V, Rabat Medical and Pharmacy School, Mohammed Vth University in Rabat, Morocco

3.1 Abstract

The SARS-CoV-2 identified as coronavirus species associated with severe acute respiratory syndrome. At the time of writing, the genetic diversity of Moroccan strains of SARS-CoV-2 is poorly documented.

The present study aims to analyze and identify the genetic variants of forty- eight Moroccan strains of SARS-CoV-2 collected from mid-March to the end of May and the prediction of their possible sources. Our results revealed 108 mutations in Moroccan SARS-CoV-2, 50% were non-synonymous were present in seven genes (S, M, N, E, ORF1ab, ORF3a, and ORF8) with variable frequencies. Remarkably, eight non-synonymous mutations were predicted to have a deleterious effect for (ORF1ab, ORF3a , and the N protein. The analysis of the haplotype network of Moroccan strains suggests different sources of SARS-CoV-2 infection in Morocco. Likewise, the phylogenetic analysis revealed that these Moroccan strains were closely related to those belonging to the five continents, indicating no specific strain dominating in Morocco.

These findings have the potential to lead to new comprehensive investigations combining genomic data, epidemiological information, and clinical characteristics of SARS-CoV-2 patients in Morocco and could indicate that the developed vaccines are likely to be effective against Moroccan strains.

Keywords: SARS-CoV-2, Morocco, Genetic diversity, Deleterious mutations,Network of haplotypes, Phylogenetic analysis

3.2 Introduction

The novel coronavirus 2019, also known as severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) [1], is the causative agent of coronavirus disease-2019 (COVID-19), a new type of pneumonia that has caused an epidemic in Wuhan, China, in late December 2020. The virus's fast transmission worldwide and a large number of confirmed cases made the World Health Organization (WHO) declare COVID-19 as a global pandemic on March 11, 2020 [2]. As of October 11, 2020, the virus has spread to 235 different countries, infected more than 37 million people, and caused more than one million deaths (https://covid19.who.int/). It should be noted that mortality from SARS-CoV-2 differs considerably by geographic region. In Morocco, the first case of COVID-19 was identified on March 02. Since then, the number of infections and the number of deaths has been increasing continuously; by October 11, 2020, the Moroccan ministry of Health announced 149,841 confirmed cases, including 2,572 deaths.

SARS-CoV-2 is a positive-sense single-stranded RNA virus, encoding four structural proteins (spike (S), envelope (E), membrane (M) and nucleocapsid (N), 16 non-structural proteins (nsp1 to nsp16), and five accessory proteins (ORF3a, ORF6, ORF7a, ORF7b, and ORF8) [3][4]. Among these, two genes are considered to be the most important targets for candidate vaccines: the S protein, which is responsible for the binding to host cells membrane receptors (ACE2) via its receptor-binding domain (RBD), and the RNA-dependent RNA polymerase (RdRp, also called nsp12) which is a key part of the virus replication/transcription machinery [5][6][7].

It is known that the mutation rate of the RNA virus contributes to viral adaptation, creating a balance between the integrity of genetic information and the variability of the genome, thus allowing viruses to evade the host's immune system and develop drug resistance [8][9]. Indeed, recent studies have reported specific genotypes at particular geographic locations and that the evolution of SARS-CoV-2 over time shows a mechanism of co-accumulation of mutations that could potentially affect the spread and the severity of this virus .

At the time of writing, genetic variants, their impact, and distribution along the viral genome of Moroccan strains are poorly documented. Few studies investigating the genetic characteristics have been done so far. The analysis of 20 Moroccan SARS-CoV-2 genomes suggested that the epidemic spread in Morocco did not show a predominant SARSCoV-2 route and took origin from multiple and unrelated sources [8].

In this study, we investigated the level of diversity of 48 strains of SARS-CoV-2 that were collected in Morocco between mid-March and the end of May 2020, including nine new strains. The potential source of these strains was also predicted by comparison to other genomes from six continents.

3.3 Materials and Methods

3.3.1 Data collection and Genomes sequencing

Thirty-nine Moroccan genomes were collected from the GISAID database (http://gisaid.org) [10]. Also, nine genomes were sequenced in-house recently to build a dataset counting 48 Moroccan genomes. The nine genomes' sequencing was achieved by the routine workflow on the MinION Mk1B Nanopore platform using the R9.4.1 flowcell. The viral RNA was extracted from nine clinical samples, and the cDNA was synthesized using a kit (Roche) with random hexamers.

Genome enrichment was done by Q5 Hot Start High-Fidelity DNA Polymerase (NEB), following the manufacturer's specifications, using a set of primers designed by the ARTIC network (https://artic.network/ncov-2019) that target overlapping regions of the SARS-CoV-2 genome. The PCR products were purified by adding an equal volume of AMPure XP beads (Beckman Coulter).

The sequencing was performed according to the eight-hour routine workflow, and amplicons were repaired with NEBNext FFPE Repair Mix (NEB), followed by the DNA ends preparation using NEBNext End repair/ dA-tailing Module (NEB) before adding native barcodes and sequencing adapters supplied in the EXP-NBD104/114 kit (Nanopore) to the DNA ends. After priming the flow cell, 60 ng DNA per sample was pooled with a final volume of 65 uL. Following the ligation sequencing kit (SQK-LSK109) protocol, the sequencing was performed using the MinION Mk1B device.

3.3.2 The assembly

The gupplyplex and minion scripts of the ARTIC Network bioinformatics protocol (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html) were used for reads Preprocessing and Consensus Building for nanopore sequencing. Gupplyplex (default parameters) was used for quality control and filtering of reads (–min-length 400 –max-length 700) followed by MinION pipeline (default parameters) to perform the sequences mapping, primers trimming, variations calling, and consensus assembly building.

3.3.3 Variant calling analysis

A set of 39 SARS-CoV-2 genomes were downloaded from the GISAID database (http://www.gisaid.org/) and added to the nine newly sequenced genomes [10](Table 1). The reads generated by MinION Nanopore-Oxford of the nine isolates were mapped to the reference sequence genome Wuhan-Hu-1/2019 using BWA-MEM v0.7.17-r1188 [11] with default parameters, while the data downloaded from the GISAID database were mapped using Minimap v2.12-r847 [12] The BAM files were sorted using SAMtools [13] and were subsequently used to call the genetic variants in variant call format (VCF) by BCFtools [13].

The final call set of the 48 genomes was annotated, and their impact was predicted using SnpEff v 4.3t [14] First, the SnpEff databases were built locally using annotations of the reference genome NC_045512.2 obtained in the GFF format from the NCBI database. Then, the SnpEff database was used to annotate SNPs and with putative functional effects according to the categories defined in the SnpEff manual (http://snpeff.sourceforge.net/SnpEff_manual.html). The variants were also evaluated for functional consequences using the SIFT algorithm. The SIFT prediction is given as a Tolerance Index score ranging from 0.0 to 1.0, which is the normalized probability that the amino acid change is tolerated. SIFT scores less than or equal to 0.05 are predicted by the algorithm to be intolerant or deleterious amino acid substitutions, while scores greater than 0.05 are considered tolerant.

3.3.4 Phylogenetic and haplotype network analysis

In order to determine the source(s) of strains circulating in morocco, we performed a multiple sequence alignment using Muscle v 3.8 [15] for the 48 Moroccan strains with 225 genomes of SARS-CoV-2 from Africa, Asia, Europe, North, South America, and Oceania (Supplementary Material ;Table S2).

Maximum-likelihood trees were inferred with IQ-TREE v1.5.5 under the GTR model [16]. Generated trees were visualized using FigTree 1.4.3 To generate haplotypes for the 48 Moroccan genomes, we aligned the viruses' complete genomes using Muscle v 3.8 [15]. The generated (.fasta) file was converted to a (.meg) file using a home-made script. Using the dnasp5 tool [17], we generated the file (.rdf), compatible with the NETWORK Package, which allowed the network's tracing linking the 48 genomes' haplotypes, using the MJ (Median Joining). In order to estimate genealogical relationships of haplotype groups, the phylogenetic networks were inferred by PopART package v1.7.2 [18] using the TCS method and MSN, respectively.

3.4 Results

3.4.1 Genetic diversity of Moroccan strains of SARS-CoV-2

To identify and study the genetic variants of the SARS-COV-2 genomes from Morocco, 48 genomes were collected from mid-March to the end of May; including nine strains sequenced in the present study (Supplementary Material , Table S1).

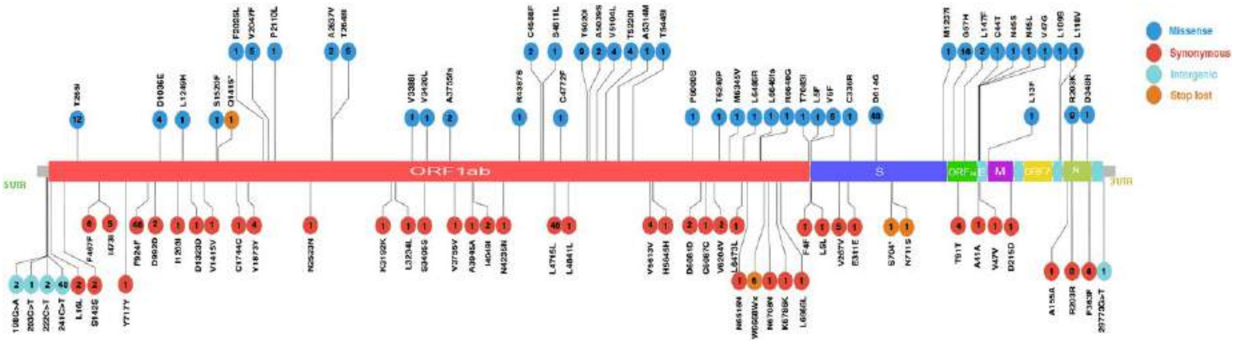

Figure 3.2: **SARS-CoV-2 genomes landscape illustration representing mutations identified in 48 Moroccan genomes**. Colored circles represent gene distribution across the genomes.

94.9% to 99.93% of the reads produced for the nine genomes were correctly mapped to the Wuhan-Hu-1/2019 reference sequence. Analysis of genetic variants revealed a total of 108 variant sites in the 48 genomes analyzed **(Figure 3.2)**; including 54 (50%) non-synonymous, 42 (38.89%) synonymous, and 3 frame-shifts (2.78%). The remaining (8.33%) were distributed along the intergenic regions. Interestingly, 36.11% of the total mutations were shared between at least two genomes, while the rest were singleton mutations (unique to one genome). Mutational distribution along the viral genome revealed seven affected genes (ORF1ab, S, N, M, E, ORF8, and ORF3a) with varying mutational frequencies. Among the non-synonymous mutations, 28 were located in the ORF1ab gene and were present in eight non-structural proteins; 9 mutations in the nsp12-RNA-dependent RNA polymerase (RdRp) (C4588F, S4611L, C4772F, T5020I, A5039S, V5104L, T5220I, R5314M, and T5448I), eight mutations in the nsp3-Multidomain (D1036E, L1249H, S1520F, F202526 P21V47, V204710, V2047, and T2648I), three mutations in nsp5-main proteinase (V3388I, N3405S, V3420L), three mutations in nsp14-Exonuclease (P6000S, T6249P, M6345V), 2 in nsp15-one (EndoRNAse) (K6486R and R6648G), 1 in nsp2 (T265I), one in nsp10-CysHis (R4387S), and 1 in nsp16 Methyltransferase (T7083I) (Table S2).

We systematically performed the analysis predicting the deleterious effect of missense mutations in two genomes or more. Functional evaluation based on the SIFT (Sorting Intolerant From Tolerant) algorithm broke the tolerance index score of 15 missense mutations (found in two genomes or more), revealing its deleterious effect (**Table 3.2**).

Table 3.2: **List of 15 non-synonymous amino acid substitutions with their tolerant effects**

	position	**(n=48)**	**(%)**	**protein**	**index (SIFT)**
nsp2	T265I	12	11.11	Benign/Tolerated	0.62
nsp3	D1036E	4	3.7	Benign/Tolerated	1
nsp3	V2047F	5	4.63	Deleterious	0.04
nsp3	A2637V	2	1.85	Benign/Tolerated	0.31
nsp3	T2648I	5	4.63	Deleterious	0
nsp12	C4588F	2	1.85	Deleterious	0.02
nsp12	T5020I	9	8.33	Deleterious	0
nsp12	A5039S	2	1.85	Benign/Tolerated	1
nsp12	V5104L	4	3.7	Deleterious	0
S	V6F	5	4.63	Benign/Tolerated	0.15
S	D614G	48	44.44	Benign/Tolerated	0.62
ORF3a	Q57H	16	14.81	Deleterious	0
ORF3a	L147F	2	1.85	Deleterious	0.03
N	R203K	13	12.04	Benign/Tolerated	0.11
N	G204R	13	12.04	Deleterious	0.02

Eight missense mutations were found with deleterious effects, while the rest predicted to be benign/tolerated. Of the eight deleterious mutations, six were observed in the non-structural protein of the ORF1ab region; three (C4588F, T5020I, V5104L) could be deleterious for the nsp12-RdRp activity and two (V2047F, T2648I) in nsp3-Multidomain. Likewise, two deleterious mutations (Q57H, L147) were found in the accessory protein ORF3a and one in the N protein (G204R).

3.4.2 Haplotype network analysis

To estimate the number of introductions of the SARS-CoV-2 virus in Morocco, an analysis of the haplotype network was carried out using the 48 strains. Our results showed that these strains were grouped into five distinct clades, harboring 28 haplotypes from which three were dominant **(Figure 3.3)**.

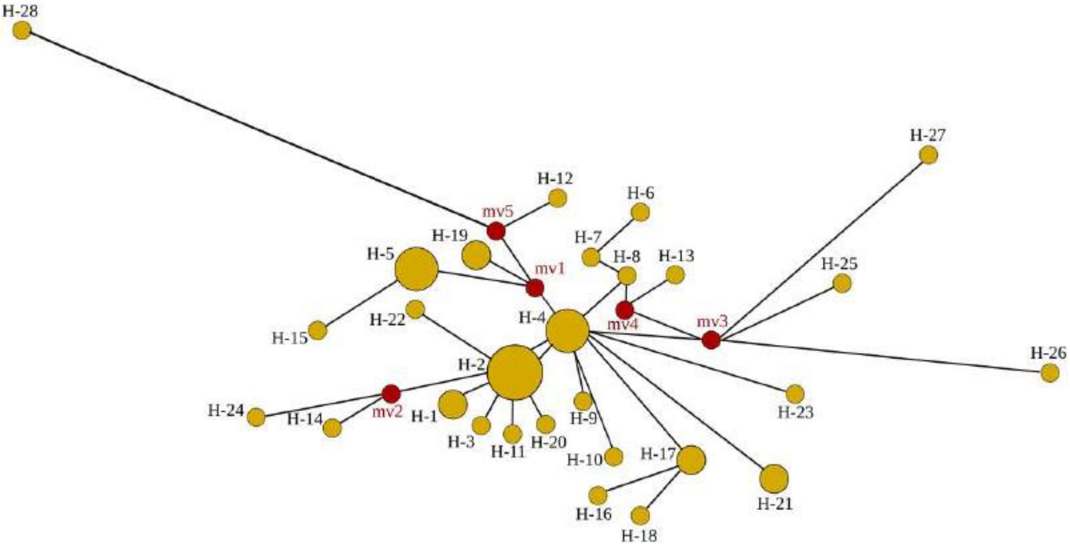

Figure 3.3: **Haplotype network using genome-wide single-nucleotide variations (HN-GSNVs) of SARS-CoV-2 isolates from Morocco.**

Current samples have shown three major haplotypes harboring the majority of strains, H2 (11 strains), H4 (5 strains), H5 (4 strains), followed by H17, H21, and H1, indicating the predominance of H2 among Moroccan strains. In **Figure 3.4**, the haplotypes are distributed in potential haplotype groups. Specifically, 43, 16, 47, and 88 haplotypes may be sub-haplotypes of primary ancestral haplotypes H1, H89, H25, and H59. Three of the 194 haplotypes that were found in Moroccan isolates were Moroccan-specific. The majority of Moroccan sequences were associated with haplotype group H25. The haplotype groups' distribution patterns differed in various geographic regions, with a few countries/territory specific. The 25 haplotypes harbored mainly strains from Asia, North American, and Africa, respectively, and are considered the ancestral strains.

The number of haplotypes increased over time as new variants were continuously acquired in haplotype group 2, harboring the second biggest haplotype, including strains from Europe, Morocco, Asia, Africa, and South America, which originated 17 Moroccan haplotypes. H89 contained mainly strain from Asia, Africa, and Europe and give birth to five haplotypes from North America and Morocco. Lastly, group 4, which contains mostly strains from Asia, harbored fewer strains from Morocco. Phylogenetic analysis of Moroccan SARS-CoV-2 genomes with genomes belonging to six geographic areas.

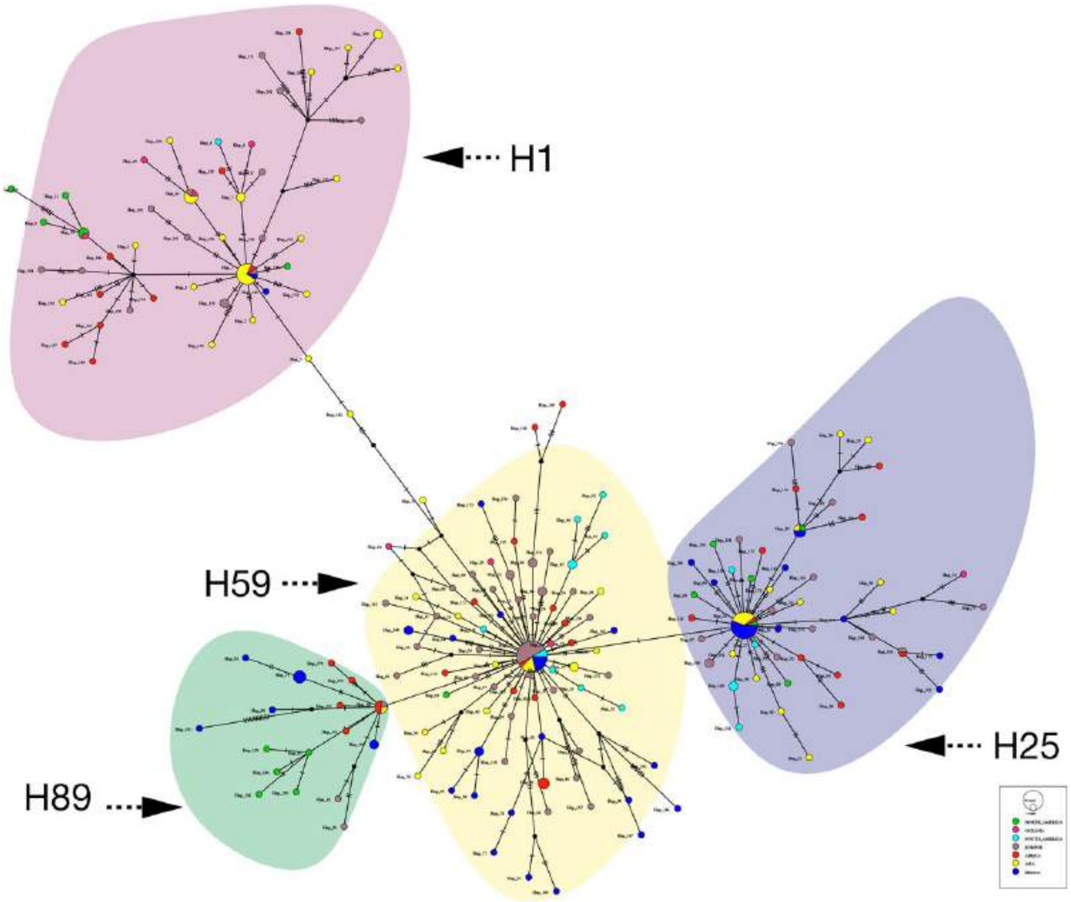

Figure 3.4: **Haplotype network using genome-wide single-nucleotide variations (HN-GSNVs) of SARS-CoV-2 isolates in the world.**

The phylogenetic analysis was carried out using 272 genomes from different countries belonging to the six continents (Supplementary Material ; Table S3) to study the possible origin of SARSCoV-2 strains circulating in Morocco. According to the GISAID nomenclature (**Figure 3.5**), our results revealed seven main clades: the ”L” clade mainly contained genomes from Asia, while the other clades contained genomes belonging to different geographical areas.

We observed that all Moroccan SARS-CoV-2 strains belonged to three close clades ”G, GH, and GR,” harboring all the D614G mutation. These clades are also subdivided into several subclades. The G clade housed about half of the Moroccan strains and is mainly closely related to European origin strains, except three strains (Morocco / S6S, Morocco / ref1, Morocco / HMMV4) are closer to strains from Kenya and the USA. Meanwhile, among the Moroccan strains of the GH clade, five (Maroc / RMPS-01, Maroc / RMPS-14, Maroc / RMPS-04, Maroc / RMPS-10, Maroc / RMPS-13, Maroc / 6901) grouped close to the strains of Tunisia; In contrast, the other Moroccan strains of this same clade appear to share close sequence similarity with strains from different geographic areas, including Asia (Israel, Taiwan, Saudi Arabia), North America (USA), Europe (France, England), and South America (Colombia).

On the other hand, the GR clade contained four Moroccan strains (Maroc / 6906, Maroc / 6899, Maroc / 6904, Maroc / 6900, Maroc / 6887) grouped with strains from the Gambia, Senegal (West Africa), and Bangladesh (Asia). Moreover, nine strains were grouped with European strains, mainly from Russia, the Czech Republic, Spain, and Cyprus. Overall, Moroccan strains were closely related to those from different continents, indicating different infection sources.

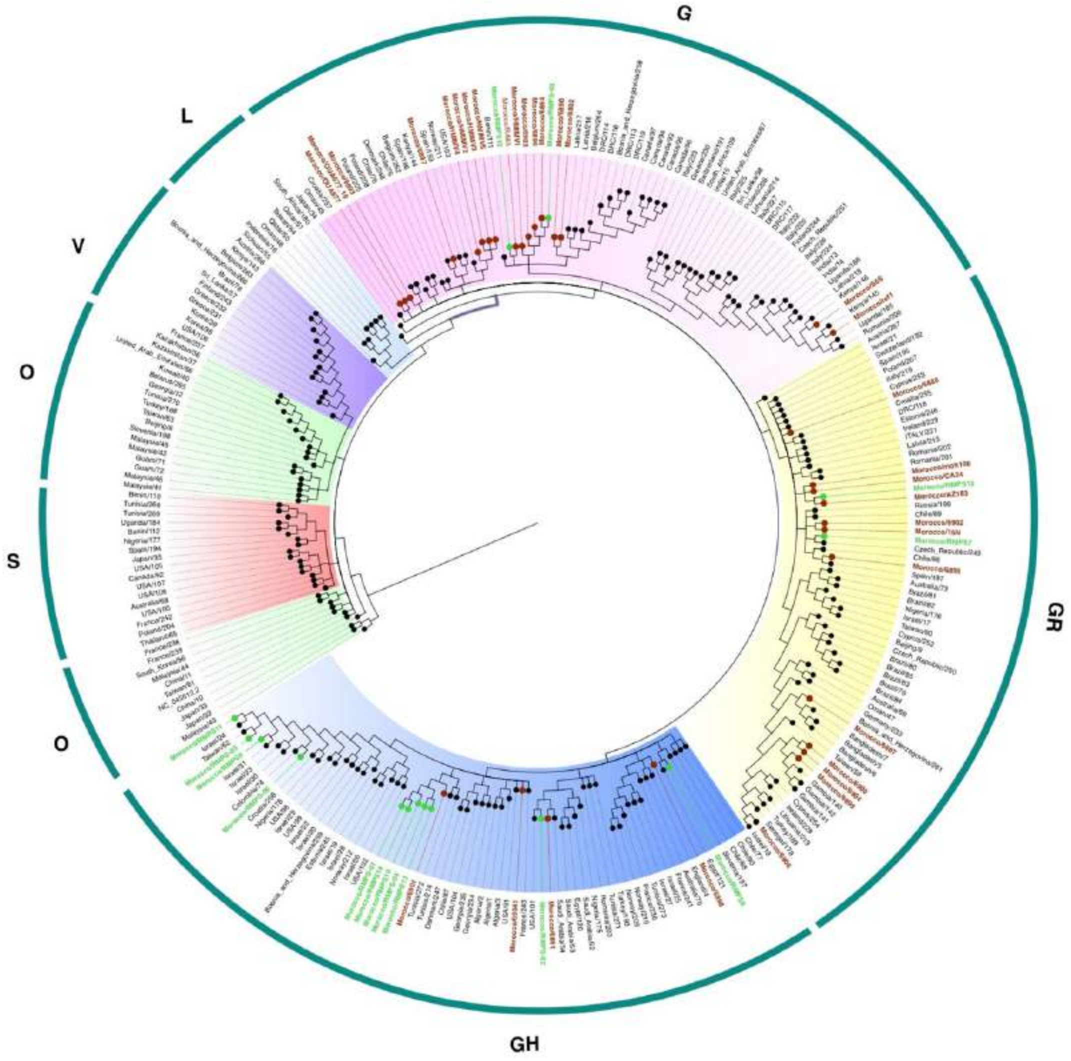

Figure 3.5: **Phylogenetic tree based on 271 complete SARS-COV2 genomes from different geographic areas.** The scale bar shows the branch's length, which represents the change of nucleotides in the genome. The six Moroccan isolates newly sequenced in this study are represented by turquoise, and the other genomes from the same country (retrieved from the GISAID database), represented by red

3.5 Discussion

The monitoring of genetic variants plays a significant role in orienting the therapeuticapproach for the development of candidate vaccines to limit the SARS-CoV-2 pandemic[19], as there is currently no proven effective treatment for SARS-CoV-2. To date,the genetic diversity of strains of SARS-CoV-2 from Morocco is poorly documented.In this study, we performed a detailed analysis of genetic variants of forty-eight Mo-roccan strains, including nine newly sequenced, to provide new information on geneticdiversity and transmission of SARS-CoV-2 in Morocco. Genetic diversity could poten-tially increase the physical shape of the viral population and make it difficult to fight,or the opposite, make the virus weaker, which could be translated with a loss of its virulence and a decrease in the number of critical cases [20].

Compared to the Wuhan-Hu-1/2019 reference sequence, Moroccan strains harbored 4 to 15 genetic variants perstrain, of which 1 to 11 involve a change of amino acids. These results are consistentwith the mutation rate previously reported in SASR-CoV-2 from different geographicareas [21][22][23][24], reporting a low frequency of recurrent mutations in thousands ofSARS-CoV-2 genomes[21] . In total, 108 variant sites have been identified, of whichonly 36% are present in two or more genomes. This result correlates with previousstudies that SARS-CoV-2 evolved and diversified mainly by a random genetic drift,which plays a dominant role in propagating single mutations [25][26][27]. The mutationswere distributed along the virus genome. The ORF1ab polyprotein region, known tobe express 16 non-structural proteins (nsp1-nsp16), had several mutations that couldaffect their activity [3].

Three non-synonymous deleterious mutations were found in theRdRp region (also called nsp12), a key part of the replication/transcription machinery,and which is proposed as a potential therapeutic target to inhibit viral infection (28,29). Likewise, a deleterious mutation has been predicted in the NSP3 protein, In addition, two deleterious mutations have been observed in ORF3a, an accessory protein that makes it possible to regulate the interferon signaling pathway and the production of cytokines [28], and the structural N protein that plays a crucial function in the virus genome by regulating RNA transcription and modulating biological processes in infected cells. RNA viruses acquire mutations readily, most of which are deleterious, and viruses carrying such mutations are eliminated. If a mutation reaches a high frequency, the mutation is expected to provide a selective advantage to the virus, usually manifested by a higher transmission efficiency. Three of these deleterious mutations (T5020I-nsp3,G204R-N, Q57H-ORF3a) have been previously reported (N-P) as hotspot mutations a large population belonging to different geographical areas. Five non-synonymous mutations were common within at least two genomes, among them, D614G (in S protein)and Q57H (in OR3a). The D614G mutation is proximal to the S1 cleavage domain of advanced glycoprotein [29] and was of great interest due to their predominance in the six continents [30][31] Alouane et al. [21] showed that this mutation appeared for thefirst time on January 24, 2020, in the Asian region (China); after a week, it was also observed in Europe (Germany).

The Q57H mutation was taken away end of February in Africa (Senegal), Europe (France and Belgium), and North America (the USA and Canada). Likewise, our previous study [21] showed that D614G had no impact on the two-dimensional and three-dimensional advanced glycoprotein structures. Furthermore,four of these five mutations (D614G, Q57H, T265I, and T5020I) have been consideredhotspot mutations in a large population [21][22]. With this in mind, we believe that this will not present a serious issue for vaccine development and that a universal candidate vaccine for all circulating strains of SARS-CoV-2 may be possible. It should be noted that our predicted deleterious variants in protein structure lack experimental validation.

Further exploration would be needed to confirm their potential effects further. Our results of the haplotype network of Moroccan strains revealed five diverse clades with 28 haplotypes, indicating different infection sources. Besides the haplotype network, the phylogenetic analysis using a set of 273 strains representing the six continents revealed seven main clades. The most important clade contained approximately three-quarters of all the strains. All SARS-CoV-2 strains from North Africa harboring the D614G mutation belonged to this clade, except for three Tunisian strains. Interestingly, the Moroccan and Tunisian strains were closely related to those from Asia, Europe, South,and North America, which could indicate different sources of SARS-CoV-2 infection in these two countries.

3.6 Conclusion

The results of this study provide valuable information on the diversity and impact of genetic variants of Moroccan strains of SARS-CoV-2 and their possible origins. This discovery could contribute to further in-depth investigations combining genomic data and clinical epidemiology of SARS-CoV -2 patients in Morocco, and suggest that, to date, the limited diversity seen in Moroccan SARS-CoV-2 should not preclude a single vaccine from providing global protection.

3.7 Reference

1. Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song,Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirusfrom patients with pneumonia in china, 2019.New England journal of medicine,2020.

2. Domenico Cucinotta and Maurizio Vanelli. Who declares covid-19 a pandemic.ActaBio Medica: Atenei Parmensis, 91(1):157, 2020.

3. Aiping Wu, Yousong Peng, Baoying Huang, Xiao Ding, Xianyue Wang, Peihua Niu,Jing Meng, Zhaozhong Zhu, Zheng Zhang, Jiangyuan Wang, et al. Genome com-position and divergence of the novel coronavirus (2019-ncov) originating in china.Cell host and microbe, 27(3):325–328, 2020.

4. Yasmin A Malik. Properties of coronavirus and sars-cov-2.The Malaysian journalof pathology, 42(1):3–11, 2020.

5. Ning Wang, Jian Shang, Shibo Jiang, and Lanying Du. Subunit vaccines againstemerging pathogenic human coronaviruses.Frontiers in microbiology, 11:298, 2020.

6. Wanbo Tai, Lei He, Xiujuan Zhang, Jing Pu, Denis Voronin, Shibo Jiang, YusenZhou, and Lanying Du. Characterization of the receptor-binding domain (rbd)of 2019 novel coronavirus: implication for development of rbd protein as a viralattachment inhibitor and vaccine.Cellular and molecular immunology, 17(6):613–620, 2020.

7. Wen-Hsiang Chen, Ulrich Strych, Peter J Hotez, and Maria Elena Bottazzi. Thesars-cov-2 vaccine pipeline: an overview.Current tropical medicine reports, pages1–4, 2020.

8. Esteban Domingo. Viruses at the edge of adaptation.Virology, 270(2):251–253,2000.

9. EJJH Domingo and JJ Holland. Rna virus mutations and fitness for survival.Annual review of microbiology, 51(1):151–178, 1997.12 Bibliography13

10. Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenzadata–from vision to reality.Eurosurveillance, 22(13):30494, 2017.

11. Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform.bioinformatics, 25(14):1754–1760, 2009.

12. Heng Li. Minimap2: pairwise alignment for nucleotide sequences.Bioinformatics,34(18):3094–3100, 2018.

13. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Ga-bor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/mapformat and samtools.Bioinformatics, 25(16):2078–2079, 2009.

14. Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, LuanWang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotatingand predicting the effects of single nucleotide polymorphisms, snpeff: Snps in thegenome of drosophila melanogaster strain w1118; iso-2; iso-3.Fly, 6(2):80–92, 2012.

15. Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and highthroughput.Nucleic acids research, 32(5):1792–1797, 2004.

16. Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh.Iqtree: a fast and effective stochastic algorithm for estimating maximum-likelihoodphylogenies.Molecular biology and evolution, 32(1):268–274, 2015.

17. Julio Rozas, Albert Ferrer-Mata, Juan Carlos Saanchez-DelBarrio, Sara Guirao-Rico,Pablo Librado, Sebastian E Ramos-Onsins, and Alejandro Saanchez-Gracia. Dnasp6: Dna sequence polymorphism analysis of large data sets.Molecular biology andevolution, 34(12):3299–3302, 2017.

18. Jessica W Leigh and David Bryant. popart: full-feature software for haplotypenetwork construction.Methods in Ecology and Evolution, 6(9):1110–1116, 2015.

19. Yung-Fang Tu, Chian-Shiu Chien, Aliaksandr A Yarmishyn, Yi-Ying Lin, Yung-Hung Luo, Yi-Tsung Lin, Wei-Yi Lai, De-Ming Yang, Shih-Jie Chou, Yi-Ping Yang,et al. A review of sars-cov-2 and the ongoing clinical trials.International journalof molecular sciences, 21(7):2657, 2020.

20. Colin R Parrish, Edward C Holmes, David M Morens, Eun-Chung Park, Donald SBurke, Charles H Calisher, Catherine A Laughlin, Linda J Saif, and Peter Daszak.Cross-species virus transmission and the emergence of new epidemic diseases.Microbiology and Molecular Biology Reviews, 72(3):457–470, 2008. Bibliography1

21. Tarek Alouane, Meriem Laamarti, Abdelomunim Essabbar, Mohammed Hakmi,El Mehdi Bouricha, MW Chemao-Elfihri, Souad Kartti, Nasma Boumajdi, HoudaBendani, Rokia Laamarti, et al. Genomic diversity and hotspot mutations in 30,983sars-cov-2 genomes: moving toward a universal vaccine for the "confined virus"?Pathogens, 9(10):829, 2020.

22. Meriem Laamarti, Tarek Alouane, Souad Kartti, MW Chemao-Elfihri, Mohammed-Hakmi, Abdelomunim Essabbar, Mohamed Laamarti, Haitam Hlali, Houda Bendani, Nassma Boumajdi, et al. Large scale genomic analysis of 3067 sars-cov-2genomes reveals a clonal geo-distribution and a rich genetic variations of hotspotsmutations.Plos one, 15(11):e0240345, 2020.

23. Sk Sarif Hassan, Pabitra Pal Choudhury, and Bidyut Roy. Sars-cov2 envelopeprotein: non-synonymous mutations and its consequences.Genomics, 112(6):3890–3892, 2020.

24. Wen-Bin Yu, Guang-Da Tang, Li Zhang, and Richard T Corlett. Decoding theevolution and transmissions of the novel pneumonia coronavirus (sars-cov-2/hcov-19) using whole genomic data.Zoological Research, 41(3):247, 2020.

25. Erik Volz, Verity Hill, John T McCrone, Anna Price, David Jorgensen, AnineO'Toole, Joel Southgate, Robert Johnson, Ben Jackson, Fabricia F Nascimento,et al. Evaluating the effects of sars-cov-2 spike mutation d614g on transmissibilityand pathogenicity.Cell, 184(1):64–75, 2021.

26. Yi Xu, Lu Kang, Zijie Shen, Xufang Li, Weili Wu, Wentai Ma, Chunxiao Fang,Fengxia Yang, Xuan Jiang, Sitang Gong, et al. Dynamics of severe acute respiratorysyndrome coronavirus 2 genome variants in the feces during convalescence.Journalof Genetics and Genomics, 2020.

27. Cheryl Yi-Pin Lee, Siti Naqiah Amrun, Rhonda Sin-Ling Chee, Yun Shan Goh,Tze Minn Mak, Sophie Octavia, Nicholas Yeo, Zi Wei Chang, Matthew Zirui Tay,Anthony Torres, et al. Neutralizing antibodies from early cases of sars-cov-2 infec-tion offer cross-protection against the sars-cov-2 d614g variant.bioRxiv, 2020.

28. Estela Gimeenez, Eliseo Albert, Ignacio Torres, Remigia, Mrıındoa, Luisa Blasco, Carlos Solanon, et al. Sars-cov-2-reactive interferon-gamma-producing cd8+ t cells in pa-tients hospitalized with coronavirus disease 2019.Journal of medical virology, 93(1):375–382, 2021.

29. Veljko Veljkovic, Vladimir Perovic, and Slobodan Paessler. Prediction of the effectiveness of covid-19 vaccine candidates.F1000Research, 9(365):365, 2020. Bibliography15

30. Muthukrishnan Eaaswarkhanth, Ashraf Al Madhoun, and Fahd Al-Mulla. Couldthe d614g substitution in the sars-cov-2 spike (s) protein be associated with highercovid-19 mortality?International Journal of Infectious Diseases, 96:459–460, 2020.

31. Kazuma Kiyotani, Yujiro Toyoshima, Kensaku Nemoto, and Yusuke Nakamura.Bioinformatic prediction of potential t cell epitopes for sars-cov-2.Journal of humangenetics, 65(7):569–575, 2020

4 Discussion

Nanopore sequencing can be successfully and cost-effective for whole-genome sequencing of viral genomes. This platform is comparable in accuracy to short read (Illumina) sequencing to generate a viral consensus sequence for each subject, provided the minimum coverage exceeds 300 reads per nucleotide position.

There is an urgent need for rapid identification and traceability of pathogens for disease control and prevention. A deep understanding of the novel virus is first obtained through the analysis of the genome sequence. This study demonstrated the utility of nanopore sequencing for SARS-CoV-2 genomes from clinical specimens based on a modified ARTIC protocol. The genomic characteristics and the origin of the virus could be quickly determined with the use of ONT devices for viral surveillance, as demonstrated during Ebola, Zika and other diseases outbreaks.

The adopted approach allowed the confirmation of SARS-CoV-2 infections at the genomic level within a few minutes by sequencing and simultaneously mapping the reads to the reference genome and analyzing the output data in real-time. To characterize the genomic variations, we found that 50% of mutation are non-synonymous distributed in seven coding regions (S, M, N, E, ORF1ab, ORF3a, and ORF8)among 48 SARS-CoV-2 genomes, without any recombination events. Genomic evidence supported the linkage of most the diverts os sars-cov-2 in Morocco. All genomes from the imported infections that occurred in Morocco exhibited the specific D614G variation in the spike protein that belongs to clade G.

Thus, The rapid 5-h workflow, with 15-min fast library preparation, could be a robust tool for backward tracing and outbreak control despite the elevated error rates. Other studies have also ensured the suitability of ONT sequencing for SARS-CoV-2 genome analysis. However, ONT sequencing failed to detect short indels and variants at low read-count frequencies accurately.

Chapter 4

Comparative genomic analysis of SARS-CoV-2 world-wide

1 Preface

Some viruses, such as influenza and HIV, have a high rate of genetic mutations, making them prone to antigenic leakage. Therefore, it is essential to assess the genetic evolution of the virus and, more specifically, the regions responsible for its interaction and replication within the host cell.

This final chapter covers an ongoing exploration of the recent pandemic. Understanding and monitoring the genetic evolution of the virus, its geographical characteristics, and its stability are particularly important for controlling the spread of the disease and especially for the development of a universal vaccine covering all circulating strains. From this perspective, we performed a large scale genomic analysis of SARS-CoV-2 genomes from 79 countries located on six continents and collected from 24 December 2019 to 13 May 2020.

This analysis was conducted in two steps, primary analysis of 3067 SARS-CoV-2 genomes starting from December to April. During this study, we characterized the genetic variants to have a detailed understanding of their genetic diversity and monitor the accumulation of mutations over time with a particular focus on the geographic distribution of recurrent mutations. After the high-frequency nonsynonymous variants, we explored the implications of this result on our purification selection hypothesis. A preliminary dN/dS analysis highlighted the number of negatively selected residues within the spike protein, specifically within the rpdb region, that could be considered therapeutic targets.

A following up analysis included 30,983 genomes, help understand the intragenomic diversity of the virus and mutations distribution, emphasising how these mutations would affect the development of a universal vaccine.

2 Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations

M.Laamarti[a], T. Alouane[a], S. Kartti[a], W Chemao[a], M. Hakmi[a], A. ESSABBAR[b], M. Laamarti[a], H. hlali[a], N. BOUMAJDI[a], H. BENDANI[a], R. L. ALLAM[a], F. GHRIFI[a], I. SMYEJ[b], J. RAHOUI[b], H. BENRAHMA[b], L. DIAWARA[b], T. AANNIZ[a],N. EL HAFIDI[a],R. EL JAOUDI[a],C. NEJJARI[c],S. AMZAZI[d],R. MENTAG[e],L. BELYAMANI[f], A. Ibrahimi
[a] [a] Biotechnology Lab (MedBiotech Center), Rabat Medical and Pharmacy School, University Mohammed V, Rabat, Morocco
[b] National Reference Laboratory, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
[c] nternational School of Public Health, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco
[d]Laboratory of Human Pathologies Biology, Faculty of Sciences, Mohammed V University in Rabat, Morocco
[e] Biotechnology Unit, Regional Center of Agricultural Research of Rabat, National Institute of Agricultural Research, Rabat, Morocco
[f]Emergency Department, Military Hospital Mohammed V, Rabat Medical and Pharmacy School, Mohammed Vth University in Rabat, Morocco

2.1 Abstract

In late December 2019, an emerging viral infection COVID-19 was identified in Wuhan, China, and became a global pandemic.

Characterization of the genetic variants of SARS-CoV-2 is crucial in following and evaluating it spread across countries. In this study, we collected and analyzed 3,067 SARS-CoV-2 genomes isolated from 55 countries during the first three months after the onset of this virus. Using comparative genomics analysis, we traced the profiles of the whole-genome mutations and compared the frequency of each mutation in the studied population. The accumulation of mutations during the epidemic period with their geographic locations was also monitored. The results showed 782 variants sites, of which 512 (65.47%) had a non-synonymous effect.

Frequencies of mutated alleles revealed the presence of 68 recurrent mutations, including ten hotspot non-synonymous mutations with a prevalence higher than 0.10 in this population and distributed in six SARS-CoV-2 genes. The distribution of these recurrent mutations on the world map revealed that certain genotypes are specific to geographic locations. We also identified co-occurring mutations resulting in the presence of several haplotypes. Moreover, evolution over time has shown a mechanism of mutation co-accumulation which might affect the severity and spread of the SARS-CoV-2. The phylogentic analysis identified two major Clades C1 and C2 harboring mutations L3606F and

G614D, respectively and both emerging for the first time in China. On the other hand, analysis of the selective pressure revealed the presence of negatively selected residues that could be taken into considerations as therapeutic targets. We have also created an inclusive unified database (http://covid-19.medbiotech.ma) that lists all of the genetic variants of the SARS-CoV-2 genomes found in this study with phylogeographic analysis around the world.

2.2 Introduction

The recent emergence of the novel, human pathogen Severe Acute Respiratory SyndromeCoronavirus 2 (SARS-CoV-2) in China with its rapid international spread poses a globalhealth emergency. On March 11, 2020, the World Health Organization (WHO) publiclyannounced the SARS-CoV-2 epidemic as a global pandemic. As of Mai 01, 2020, theCOVID-19 pandemic had affected more than 200 countries and territories, with morethan 3,175,207 confirmed cases and 224,172 deaths [1]. The new SARS-CoV-2 coronavirus is an enveloped positive-sense single-stranded RNA virus belonging to a largefamily named coronavirus which have been classified under three groups two of themare responsible for infections in mammals [2], such us: bat SARS-CoV- like; MiddleEast respiratory syndrome coronavirus (MERS-CoV).

Many recent studies have sug-gested that SARS-CoV-2 was diverged from bat SARS-CoV-like [3][4]. The size of the SARS-CoV2 genome is approximately 30 kb and its genomicstructure has followed the characteristics of known genes of Coronavirus; the polypro-tein orf1ab also known as the polyprotein replicase covers more than 2 thirds of thetotal genome size while the rest contains structural proteins, including spike protein,membrane protein, envelope protein and nucleocapsid protein. In addition to six ORFs(ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10) predicted as hypothetical proteinswith no associated function [5].
Characterization of viral mutations can provide valu-able information for assessing the mechanisms linked to pathogenesis, immune evasionand viral drug resistance. A previous study [6] based on an analysis of 103 genomes ofSARS-CoV-2 indicates that this virus has evolved into two main types, type L beingmore widespread than type S, and type S represent- ing the ancestral version. Anotherstudy [7] conducted on 32 genomes of strains sampled from China, Thailand and theUnited States between December 24, 2019 and January 23, 2020 sug- gested increasingtree-like signals from 0 to 8.2%, 18.2% and 25, 4% over time, which may indi- cate anincrease in the genetic diversity of SARS-CoV-2 in human hosts. Therefore, the analysis of mutations and monitoring of the evolutionary capacity of SARS– CoV-2 over time-based on a large population is necessary.

In this study, we characterized the geneticvariants in 3067 SARS-CoV-2 genomes for a detailed understanding of their genetic di-versity and to monitor the accumulation of mutations over time with particular focus onthe geographic distribution of recurrent mutations. On the other hand, we establishedselective pressure analysis to predict negatively selected residues which could be usefulfor the design of therapeutic targets. We have also created a database to share, exploitand research knowledge of genetic variants to facilitate SARS-CoV-2 comparison forthescientific community.

2.3 Methods

2.3.1 Data collection and variant calling analysis

3067 sequences of SARS-CoV-2 were collected from the GISAID EpiCovTM (update:02-04- 2020) and NCBI (update: 20-03-2020) databases. Only complete genomes wereused in this study (S1 Table). Genomes were mapped to the reference sequence Wuhan-Hu-1/2019 (NC045512) using Minimap v2.12-r847-dirty [8]. The BAM files were sortedby SAMtoolssort [9], then used to call the genetic variants in variant call format (VCF) by SAM-tools mpi- leup [9] and bcftools v1.8 [9].
The final call set of the 3067 genomes, wasannotated and their impact was predicted using SnpEff v 4.3t [10]. First, the SnpEffdatabases were built locally using annotations of the reference genome NC045512.2 ob-tained in GFF format from the NCBI database. Then, the SnpEff database was usedto annotate single nucleotide polymor- phisms (SNPs) and insertions/deletions (InDels)with putative functional effects according to the categories defined in the SnpEff manual(http://snpeff.sourceforge.net/SnpEffmanual.html).

2.3.2 Phylogentic analysis and geodistribution

The downloaded full-length genome sequences of coronaviruses isolated from differ-ent hosts from public databases were subjected to multiple sequence alignments usingMAFFT v7.4 [11]. Maximum-likelihood trees were inferred with IQ-TREE v1.5.5 [12]under the GTR model using NextStrain tools. Heatmap for correlation analysis was per-formed on countries and hotspots mutations using CustVis with default settings rows scaling = variance scaling, PCA method = SVD with imputation, clustering distance for rows = correlation clustering, the method for rows = average, tree ordering for rows= tightest cluster first.

2.3.3 Selective pressure and modelling

We used Hyphy v2.5.8 [13] to estimate synonymous and non-synonymous ratio dN / dS(). Two datasets of 191 and 433 for orf1ab and genes respectively were retrieved fromGen-bank (http://www.ncbi.nlm.nih.gov/genbank/). After deletion of duplicated andcleaning the sequences, only 91 and 39 for orf1ab and spike proteins, respectively, were used for the analysis (Appendix **Table 6.1**).

The selected nucleotide sequences of each dataset were aligned using Clustalw codon-by-codon and the phylogenetic tree was obtained using ML(maximum likelihood) available in MEGA X [14]. For this analysis, four Hyphy's meth-ods were used to study site-specific selection: SLAC (Single-Likelihood Ancestor Count-ing) , FEL (Fixed Effects Likeli- hood), FUBAR (Fast, Unconstrained Bayesian AppRox-imation) [15] and MEME (Mixed Effects Model of Evolution) [16]. For all the methods,values supplied by

default were used for statistical confirmation and the overall value was calculated according to ML trees under General time reversible model (GTR model).
The CI- TASSER generated models (https:// zhanglab.ccmb.med.umich.edu/COVID-19/) of nonstructural proteins (nsp3, nsp4, nsp6, nsp12, nsp13, nsp14 and nsp16 oforf1ab were used to highlight the sites under selective pres- sure on the protein. On the other hand, the cryo-EM structure with PDB id 6VSB was used as a model for the spike protein in its prefusion conformation. Structure visualization and image rendering were performed in PyMOL 2.3 (Schrodinger LLC)

2.3.4 Pangenome construction

115 proteomes of the genus Betacorononavirus were obtained from the NCBI database(update: 20-03-2020), of which 83 genomes belonged to the SARS-CoV-2 species and the rest distributed to other species of the same genus publicly available (S3 Table).These proteomes were used for the construction of pangenome at the inter-specific scale of Betacoronavirus and intra-genomic of SARS-CoV-2. The strategy of best reciprocal BLAST results [17] was imple- mented to identify all of the orthologous genes using Proteinortho v6.0b [18]. Proteins with an identity above 60% and sequence coverageabove 75% with an e-value threshold below 1e-5 were used to be considered as significant hits.

2.4 Results

2.4.1 SARS-CoV-2 genomes used in this study

In this study, we used 3,067 SARS-CoV2 complete genomes collected from GISAID Epi-CovTM (update: 02-04-2020) and NCBI (update: 20-03-2020) databases. These strains were isolated from 55 countries **(Figure 4.1 - A)**. The most represented origin was American strains with783 (25.53%), followed by strains from England, Iceland, and China with

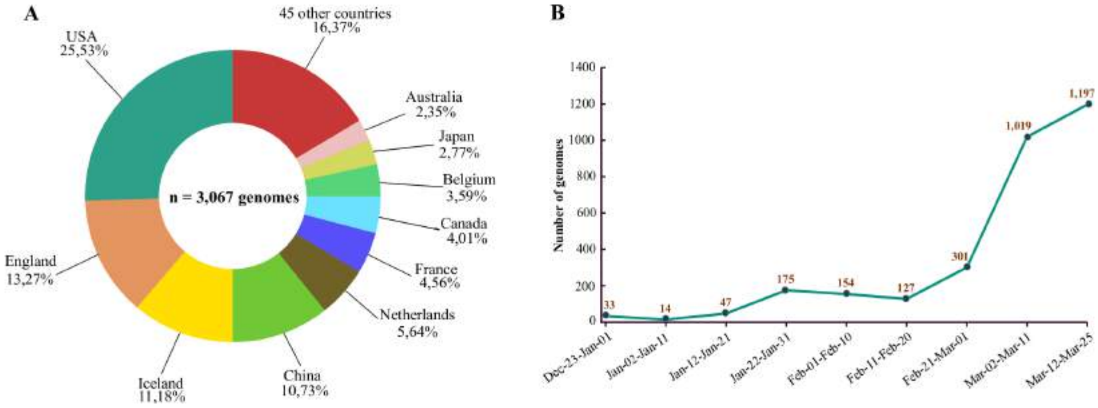

Figure 4.1: **Distribution of the 3,067 genomes used in this study by country and date of isolation.**(**A**) The pie chart represents the percentage of genomes used in this study according to their geographic origins. The colors indicate different countries. (**B**) Number of genomes of complete pathogens, distributed over a period of 3 months from the end of December to the end of March.

407 (13.27%),343 (11.18%), 329 (10.73%), respectively. The date of isolation was during the first three months after the appearance of the SARS-CoV-2 virus, from December 24, 2019,to March 25, 2020(**Figure 4.1 - B**). Likewise, about two-thirds of these strains collected inthis work were isolated during March

2.4.2 Mutational frequency analysis revealed a diversity of genetic variants in six SARS-Cov-2 genes

A total of 782 variant sites were detected compared to the Wuhan-Hu-1/2019 reference sequence, of which 65.98% having a non-synonymous effect, 28.39% synonymous mutations, and 5.63% are distributed in the regions intergenic. Mutational frequency analysis revealed the presence of 68 mutations with a frequency greater than 0.06%of the total genomes, which cor- responds to at least 20/3067 genomes. Focusing on non-synonymous mutations (freq$\geq$ 0.06$\%ofthetotalgenomes),38wereidentifiedanddistributedinsixSARS-CoV-2genesatvariablefrequencies$(**Figure 4.2**).

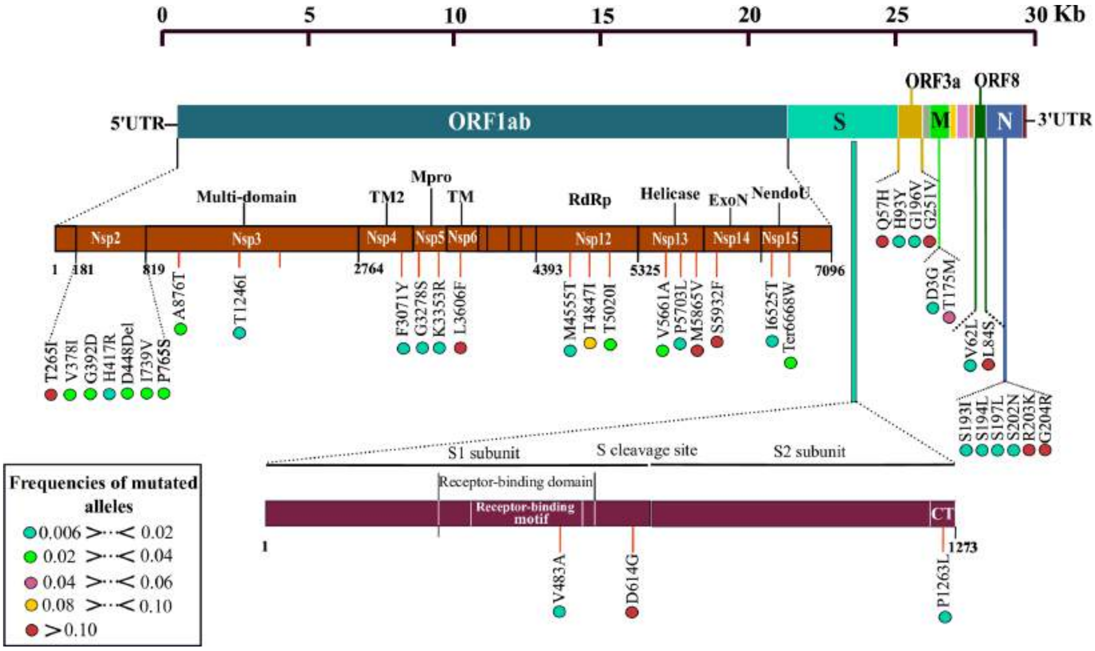

Figure 4.2: **Distribution of the 3,067 genomes used in this study by country and date of isolation.**(**A**) The pie chart represents the percentage of genomes used in this study according to their geographic origins. The colors indicate different countries. (**B**) Number of genomes of complete pathogens, distributed over a period of 3 months from the end of December to the end of March.

Among them, the ORF1ab polyprotein harbored 22 non-synonymous mutations: seven in nsp2 (T265I, V378I, G392D, H417R, I739V, P765S, and D448Del) three in nsp12-RdRp (M4555T, T4847I and T5020I), three in nsp13-nsp13 (V5661A, P5703L, and M5865V), two in nsp3-multi-domains (A876T and T1246I), two in nsp5-main proteinase (G3278S and K3353R), two in nsp15-EndoRNAse (I6525T, Ter6668W), and one in each of three proteins; nsp6-transmembrane domain (L3606F), nsp4-transmembrane domain-2 (F3071Y), and nsp14-exonuclease (S5932F). Likewise, the spike protein harbored three non-synonymous mutations, including V483A in the receptor-binding domain (RBD). The remainder was found in the core phosphoprotein (S193I, S194L, S197L, S202N, R203K, and G204R), membrane glycoprotein (D3G, T175M), ORF3a (Q57H, H93Y, G196V, and G251S V62) and ORF62884).

2.4.3 Identification of ten hyper-variable genomic hotspot in SARS-CoV-2 genomes

Interestingly, among all recurrent mutations, ten were found as hotspot mutations with a fre- quency greater than 0.10 in this study population (Fig 2). The most represented was D614G mutation at spike protein with 43.46% (n = 1.333) of the genomes, the second was L84S (at ORF8) found in 23.21% (n = 712). Thus, the gene coding for orf1ab had four mutations hot- spots, including S5932F of nsp14-exonuclease, M5865V of nsp13 helicase L3606F of nsp6 transmembrane domain and T265I of nsp2 found with 17.02%, 16.56%, 14.38% and 10.66% of the total genomes, respectively. For the four other hotspot mutations were distributed in ORF3a (Q57H and G251V) and nucleocapsid phosphoprotein (R203K and G204R).

2.4.4 Geographical distribution and origin of mutations worldwide

3067 genomes were dispersed in different countries with different genotype profiles. We performed a geo-referencing mutation analysis to identify region-specific loci. China and the USA were the countries with the highest number of mutations 301 and 296 correspondings, respectively to 38,19% and 37,56% of the total number of mutations.

These mutations include 140 (17,76%) and 229 (29%) singleton mutations specific to China and USA genomes, respectively, which is mainly due to the high number of genomes available in these countries. How- ever, England contains more than 300 genomes and harbored only 116 mutations only 23 of theme were singleton mutations. Data normalization shows that the high mutational rate was observed in New Zealand followed by Malaysia and Vietnam, respectively, While the number of singleton mutations per genome was higher in Malaysia (20,16%), China (5,2%), and the USA (2%), respectively. It is interesting to note that among the 55 countries, 21 harbored singleton mutations. S4 and S5 Tables (Supplementary material) illustrates the detailed singleton mutations found in these countries (**Table 6.2**).

The majority of the genomes analyzed carried more than one mutation. How-ever, among the recurrent non-synonymous, synonymous, deletion and intergenic mutations, we found G251V (in ORF3a), and S5932F (in ORF1ab) present on all continents except Africa**(Figure 4.3)**. While F924F, L4715L (in orf1ab), D614G (in spike) appeared in all strains except those from Asia. In Algeria, the genomes harbored mutations very similar to those in Eu-rope, including two recurrent mutations T265I and Q57H of the ORF3a. Likewise,the European and Dutch genomes also shared ten recurrent mutations. On the other hand, continent-specific mutations have also been observed, for example in America,we found seven mutations shared in almost all genomes.

Besides, two mutations at positions 28117 and 28144 were shared by the Asian genomes,

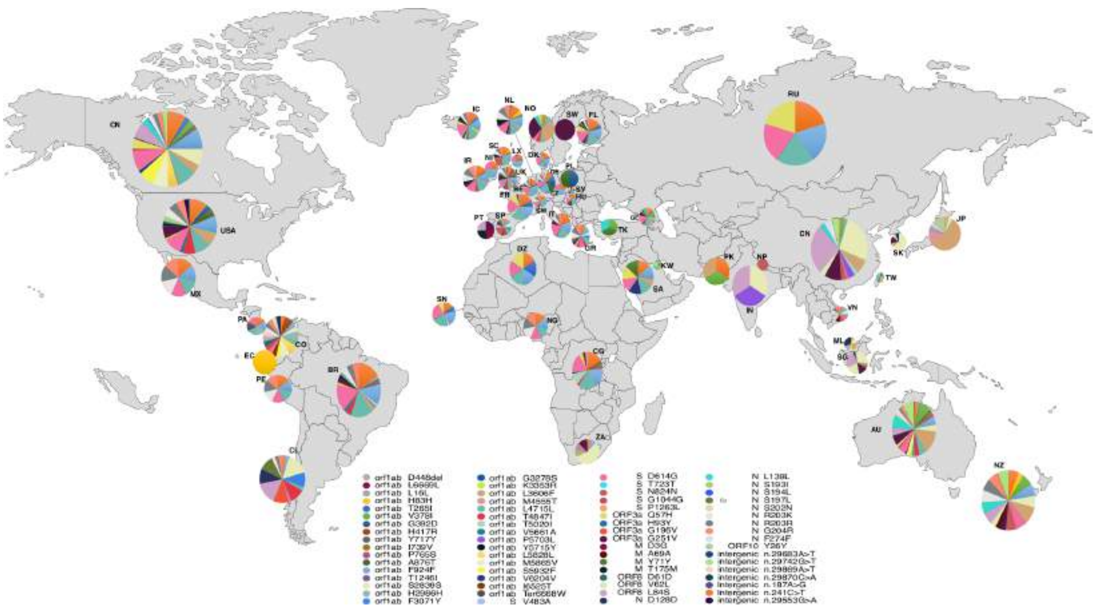

Figure 4.3: **Map showing geographical distribution of recurrent mutation in the studied population worldwide**. The pie charts show the relative frequencies of haplotype for each population. The haplotypes are color coded as shown in the key. The double-digit represent countries' two letters code. The circle's size was randomly generated with no association with the number of genomes in each country. Abbreviations: S, spike; E, enveloppe; M, membrane protein; N, nucleocapsid protein.

while four different positions1059, 14408, 23403, 25563 and 1397, 11083, 28674, 29742 were shared by African and Australian genomes (S1 Table). The majority of these mutations are considered to be transition mutations with a high ratio of A substituted by G. The genome variability was more visible in China and USA than in the rest of the world. SARS-CoV-2 genomes also harbored three co-occurrent mutations R203K, R203R and G204R in the N protein and were present in all continents except Africa and Asia (besides Taiwan).

2.4.5 Evolution of mutations over time

We selected the genomes of the SARS-CoV-2 virus during the first three months after the emergence of this virus (December 24 to March 25). We have noticed that the mutations have accumulated at a relatively constant rate **(Figure 4.4 - A)**. The strains selected at the end of March showed a slight increase in the accumulation of mutations with an average of 11.34 mutations per genome, compared to the gnomes of February, December and January with an average number of mutations of 9.26, 10.59 and 10.34 respectively. The linear curve in **Figure 4.4** suggests a continuous accumulation of single SNPs in the SARS-CoV-2 genomes in the coming months. This pointed out that many countries had multiple entries for this virus that could be claimed. Thus in the deduced network demonstrated transmission routes in different countries. The study of mutations accumulation over time showed a higher number of mutations in the middle of the outbreak(end of January). At the same time, an increase in the number of mutations in early April was

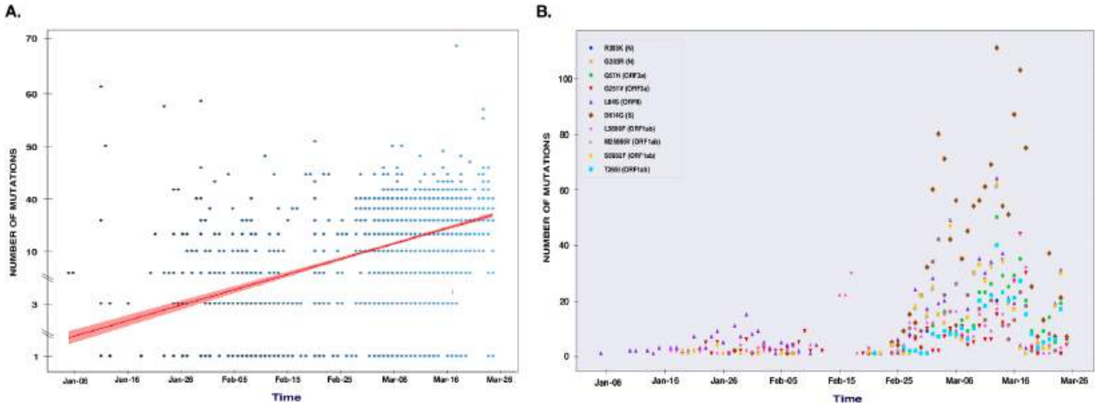

Figure 4.4: **The graph represents substitutions accumulation in a three months period.***(A)* The accumulation of mutations increases linearly with time. The dots represent the number of mutations in each genome. All substitutions have been included: non-synonymous, synonymous and intergenic mutations. **(B)** The distribution and accumulation of Hot spot mutations over time.

also observed. The first mutations to appear were mainly located in the inter-genic region linked to the nucleocapsid phospho protein and the orf8 protein. The T265I,D614G and L84S hotspot mutations located in orf1ab and Spike proteins respectively were introduced into the virus for the first time in late February **(Figure 4.4 - B)**.

2.4.6 Phylogeographical analysis of SARS-CoV-2 genomes

The phylogenetic tree based on the whole genome alignment demonstrates that SARS-CoV-2 is wildly disseminated across distinct geographical location. The results showed that several strains are closely related even though they belong to different countries.Which indicate likely transfer events and identify routes for geographical dissemination.
Phylogenetic trees based on full-genome sequences deposited and available at GISAID and NCBI revealed the diversification, and the clustering of genomes into groups, based on the genetic variants. The phylogenetic analysis revealed two main clades C1 and C2;the original clade C1 harboring the mutation F3606L and starting since the beginning of the pandemic contains mainly Chinese strains from Dec to mid-Feb. After this period,the clade has emerged in other countries all over the globe. C1 is also composed of two subclades, SCB 1 sharing the mutation G251V (ORF3a) first identified in strains from china and further emerged in European strains, such as England and Iceland. The second subclade SCB2 also stared in China at the beginning of Jan and harbored the mutation L84S (ORF8).

Following the first appearance it started emerging in other European countries mainly in Spain, this clade has also emerged in the USA in mid-Jan and gives birth to a new cluster containing 444 strains all sharing a C17747T mutation (Leu5828Leu, ORF1ab) starting from mid-Feb. Strains from the second cladC2 shared the spike mutation D614G (S) and harbored three subclades, this clade started in shanghai end of Jan. However, it contains mainly strain from Europe and North America.

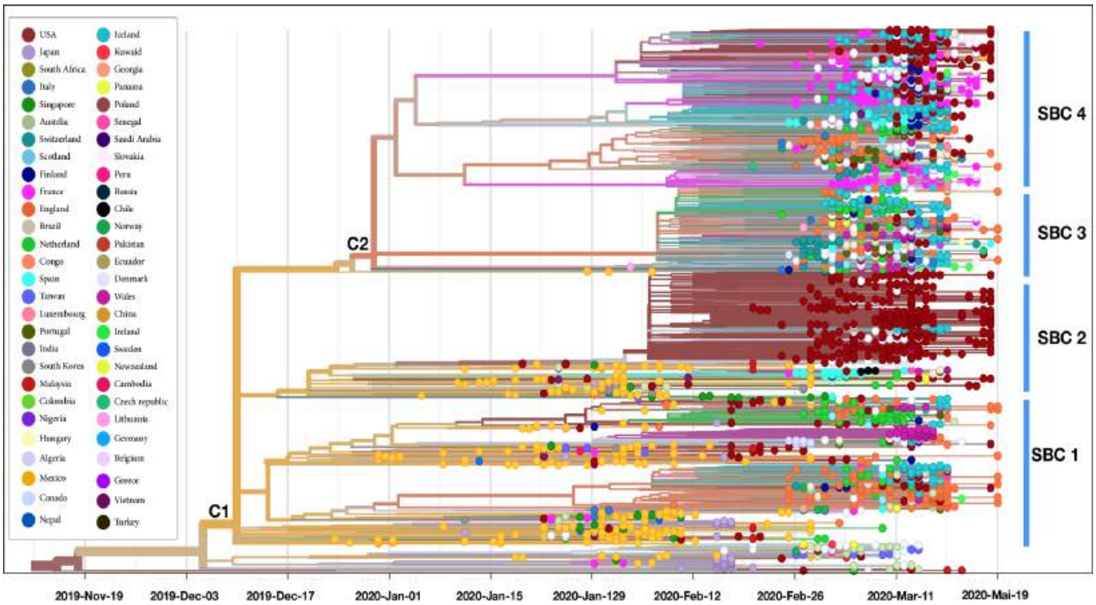

Figure 4.5: **Phylogenetic analysis of 3067 SARS-CoV 2 genomes grouped according to the country of origin** The length of the branches represents the distance in time.

The first subcluster SCB3 harbored strain sharing two mutation R203K (N)and G204R (N) harboring largely strains from Europe and some strains from North Africa (France and USA). The second subcluster SCB4 harbored strain from Europe with the Q57H (ORF3a) mutation, these clusters started in France and Netherland during mid-Feb. genomes of Countries originated from Australia, South America, and Africa are scattered through the entire tree **(Figure 4.5)**.

For phylogenetic tree (http://covid-19.medbiotech.ma) showed multiple introduction dates of the virus inside the USA with the first haplotype introduced related to the second epidemic wave in China. Using correlation analysis between most recurrent mutations and countries distribution **(Figure 4.6)**. We observed that most recurrent mutations clusters could be divided into four groups;the bigger cluster compromised nine mutations from the ten hotspots, while the first cluster harbored only the orf1ab mutation L3606F. Meanwhile, geo clustering by geo-graphic location showed two distinct clusters **(Figure 4.6)**, cluster A grouping countries from Europe with those from America and Africa.

However, Asia was only represented by Saudi Arabia. Cluster B in the other hand contained the majority of countries from the Asian and Australian continents. It is also harboring a sub-cluster containing the UK, USA, and Ireland which was previously demonstrated to contain a high number of mutations. On the other hand, mutations as V378I and L3606F (in orfab1), 29742CT (intergenic), L139L in (in nucleocapside) were mainly correlated with Pakistan,Norway, Georgia, Taiwan, Kuwait, Australia, and Turkey while (S2839S, F3071Y andT4847I), D128D and G196V mutations in orf1ab, nucleocapsid, ORF3a, respectively,were mainly present in Spain, Chile, and Greece. However, cluster harboring D614G(in spike), F924F (in orf1ab), and L4715L (in orf1ab) mutations, showed no correlation and were scatted through all countries especially those from Europe.

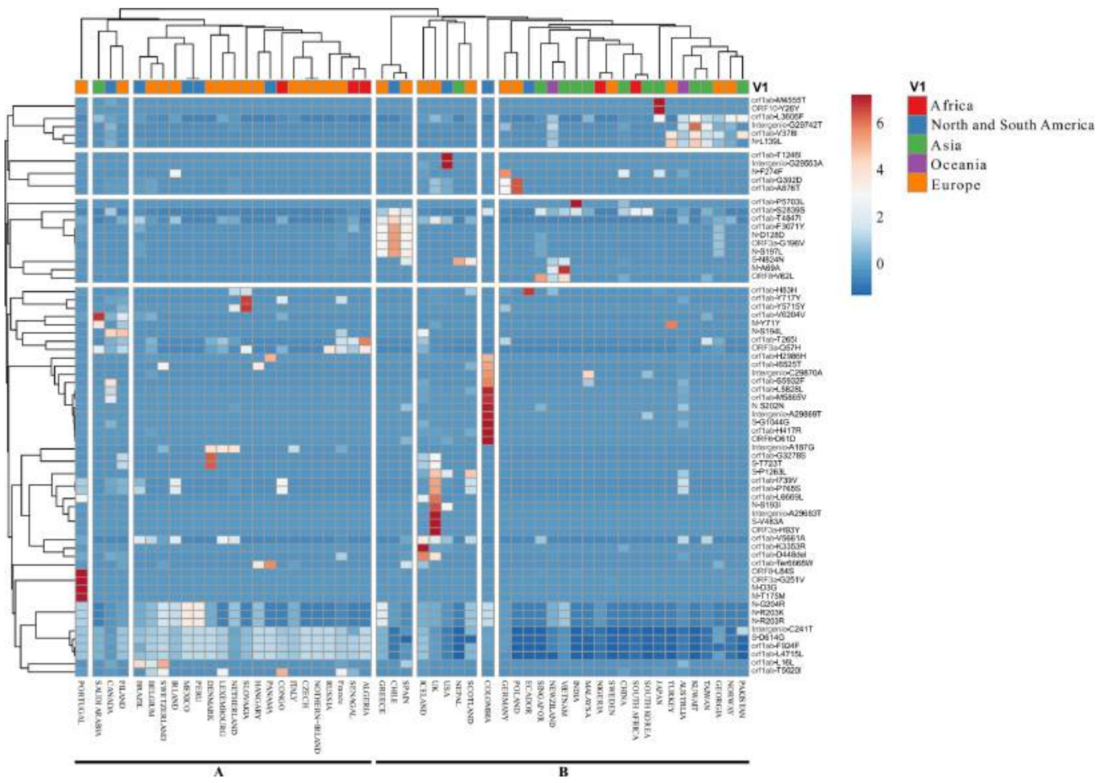

Figure 4.6: **Heatmap showing the correlation between mutations and the geographic distribution of the genomes analyzed.** The correlation was applied to a data set of 68 most recurrent mutations with different distribution in all 55 countries divided into two distinct cluster A and B. The color scale indicates the significance of correlation with blue and orange colors indicating the highest and lowest correlation. The red, yellow and orange colors in the horizontal bar represent the continent of origin. Abbreviations: S, spike; M, membrane protein; N, nucleocapsid protein.

A high correlation with a specific mutation was observed within Portugal, Saudi Arabia, Slovakia, Iceland,UK, USA, Colombia, Ecuador, Vietnam, Japan genomes.

2.4.7 Selective pressure analysis

Selective pressure on orf1ab, gene harbored a high rate of mutations and on the Spike gene, indicated a single alignment-wide ratio of 0.571391 and 0.75951 for spike and or1ab, respectively. Most sites for both genes had ¡1 values, indicating purifying selection. In orf1ab, we estimated eight sites under negative selection pressure (696,1171, 2923, 3003, 3715, 5221, 5704 and 6267) and three sites under positive selection pressure (1473, 2244 and 3090). For spike, we found seven sites under negative selection pressure (215, 474, 541, 809, 820, 921 and 1044), and only one site under negative selection pressure **(Table 4.1)**.

Table 4.1: **Selective pressure analysis on the spike and orf1ab genes of SARS-CoV-2**

Genes	O	FEL		MEME	SLAC		FUBAR	
Spike	0.571391	**PS**	**NS**	**PS**	**PS**	**NS**	**PS**	**NS**
		-	215, 474, 809, 820, 921, 1044	-	-	-	5	Codons 215,
								474, 541, 809, 820, 921, 1044
orf1ab	0.75951	**PS**	**NS**	**PS**	**PS**	**NS**	**PS**	**NS**
		2244	Codons	Codon 2244	-	-	1473,	-
		-	1171, 2923, 3003, 3715,				2244, 3090	
			5221, 5704,					
			6267, 6961					

None of the hotspot mutations was identified under negative selecion, this is moslty due sampling size and early date of sam- ple collection. The modelling results of orf1ab showed that the sites with positive selections were distributed in nsp3 and nsp4, while the negatively selected codons were located in nsp3, nsp4,nsp6, nsp12, nsp13, nsp14 and nsp16 **(Figure 4.7)**.In spike, the only negatively selected residue was observed in the RBD region **(Figure 4.8)**.

2.4.8 Inter and intra-specific pan-genome analysis

In order to highlight the proteins shared between SARS-CoV-2 and other species of the genus Betacoronavirus, Likewise, the proteins shared on the intra-genomic scale of SARS-CoV-2, we have constructed a pan-genome by clustering the sets of proteins encoded in 115 genomes distributed in 17 species, including 83 genomes belongingto SARS-Cov-2 (S3 Table). A total of 1,148 proteins were grouped into a pangenome of 94 orthologous protein clusters. Among them, ten protein clusters were shared be-tween SARS-CoV-2 and only three species of the genus *Betacoronavirus*, including; Bat-CoV RaTG13, SARS-CoV and Bat Hp-*betacoronavirus* / Zhejiang2013. The BatCoVRaTG13 genome had more orthologous proteins shared with SARS-CoV-2, followed by SARS-CoV with ten and nine orthologous proteins, respectively **(Figure 9 - A)**. It is interesting to note that among all the strains used of *Betacoronavirus*, the protein ORF8 was found in orthology only between SARS-RATG13 and SARS-CoV-2. In addition, the ORF10 protein was found as a singleton for SARS-CoV-2 genomes. On the other hand, the analysis of the pangenome at the intra-genomic scale of 83 isolates of SARS-CoV-2 **(Figure 9 - B)**, showed that ORF7b and ORF10 were two accessory genome(proteins variable between genomes) in SARS-CoV- 2 genomes, while the other proteins belonged to the core genome (present in all strains) of SARS-CoV-2.

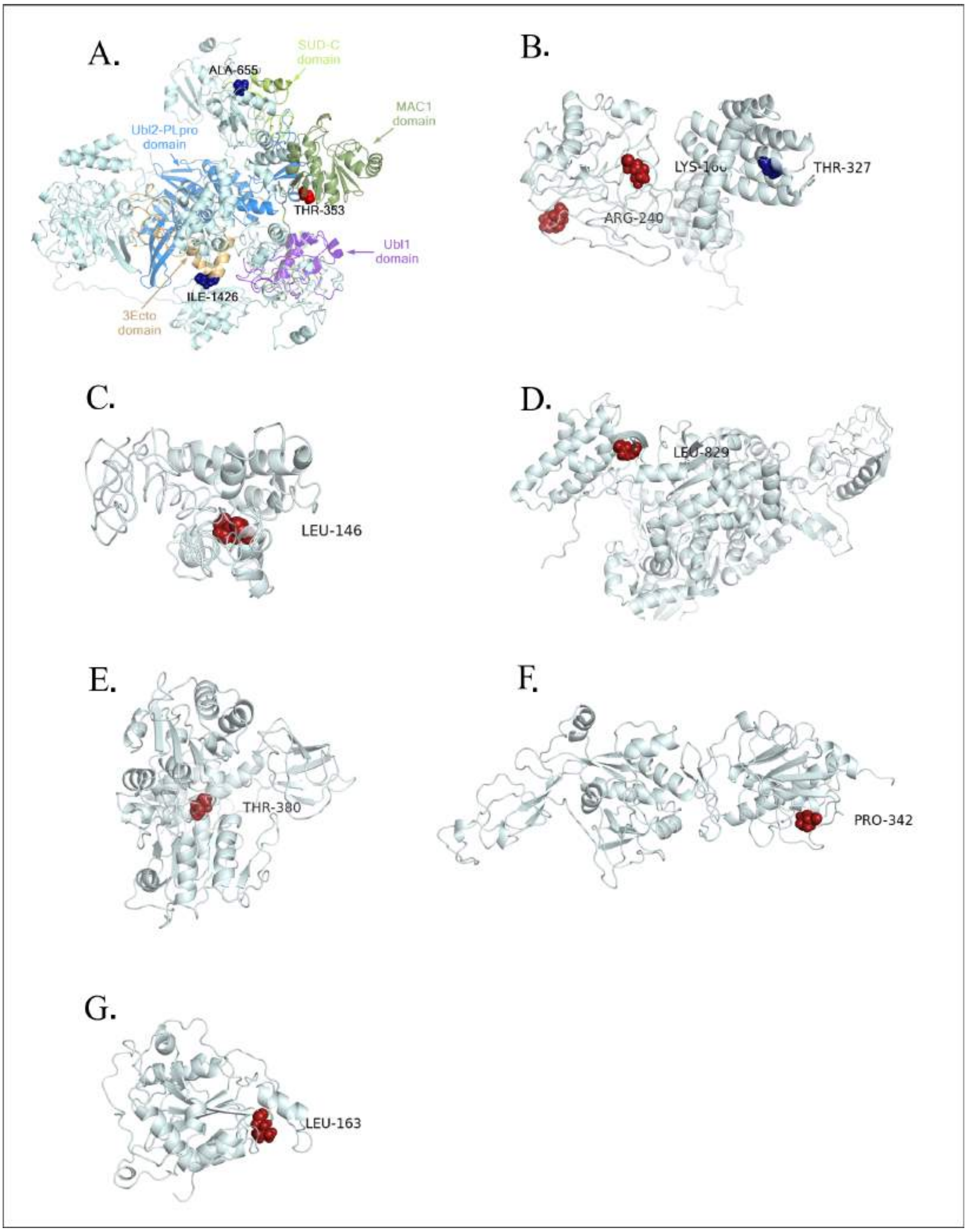

Figure 4.7: **Structural view of selective pressure in orf1ab gene.**

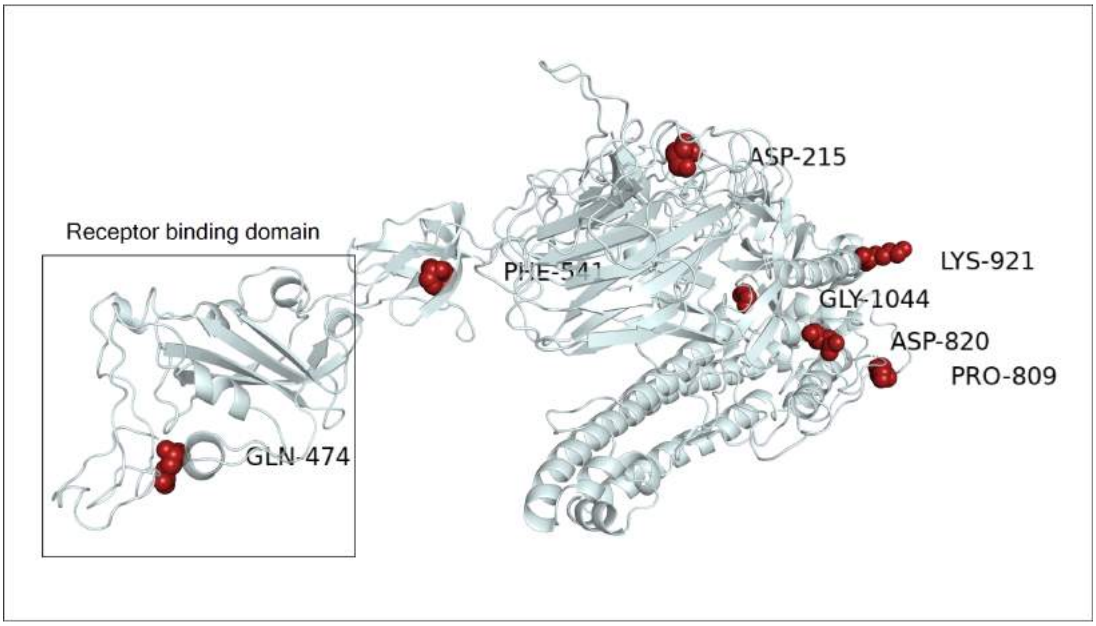

Figure 4.8: **Structural view of selective pressure in spike gene.** The negatively selected site in spike protein is highlighted in red. The only amino acid residue selected negatively on the receptor-binding domain corresponds to GLN-474. The cryo-EM structure with PDB id 6VSB was used as a model for the spike gene in its prefusion conformation.

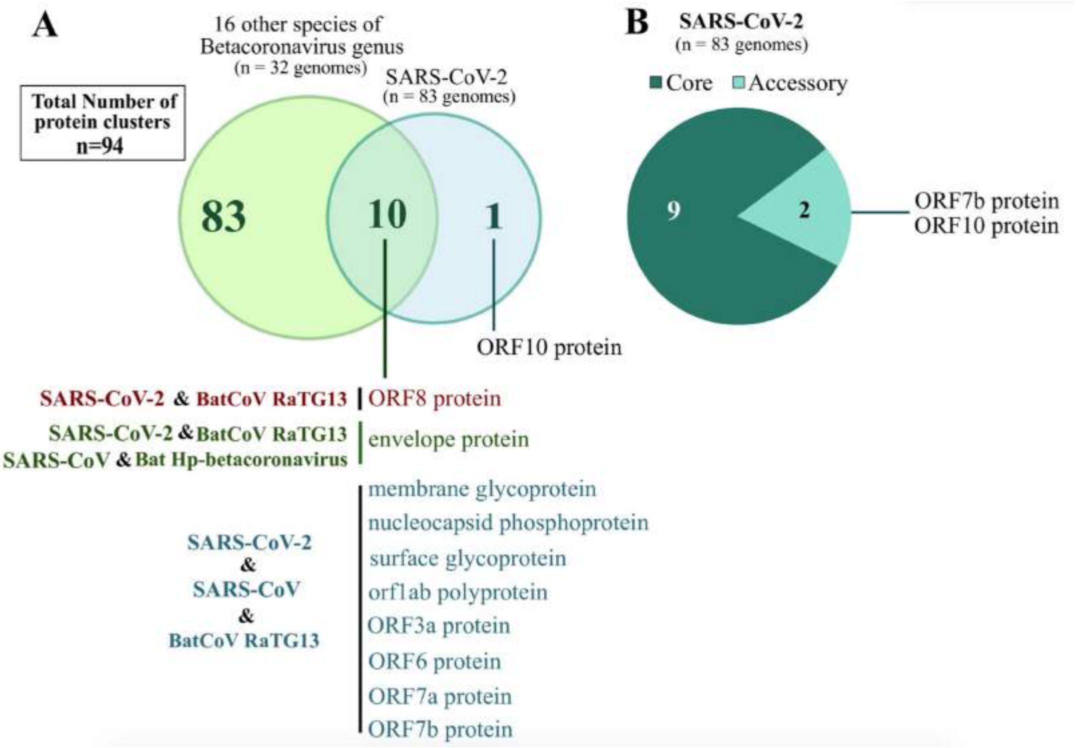
A
16 other species of Betacoronavirus genus (n = 32 genomes)
SARS-CoV-2 (n = 83 genomes)
Total Number of protein clusters n=94
83
10
1
ORF10 protein
SARS-CoV-2 & BatCoV RaTG13
ORF8 protein
SARS-CoV-2 & BatCoV RaTG13
SARS-CoV & Bat Hp-betacoronavirus
envelope protein
SARS-CoV-2 & SARS-CoV & BatCoV RaTG13
membrane glycoprotein
nucleocapsid phosphoprotein
surface glycoprotein
orf1ab polyprotein
ORF3a protein
ORF6 protein
ORF7a protein
ORF7b protein
B
SARS-CoV-2 (n = 83 genomes)
Core
Accessory
9
2
ORF7b protein
ORF10 protein

Figure 4.9: **Pangenome construction of different strains belonging to the genus Betacoronavirus (A)** The Venn diagram represents the shared and unique proteins of SARS-CoV-2 compared to the 16 species of the genus Betacoronavirus. **(B)** The pie diagram showing the core (present in all strains) and accessory proteins (not present in all strains) at the intragenomic scale of SARS-CoV-2.

2.5 Discussion

The rate of mutations results in viral evolution and variability in the genome, thus allowing viruses to escape host immunity, as well as drugs [19]. Initial published data suggests that SARS-CoV-2 is genetically stable [20] which may increase the effectiveness of vaccines under development. The study on the genomic variation of SARS-CoV-2 is very important for the investigation of pathogenesis, disease course, prevention, and treatment of SARS-CoV-2 infection. In this study, we characterized the genetic varia-tions in a large population of SARS-CoV- 2 genomes. Our results showed a diversity of mutations detected with different frequencies. Overall, more than 500 non-synonymous mutations in SARS-CoV-2 genomes have been identified. The orf1ab gene having more than half the size of the SARS-CoV-2 genome and is divided into 16 nsp (nsp1-nsp16)[21]. We found more than half of recurrent mutations in orf1ab, and a high mutation rate in nsp3, nsp12 and nsp2, with 124, 57 and 46, respectively. Nsp2 and nsp3 were both essential for correcting viral replication errors [22]. Thus, recent studies have suggested that mutations falling in the endosome—associated—protein—like domain of then sp2, could explain why this virus is more contagious than SARS [23].
The replication enzymes nsp12 to nsp16 have been reported as antiviral targets for SARS–CoV [24].In the SARS-CoV-2 genomes, we found that nsp12 to nsp15 harbored nine recurrent non-synonymous mutations. Among them, eight identified as new mutations, including three in nsp12-RNA-dependent RNA polymerase (M4555T, T4847I and T5020I),three in nsp13-Helicase (V5661A, P5703L and M5865V) and two in nsp15-EndoRNAse(I6525T and Ter6668W). However, these new mutations must be taken into account when developing a vaccine using the orf1ab protein sequences as a therapeutic target.

A high number of mutations were identified in the spike protein, an important determinant in pathogenicity that allows the virion attachment to the cell membrane by interacting with the host ACE2 (Angiotensin-converting enzyme 2) receptor [25]. Among all the frequent mutations in this protein, the V483A mutation has been identified in this receptor and found mainly in SARS-CoV-2 genomes isolated from USA. This result is consistent with the study of Junxian et al. [26]. Eight stains from china, USA and France harboredV367F mutation previously described to enhance the affinity with ACE2 receptor [26].

Interestingly, ten hyper variable genomic hotspots with high frequencies of mutated allel detected. Among them, position 11083 (L3606F) detected in nsp6, this protein works with nsp3 and nsp4 by forming double-membrane vesicles and convoluted membranes involved in viral replication [27]. Besides, three positions were previously re-ported by Pachetti et al. (2020) [19], of which the two positions 17858 (M5865V) and18060 (S5932F) in orf1ab, and 28881 (R203K) in nucleocapside. Moreover, intraspecies pangenome analysis of SARS- CoV-2 showed that the six of the genes harboring hotspot mutations belong to the core genome. Thus, under normal circumstances genomic variation increase the viruses spread and pathogenicity. This happens when the virus accumulated mutation enabling its virulence potential [28]. Genomic comparison of the studied population allowed us to gain insights into virus mutations occurrence over time and within different geographic areas. In the SARS-CoV virus, the SNPs distribution is not random, and it is more dominant in critical genes for the virus [19][29]. Our results confirmed what was previously described and elucidate the presence of numerous hotspot mutations. Besides, co-occurrence mutations were also common in different countries all along with singleton mutations. In the case of the China, the singleton mutations are driven by the single group that diverged differently due to the environment, the host, and the number of generations. These mutations are due to the low fidelity of reverse transcriptase [28][30].

China, US, France and Malaysia contain a high number of specific mutations which may be the cause of a rapid transmission, especially in the US. These specific mutations may also be correlated with the critical condition in US and France.

The clustering of these genomes revealed the spread of clades to diverse geographical regions. We observed an increase of mutations over time following the first dissemination event from China. Specific haplotypes were also predominant to a geographical location, especially in the China. This study opens up new perspectives to determine whether one of these frequent mutations will lead to biological differences and their correlation with different mortality rates.

Among the seven nsp of or1ab hosting sites under selective pressure, only nsp3 and nsp4 contains both residues underpositive and negative selection. The modelling of nsp3 domains shows that only the negative selection site 1171 (Thr- 353), was located at the conserved macro domain Mac1(previously X or ADP-ribose 1" phosphatase) [31]. This domain has been previ- ously shown to be dispensable for RNA replication in the context of a SARS-CoV replicon [32].However, it could counteract the host's innate immune response [33]. It was proposed that the 3Ecto luminal domain of nsp3 interacts with the large luminal domain of nsp4 (residues 112–164) to "zipper" the endoplasmic reticulum (ER) membrane and induce discrete membrane formations as an important step in the generation of ER-origin viral replication organelles [34][35].

As we have shown previously by the FEL, MEME and FUBAR methods, the orf1ab 2244 site coding for ILE-1426 is under positive selection pressure and since it is located on the luminal 3ecto domain of the nsp3 protein, this can be explained by a possible host influence on the virus in this domain.

The results of selective pressure analysis revealed the presence of several negatively selected residues,one of which is located at the receptor- binding domain (GLN-474) and which is known by its interaction with the GLN24 residue of the human ACE2 receptor [36]. While thisstudy allowed the identification of several site under selective pressure we would like to point that the size of the dataset could be a potential limitation for this type of study,therefore didn't allow the identification of other site under selective pressure. Hence the need for a larger dataset. In general, it is well-known that negatively selected sites could indicate a functional constraint and could be useful for drug or vac- cine target design,given their conserved nature and therefore less likely to change [37].

2.6 Conclusion

The SARS-CoV-2 pandemic has caused a very large impact on health and economy worldwide. Therefore, understanding genetic diversity and virus evolution become a priority in the fight against the disease.

Our results show several molecular facets of the relevance of this virus. We have shown that recurrent mutations are distributed mainly in six SARS-CoV-2 genes with variable mutated allele frequencies. We were able to highlight an increase in mutations accumulation overtime and revealed the existence of three major clades in various geographic regions. Finally, the study allowed us to identify specific haplotypes by geographic location and provides a list of sites under selective pressure that could serve as an interesting avenue for future studies.

2.7 References

1. Ke Wang, Wei Chen, Yu-Sen Zhou, Jian-Qi Lian, Zheng Zhang, Peng Du, Li Gong,Yang Zhang, Hong-Yong Cui, Jie-Jie Geng, et al. Sars-cov-2 invades host cells viaa novel route: Cd147-spike protein.BioRxiv, 2020.

2. Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song,Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirusfrom patients with pneumonia in china, 2019.New England journal of medicine,2020.

3. Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, AhmedAl-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declaresglobal emergency: A review of the 2019 novel coronavirus (covid-19).Internationaljournal of surgery, 76:71–76, 2020.

4. Jose M Cuevas, Ron Geller, Raquel Garijo, Jose Loopez-Aldeguer, and RafaelSanjan. Extremely high mutation rate of hiv-1 in vivo.PLoS Biol, 13(9):e1002251,2015.

5. Barry T Rouse and Sharvan Sehrawat. Immunity and immunopathology to viruses:what decides the outcome?Nature Reviews Immunology, 10(7):514–526, 2010.

6. Aiping Wu, Yousong Peng, Baoying Huang, Xiao Ding, Xianyue Wang, Peihua Niu,Jing Meng, Zhaozhong Zhu, Zheng Zhang, Jiangyuan Wang, et al. Genome com-position and divergence of the novel coronavirus (2019-ncov) originating in china.Cell host and microbe, 27(3):325–328, 2020.

7. Yasmin A Malik. Properties of coronavirus and sars-cov-2.The Malaysian journalof pathology, 42(1):3–11, 2020.

8. Lanying Du, Yuxian He, Yusen Zhou, Shuwen Liu, Bo-Jian Zheng, and Shibo Jiang.The spike protein of sars-cov—a target for vaccine and therapeutic development.Nature Reviews Microbiology, 7(3):226–236, 2009.19 Bibliography20

9. Oliver C Grant, David Montgomery, Keigo Ito, and Robert J Woods. Analysis of thesars-cov-2 spike protein glycan shield reveals implications for immune recognition.Scientific reports, 10(1):1–11, 2020.

10. Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenzadata–from vision to reality.Eurosurveillance, 22(13):30494, 2017.

11. Heng Li. Minimap2: pairwise alignment for nucleotide sequences.Bioinformatics,34(18):3094–3100, 2018.

12. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Ga-bor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/mapformat and samtools.Bioinformatics, 25(16):2078–2079, 2009.

13. Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, LuanWang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotatingand predicting the effects of single nucleotide polymorphisms, snpeff: Snps in thegenome of drosophila melanogaster strain w1118; iso-2; iso-3.Fly, 6(2):80–92, 2012.

14. Meriem Laamarti, Tarek Alouane, Souad Kartti, MW Chemao-Elfihri, MohammedHakmi, Abdelomunim Essabbar, Mohamed Laamarti, Haitam Hlali, Houda Bendani, Nassma Boumajdi, et al. Large scale genomic analysis of 3067 sars-cov-2genomes reveals a clonal geo-distribution and a rich genetic variations of hotspotsmutations.Plos one, 15(11):e0240345, 2020.

15. Kazuma Kiyotani, Yujiro Toyoshima, Kensaku Nemoto, and Yusuke Nakamura.Bioinformatic prediction of potential t cell epitopes for sars-cov-2.Journal of humangenetics, 65(7):569–575, 2020.

16. Sujay Chattopadhyay, Scott J Weissman, Vladimir N Minin, Thomas A Russo,Daniel E Dykhuizen, and Evgeni V Sokurenko. High frequency of hotspot mutationsin core genes of escherichia coli due to short-term positive selection.Proceedings ofthe National Academy of Sciences, 106(30):12412–12417, 2009.

17. Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch,Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a vi-sualization system for exploratory research and analysis.Journal of computationalchemistry, 25(13):1605–1612, 2004.

18. Gaoqi Weng, Ercheng Wang, Zhe Wang, Hui Liu, Feng Zhu, Dan Li, and TingjunHou. Hawkdock: a web server to predict and analyze the protein–protein complexbased on computational docking and mm/gbsa.Nucleic acids research, 47(W1):W322–W330, 2019. Bibliography21

19. Michael Levandowsky and David Winter. Distance between sets.Nature, 234(5323):34–35, 1971.

20. Esko Ukkonen. Approximate string-matching with q-grams and maximal matches.Theoretical computer science, 92(1):191–211, 1992.

21. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.Bioinformatics, 22(13):1658–1659,2006.

22. James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Pot-ter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher.Nextstrain: real-time tracking of pathogen evolution.Bioinformatics, 34(23):4121–4123, 2018.

23. Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh.Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihoodphylogenies.Molec biology and evolution, 32(1):268–274, 2015.

24. Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis.Virus evolution, 4(1):vex042, 2018.

25. Fatima Amanat and Florian Krammer. Sars-cov-2 vaccines: status report.Immunity, 52(4):583–589, 2020.

26. Jinyong Zhang, Hao Zeng, Jiang Gu, Haibo Li, Lixin Zheng, and Quanming Zou.Progress and prospects on vaccine development against sars-cov-2.Vaccines, 8(2):153, 2020.

27. Yung-Fang Tu, Chian-Shiu Chien, Aliaksandr A Yarmishyn, Yi-Ying Lin, Yung-Hung Luo, Yi-Tsung Lin, Wei-Yi Lai, De-Ming Yang, Shih-Jie Chou, Yi-Ping Yang,et al. A review of sars-cov-2 and the ongoing clinical trials.International journalof molecular sciences, 21(7):2657, 2020.

28. Jason W Rausch, Adam A Capoferri, Mary Grace Katusiime, Sean C Patro, and-Mary F Kearney. Low genetic diversity may be an achilles heel of sars-cov-2.Proceedings of the National Academy of Sciences, 117(40):24614–24616, 2020.

29. Carol A Abidha, Joyce Nyiro, Everlyn Kamau, Osman Abdullahi, David James-Nokes, and Charles N Agoti. Transmission and evolutionary dynamics of human-coronavirus oc43 strains in coastal kenya investigated by partial spike sequenceanalysis, 2015–16.Virus Evolution, 6(1):veaa031, 2020. Bibliography22

30. Takahiko Koyama, Dilhan Weeraratne, Jane L Snowdon, and Laxmi Parida. Emergence of drift variants that may affect covid-19 vaccine development and antibodytreatment.Pathogens, 9(5):324, 2020.

31. Alessia Lai, Annalisa Bergna, Sara Caucci, Nicola Clementi, Ilaria Vicenti, FilippoDragoni, Anna Maria Cattelan, Stefano Menzo, Angelo Pan, Annapaola Callegaro,et al. Molecular tracing of sars-cov-2 in italy in the first three months of theepidemic.Viruses, 12(8):798, 2020.

32. M Rafiul Islam, M Nazmul Hoque, M Shaminur Rahman, ASM Rubayet Ul Alam,Masuda Akther, J Akter Puspo, Salma Akter, Munawar Sultana, Keith A Crandall,and M Anwar Hossain. Genome-wide analysis of sars-cov-2 virus strains circulatingworld-wide implicates heterogeneity.Scientific reports, 10(1):1–9, 2020.

33. Erik Alm, Eeva K Broberg, Thomas Connor, Emma B Hodcroft, Andrey B Komissarov, Sebastian Maurer-Stroh, Angeliki Melidou, Richard A Neher, Aine O'Toole,Dmitriy Pereyaslov, et al. Geographical and temporal distribution of sars-cov-2clades in the who european region, january to june 2020.Eurosurveillance, 25(32):2001410, 2020.

34. Paola Stefanelli, Giovanni Faggioni, Alessandra Lo Presti, Stefano Fiore, Antonella-Marchi, Eleonora Benedetti, Concetta Fabiani, Anna Anselmo, Andrea Ciammaruconi, Antonella Fortunato, et al. Whole genome and phylogenetic analysis of twosars-cov-2 strains isolated in italy in january and february 2020: additional clueson multiple introductions and further circulation in europe.Eurosurveillance, 25(13):2000305, 2020.

35. Michael Worobey, Jonathan Pekar, Brendan B Larsen, Martha I Nelson, Verity Hill,Jeffrey B Joy, Andrew Rambaut, Marc A Suchard, Joel O Wertheim, and PhilippeLemey. The emergence of sars-cov-2 in europe and north america.Science, 370(6516):564–570, 2020.

36. Javaid Ahmad Sheikh, Jasdeep Singh, Hina Singh, Salma Jamal, Mohd Khubaib,Sunil Kohli, Ulrich Dobrindt, Syed Asad Rahman, Nasreen Zafar Ehtesham, andSeyed Ehtesham Hasnain. Emerging genetic diversity among clinical isolates ofsars-cov-2: Lessons for today.Infection, Genetics and Evolution, 84:104330, 2020.

37. Leolin Katsidzira, Lenon Gwaunza, and James G Hakim. The severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic in zimbabwe: Quo vadis?Clinical Infectious Diseases, 71(16):2180–2183, 2020. Bibliography23

38. Marguerite Massinga Loembe, Akhona Tshangela, Stephanie J Salyer, Jay KVarma, Ahmed E Ogwell Ouma, and John N Nkengasong. Covid-19 in africa:the spread and response.Nature Medicine, 26(7):999–1003, 2020.

39. Julio A Poterico and Orson Mestanza. Genetic variants and source of introductionof sars-cov-2 in south america.Journal of medical virology, 92(10):2139–2145, 2020.

40. Bette Korber, Will Fischer, S Gnana Gnanakaran, Heyjin Yoon, James Theiler,Werner Abfalterer, Brian Foley, Elena E Giorgi, Tanmoy Bhattacharya, Matthew DParker, et al. Spike mutation pipeline reveals the emergence of a more transmissibleform of sars-cov-2.BioRxiv, 2020.

41. Jie Hu, Chang Long He, Qingzhu Gao, Gui Ji Zhang, Xiao Xia Cao, Quan Xin Long,Hai Jun Deng, Lu Yi Huang, Juan Chen, Kai Wang, et al. The d614g mutation ofsars-cov-2 spike protein enhances viral infectivity.BioRxiv, 2020.

42. T Thanh Le, Zacharias Andreadakis, Arun Kumar, R Gomez Roman, Stig Tollefsen, Melanie Saville, Stephen Mayhew, et al. The covid-19 vaccine development-landscape.Nat Rev Drug Discov, 19(5):305–306, 2020.

43. Sandra Isabel, Lucıa Grna-Miraglia, Jahir M Gutierrez Helen E Groves, Marc R Isabel, AliReza Eshaghi, Samir N Patel,Jonathan B Gubbay, Tomi Poutanen, et al. Evolutionary and structural analysesof sars-cov-2 d614g spike protein mutation now documented worldwide.Scientificreports, 10(1):1–9, 2020.

44. Xianding Deng, Wei Gu, Scot Federman, Louis Du Plessis, Oliver G Pybus, Nuno RFaria, Candace Wang, Guixia Yu, Brian Bushnell, Chao-Yang Pan, et al. Genomicsurveillance reveals multiple introductions of sars-cov-2 into northern california.Science, 369(6503):582–587, 2020.

45. Leyan Tang, Allison Schulkins, Chun-Nan Chen, Kurt Deshayes, and John S Kenney. The sars-cov-2 spike protein d614g mutation shows increasing dominance andmay confer a structural advantage to the furin cleavage domain. 2020.

46. Xiaoli Xiong, Kun Qu, Katarzyna A Ciazynska, Myra Hosmillo, Andrew P Carter,Soraya Ebrahimi, Zunlong Ke, Sjors HW Scheres, Laura Bergamaschi, Guinevere LGrice, et al. A thermostable, closed sars-cov-2 spike protein trimer.Nature Structural and Molecular Biology, 27(10):934–941, 2020.

47. Nathan D Grubaugh, William P Hanage, and Angela L Rasmussen. Making senseof mutation: what d614g means for the covid-19 pandemic remains unclear.Cell,182(4):794–795, 2020. Bibliography24

48. Wanbo Tai, Lei He, Xiujuan Zhang, Jing Pu, Denis Voronin, Shibo Jiang, YusenZhou, and Lanying Du. Characterization of the receptor-binding domain (rbd)of 2019 novel coronavirus: implication for development of rbd protein as a viralattachment inhibitor and vaccine.Cellular and molecular immunology, 17(6):613–620, 2020.

49. Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, QibinGeng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition bysars-cov-2.Nature, 581(7807):221–224, 2020.

50. Alan HM Wong, Aidan CA Tomlinson, Dongxia Zhou, Malathy Satkunarajah,Kevin Chen, Chetna Sharon, Marc Desforges, Pierre J Talbot, and James M Rini.Receptor-binding loops in alphacoronavirus adaptation and evolution.Nature com-munications, 8(1):1–10, 2017.

51. Barry Rockx, Eric Donaldson, Matthew Frieman, Timothy Sheahan, Davide Corti,Antonio Lanzavecchia, and Ralph S Baric. Escape from human monoclonal anti-body neutralization affects in vitro and in vivo fitness of severe acute respiratorysyndrome coronavirus.The Journal of infectious diseases, 201(6):946–955, 2010.

52. Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthenedsars-cov-2 infectivity.Journal of molecular biology, 432(19):5212–5226, 2020.

53. Xiaojun Li, Elena E Giorgi, Manukumar Honnayakanahalli Marichannegowda,Brian Foley, Chuan Xiao, Xiang-Peng Kong, Yue Chen, S Gnanakaran, Bette Ko-rber, and Feng Gao. Emergence of sars-cov-2 through recombination and strongpurifying selection.Science Advances, 6(27):eabb9153, 2020.

54. Junxian Ou, Zhonghua Zhou, Ruixue Dai, Jing Zhang, Wendong Lan, Shan Zhao,Jianguo Wu, Donald Seto, Lilian Cui, Gong Zhang, et al. Emergence of rbd muta-tions in

circulating sars-cov-2 strains enhancing the structural stability and humanace2 receptor affinity of the spike protein.BioRxiv, 2020.

55. Hasan Uludag, Kylie Parent, Hamidreza Montazeri Aliabadi, and Azita Haddadi.Prospects for rnai therapy of covid-19.Frontiers in bioengineering and biotechnol-ogy, 8:916, 2020.

56. Cynthia Liu, Qiongqiong Zhou, Yingzhu Li, Linda V Garner, Steve P Watkins,Linda J Carter, Jeffrey Smoot, Anne C Gregg, Angela D Daniels, Susan Jervey,et al. Research and development on therapeutic agents and vaccines for covid-19and related human coronavirus diseases, 2020.

57. Sanhita Ghosh, Sayeed Mohammad Firdous, and Anirban Nath. sirna could be apotential therapy for covid-19.EXCLI journal, 19:528, 2020. Bibliography25

58. SHI Yi, YANG De Hua, Jie XIONG, JIA Jie, Bing HUANG, and You Xin JIN.Inhibition of genes expression of sars coronavirus by synthetic small interfering rnas.Cell research, 15(3):193–200, 2005.

59. T Li, Y Zhang, L Fu, C Yu, X Li, Y Li, X Zhang, Z Rong, Y Wang, H Ning,et al. sirna targeting the leader sequence of sars-cov inhibits virus replication.Gene therapy, 12(9):751–761, 2005.

60. Chang-Jer Wu, Hui-Wen Huang, Chiu-Yi Liu, Cheng-Fong Hong, and Yi-Lin Chan.Inhibition of sars-cov replication by sirna.Antiviral research, 65(1):45–48, 2005.

61. John Hodgson. The pandemic pipeline.Nature biotechnology, 38(5):523–532, 2020.

62. Wei Chen, Pengmian Feng, Kewei Liu, Meng Wu, and Hao Lin. Computationalidentification of small interfering rna targets in sars-cov-2.Virologica Sinica, 35(3):359–361, 2020

3 Genomic Diversity and Hotspot Mutations in 30,983 SARS-CoV-2 Genomes: Moving Toward a Universal Vaccine for the "Confined Virus"?

M.Laamarti[a], T. Alouane[a], A. ESSABBAR[a], M. Hakmi[a], W Chemao[a], S. Kartti[a], N. BOUMAJDI[a], H. BENDANI[a], R. Laamarti[b], F. GHRIFI[a], L. ALLAM[a], T. AANNIZ[a], M. OUADGHIRI[a], N. EL HAFIDI[a],R. EL JAOUDI[a], H. Benrahma[c], J. EL EL ATTAR[d], R. MENTAG[e], L. SBABOU[f], C. NEJJARI[j],S. AMZAZI[h], L. BELYAMANI[i], A. Ibrahimi [a]

[a] Biotechnology Lab (MedBiotech Center), Rabat Medical and Pharmacy School, University Mohammed V, Rabat, Morocco
[b]Medical Biotechnology Center, Moroccan Foundation for Science, Innovation and Research (MAScIR)
[c]Faculty of Medicine, Mohammed VI University of Health Sciences (UM6SS), Casablanca
R[d]riad Laboratory, City Center Hay Riad, Rabat 10112, Morocco.
[e]Biotechnology Unit, Regional Center of Agricultural Research of Rabat, National Institute of Agricultural Research, Rabat
[f]Microbiology and Molecular Biology Team, Center of Plant and Microbial Biotechnology, Biodiversity and Environment, Faculty of Sciences, Mohammed V University, Rabat.
[j]International School of Public Health, Mohammed VI University of Health Sciences (UM6SS), Casablanca
[h]Laboratory of Human Pathologies Biology, Faculty of Sciences, Mohammed V University
[i]Emergency Department, Military Hospital Mohammed V, Rabat Medical and Pharmacy School, Mohammed Vth University.

3.1 Abstract

The COVID-19 pandemic has been ongoing since its onset in late November 2019 in Wuhan, China. Understanding and monitoring the genetic evolution of the virus, its geographical characteristics, and its stability are particularly important for controlling the spread of the disease and especially for the development of a universal vaccine covering all circulating strains. From this perspective, we analyzed 30,983 complete SARS-CoV-2 genomes from 79 countries located in the six continents and collected from 24 December 2019, to 13 May 2020, according to the GISAID database.

Our analysis revealed the presence of 3206 variant sites, with a uniform distribution of mutation types in different geographic areas. Remarkably, a low frequency of recurrent mutations has been observed; only 169 mutations (5.27%) had a prevalence greater than 1% of genomes. Nevertheless, fourteen non-synonymous hotspot mutations (¿10%) have been identified at different locations along the viral genome; eight in ORF1ab polyprotein

(in nsp2, nsp3, transmembrane domain, RdRp, helicase, exonuclease, and endoribonuclease), three in nucleocapsid protein, and one in each of three proteins: Spike, ORF3a, and ORF8. Moreover, 36 non-synonymous mutations were identified in the receptor-binding domain (RBD) of the spike protein with a low prevalence (¡1%) across all genomes, of which only four could potentially enhance the binding of the SARS-CoV-2 spike protein to the human ACE2 receptor. These results along with intra-genomic divergence of SARS-CoV-2 could indicate that unlike the influenza virus or HIV viruses, SARS-CoV-2 has a low mutation rate which makes the development of an effective global vaccine very likely.

3.2 Introduction

The year 2019 ended with the appearance of groups of patients with pneumonia of unknown cause. Initial evidence suggested that the outbreak was associated with a sea food market in Wuhan, China, as reported by local health authorities [1]. The results of the investigations led to the identification of a new coronavirus in affected patients[2] Following its identification on the 7 January 2020 by the Chinese Center for Disease Control and Prevention (CCDC), the new virus and the disease were officially named SARS-CoV-2 (for severe acute respiratory syndrome coronavirus-2) and COVID-19 (for coronavirus disease 19), respectively, by the World Health Organization (WHO) [3]. On 11 March 2020, WHO publicly announced the SARS-CoV-2 epidemic as a global pandemic. This virus is likely to remain and continue to spread unless an effective vaccine is developed, or a high percentage of the population is infected in order to achieve collective immunity. The development of a vaccine is a long process and is not guaranteed for all infectious diseases. Indeed, some viruses such as influenza and HIV have a high rate of genetic mutations, which makes them prone to antigenic leakage[4][5]. It is therefore important to assess the genetic evolution of the virus and morespecifically the regions responsible for its interaction and replication within the host cell. Thus, identifying the conserved and variable regions of the virus could help guide the design and development of anti-SARS-CoV-2 vaccines.

SARS-CoV-2 is a single-stranded positive-sense RNA virus belonging to the genus Betacoronavirus. The genome size of SARS-CoV-2 is approximately 30 kb and its genomic structure has followed the characteristics of known genes of the coronavirus [6]. The ORF1ab polyprotein is covering two-thirds of the viral genome and cleaved into many nonstructural proteins(nsp1 to nsp16). The third part of the SARS-CoV-2 genome codes for the main structural proteins; spike (S), envelope (E), nucleocapsid (N), and membrane (M). In addition, sixORFs, namely ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10, are predicted ashypothetical proteins with no known function [7].

Protein S is the basis of most candidate vaccines; it binds to membrane receptors in host cells via its RBD and ensures a viral fusion with the host cells [8]. Its main receptor is the angiotens in-converting enzyme 2(ACE2), although another route via CD147 has also been described [9][10]. The glycans attached to S protein assist the initial attachment of the virus to the host cells and act asa coat that helps the virus to evade the host's immune system. In fact, a previous study has shown that glycans cover about 40% of the surface of the spike protein. However, theACE2-RBD was found to be the largest and most accessible epitope [11]. Thus, it maybe possible to develop a vaccine that targets the spike receptor-binding domain (RBD),provided it remains accessible and stable over time; hence, the importance of monitoring the introduction of any mutation that could compromise the potential effectiveness of a candidate vaccine.

This study aims to deepen our understanding of the intra-genomic diversity of SARS-CoV-2, by analyzing the mutational frequency and divergence rate in 30,983 genomes from six geographic areas (Africa, Asia, Europe, North and South America, and Oceania), collected during the first five months after the onset of the virus. These analyses generate new datasets providing a repository of genetic variants from different geographic areas, with particular emphasis on recurrent mutations and their distribution along the viral genome as well as estimating the rate of intraspecific divergence while evaluating the adaptation of SARS-CoV-2 to its host and the possibility of developing a universal vaccine.

3.3 Materials and methods

3.3.1 Data Collection

Full-length viral nucleotide sequences of 30,983 SARS-CoV-2 genomes were collected from the GISAID EpiCovTM (update: 26 May 2020) [10], belonging to the six geographic areas (according to GISAID database) and distributed in 79 countries as follows: 214from Africa, 368 from South America, 1590 from Oceania, 2111 from Asia, 6825 from North America, and 19,875 from Europe. The genomes were obtained from samples collected from 24 December 2019 to 13 May 2020.

For each geographical area the collection date of the samples is from 27 February to 1 May for Africa, 24 December to13 May for Asia, 23 January to 10 May for Europe, 19 January to 12 May for North America, 25 February to 19 April for South America, and 24 January to 21 April for Oceania (Supplementary Table S1).

3.3.2 Variant Calling Analysis

Genome sequences were mapped to the reference sequence Wuhan-Hu-1/2019 (Gen-bank ID: NC045512.2) using Minimap v2.12-r847 [11]. The BAM files were sortedby SAMtools sort [12]. The final sorted BAM files were used to call the genetic variants in variant call format (VCF) by SAMtools mpileup and BCFtools [12]. The final call set of the 30,983 genomes was annotated and their impact was predicted usingSnpEff v 4.3t [13]. For that, the SnpEff databases were first built locally using an-notations of the reference sequence Wuhan-Hu-1/2019 obtained in the GFF format from the NCBI database. Then, the SnpEff database was used to annotate SNPs and InDels with putative functional effects according to the categories defined in the SnpEff manual (http://snpeff.sourceforge.net/SnpEffmanual .html). The frequency of each identified mutation was estimated by normalizing the number of genomes harboring a given mutation, per the total number of genomes recovered from each of the six geographic areas. Non-synonymous mutations with a frequency of 10% or greater were used as a cutoff to define the most frequent mutations [14][15]. Indeed, given that hotspot mutations are known to be strong evidence of positive selection [16] and that sites harboring these mutations have been previously reported under positive selection(http://covid19.datamonkey.org/), we systematically considered the non-synonymous mutation with a frequency $\geq 10\%inthegenomesofsixgeo$

3.3.3 D614G Mutagenesis Analysis

To investigate the possible impact of the most frequent D614G mutation, we conducted an in silico mutagenesis analysis based on the CryoEM structure of the spike protein inits pre-fusion conformation (PDB id 6VSB). Modeling of the D614G mutation was done using UCSF Chimera [17]. Then, the mutant structure was relaxed by 1000 steps of steepest descent (SD) and 1000 steps of conjugate gradient (CG) energy minimization skeeping all atoms with more than 5A from G614 fixed. Comparative analysis of D614(wild type) and G614 (mutant) interactions with their surrounding residues was done in PyMOL (Schrodinger L.L.C).

3.3.4 RBD Mutations and Spike/ACE2 Binding Affinity

Modeling of RBD mutations was performed by UCSF chimera [58] using the 6M0J structure of the SARS-CoV-2 wild-type spike in complex with human ACE2 as a template.Mutant models were relaxed by 1000 steps of SD followed by 1000 steps of CG minimization keeping all atoms far by more than 5A from the mutated residues fixed. Changes in the binding affinity of the spike/ACE2 complex for each spike mutant were estimated by the MM-GBSA method using the HawkDock server [18].

3.3.5 Clustering and Divergence Analysis

In this work, we use the Jaccard distance to compare the similarity of mutational profile of SARS-CoV-2 genomes between 79 countries. It is a metric particularly suited for clustering and useful when the sets to be compared are of different sizes, because its normalization is designed to take the union of the two sets. We first calculated the mutational frequency in each country individually.

The Jaccard similarity coefficient,also known as the Jaccard index, is defined as the ratio of the size of the intersection(shared mutational profile) divided by the union (union of mutational profiles) of two sets A, B (Equation (1)) [19]:

$$\mathrm{J(A,\ B)} = (\mathrm{A}\ \cap B)/(A \cup B)$$

Then, the Jaccard index was converted into the Jaccard distance which is noted as the difference between one and the Jaccard similarity coefficient (Equation (2)), and is re-lated to the q-gram distance but without the number of occurrences [20].

$$\mathrm{d\_J(A,\ B)} = 1 \quad \mathrm{J\ (A,\ B)}$$

The similarity of the set is based on the Jaccard distance. A distance of zero is equivalent to a 100% overlap between countries.

On the other hand, to calculate the intra-genomic divergence of SARS-CoV-2, we used the Wuhan-Hu-1/2019 genome as a reference sequence, and the other 30,983 genomes were also sorted by country of origin. The divergence was first calculated by estimating the similarities of the genomes with the reference sequence by grouping genomes from the same country using CD-Hit [21]. All SARS-CoV-2 genomes used in this study were included except those from Ghana which were excluded from this analysis due to the high number of Ns. The percentage of similarity was then recovered to 100%. Then the percentage of divergence for each country was calculated using the following formula (Equation (3)):

$$\left(100 - \frac{\sum(A \times B)}{C}\right)$$

A = Similarity percentage;
B = Number of genomes with similarity value equal A;
C = Total genomes by country continent;
D = Percentage of divergence.

3.3.6 Phylogenetic and Spatio-Dynamic Analysis

We generated a phylogenetic and divergence tree, as well as a genomic epidemiology map based on the 30,983 genomes of SARS-CoV-2 using NextStrain tools (https://nextstrain.org)[22]. The tree was constructed in IQ-TREE v1.5.5 [23] using the maximum likelihood method under the GTR model. The rate of evolution and the time to the most recent common ancestor (TMRCA) were estimated using ML dating in the tree time package[24]

3.4 Results

3.4.1 Diversity of Genetic Variants of SARS-CoV-2 in Different Geo-graphic Areas

A total of 30,983 SARS-CoV2 genomes from 79 countries in six geographic areas (Africa,Asia, Europe, North and South America, and Oceania) were included in this analysis.According to the GISAID database, the date of collection of the strains was within the first five months following the onset of SARS-CoV-2 (Supplementary Table S1). A total of 3206 variant sites were detected compared to the reference genome Wuhan-Hu-1/2019 (Supplementary Table S2). Then, we analyzed the type of each mutation,highlighting the prevalence of these mutations both in all genomes (worldwide) and in each of the geographic areas studied **(Figure 4.10)**.

Worldwide, 67.96% of mutations had a non-synonymous effect (64.16% have missense effects, 3.77% produce a gain or loss of stop codon, and 0.33% produce a loss of start codon), 28.60% were synonymous, while 3.43% of the mutations were localized in the intergenic regions, mainly in the untranslated regions (UTR). Likewise, the comparison between the six geographic areas shows a similar trend with a uniform distribution of the prevalence of each type of mutation. The frequency of mutations in the six geographic areas was estimated by normalizing the number of genomes carrying given mutation per the total number of genomes recovered by geographic area. Only169 (5.27%) variant

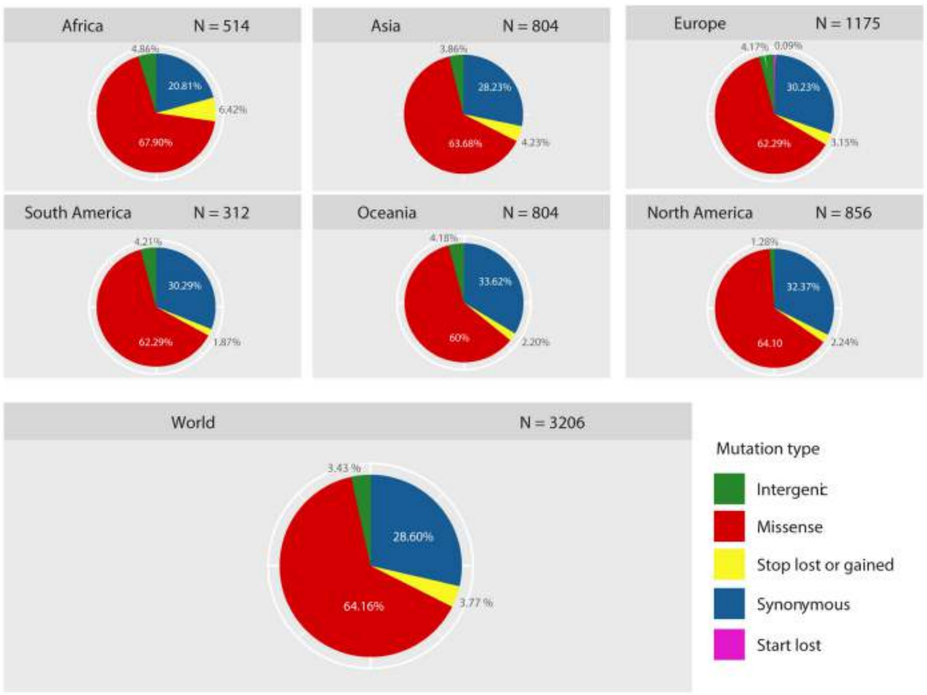

Figure 4.10: **Prevalence and distribution of types of mutations in different geographic regions**. Pie charts showing the global and continent-stratified distribution of the mutation types identified in the 30,983 SARS-CoV-2 genomes. The prevalence of each type of mutation is uniform across the six geographic areas and missense mutations were the most frequent type. Color codes represent the type of mutations.

sites were found with a frequency greater than 0.01 **(Figure 4.11 - A**,Supplementary Table S3), and were distributed in six geographic areas as follows: 69 in Oceania, 65 in Africa, 54 in Asia, 31 in Europe, 43 in North America, and 43 in South America. Focusing on non-synonymous mutations (with a frequency (0.01), 3.34% (n =107) of the total mutations were identified **(Figure 4.11 - B)**. The polyprotein ORF1ab contained approximately two thirds of these mutations (63.55%;n = 68) and distributed in thirteen non-structural proteins; nsp3-Multi-domain: 15.89%, nsp2: 11.21%, nsp12-RNA-dependent RNA polymerase (RdRp): 8.41%, nsp4-trans membrane domain-2 (TM-2): 4.67%, nsp13-helicase: 4.67%, nsp15-endoribonuclease (NendoU):4.67%, nsp5-main proteinase (Mpro):3.74%, nsp14-exonuclease (ExoN): 3.74%, nsp6-TM: 2.80%, nsp1: 0.93%, nsp7: 0.93%,nsp8: 0.93%, and nsp16-2'-O-ribose methyltransferase (OMT): 0.93%. The rest (36.45%)were distributed in eight proteins, including N (11.21%), S (8.41%), ORF3a (5.61%),ORF8 (4.67%), M (1.87%), ORF6 (1.87%), ORF7a (1.87%), and ORF7b (0.93%).

Comparative analysis of these non-synonymous mutations shows only nine that have been shared in the six geographic areas: T265I (nsp2), L3606F (in nsp6-TM) T4847I (innsp12-RdRp), D614G (in S), R203K-G204R (in N), Q57H-G251V (in ORF3a), andL84S (in ORF8). It is also interesting to note that none of the nine non-synonymous mutations (¿0.01) of S protein was localized in RBD. The 36 non-synonymous mutations (35 with a missense effect and 1 with a stop gain effect) found in this area had a low frequency (¡0.01) across all genomes (Appendix T**bale 6.3**). Among them, only two mutations were shared between genomes of different geographic areas; the V367F mutation was identified in Europe, Asia, and North America, the V367F mutation has been identified in Europe,Asia, and North America, while P491L in Asia and Oceania.

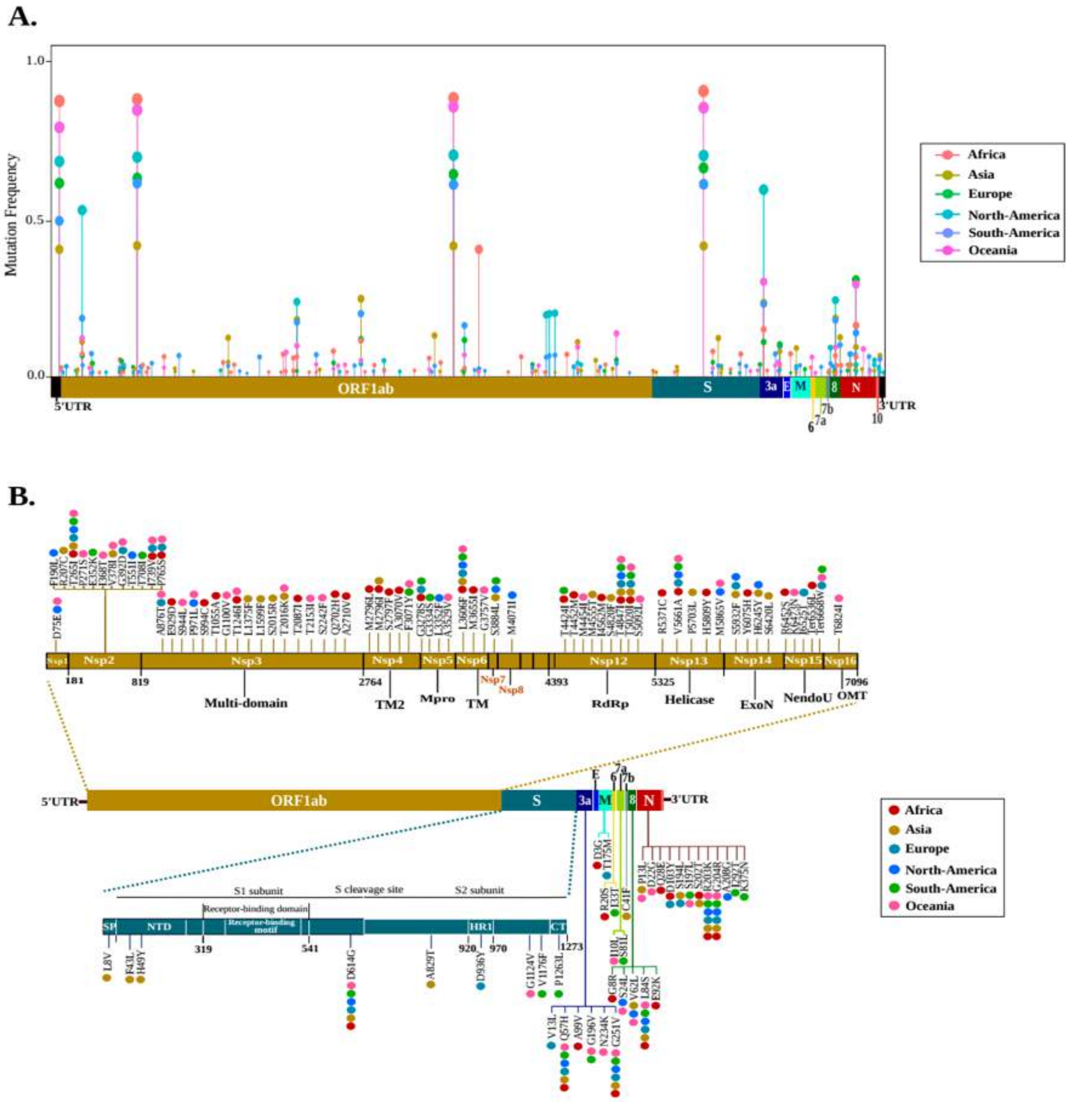

Figure 4.11: **Distribution of recurrent mutations across the SARS-CoV-2 genome**. **(A)** Lollipop plot illustrating the location of mutations with a frequency greater than 0.01 of the total genomes of each geographic area. All types of mutations are included (non-synonymous, synonymous, and intergenic). The mutation frequency was estimated for each of them, by normalizing the number of genomes harbored in a given mutation in a geographic area, per the total number of genomes recovered by geographic area. **(B)** Schematic representation illustrating the distribution of non-synonymous mutations (with a frequency ¿0.01) along the viral genome. Amino acid mutations are shown by vertical lines. Colored dots represent geographic areas

3.4.2 Geographical Distribution of the SARS-CoV-2 Hotspot Mutations

Comparative genomic analysis of each geographic area revealed fourteen non-synonymous mutations with a frequency greater than 0.1 and considered as hotspot mutations (**Figure 4.12**). Eight mutations of them were found in the ORF1ab polyprotein, distributed in seven regions coding for nsp2 (T265I), nsp3-Multi-domain (T2016K), nsp6-TM (L3606F),nsp12-RdRp (T5020I and T4847I), nsp13-helicase (M5865V), nsp14-ExoN (D5932T) and nsp15-NendoU (Ter6668W). Moreover, three mutations in N protein (R203K, G204R,and P13L) and one in each of the three proteins; S (D614G), ORF3a (Q57H), and ORF8(L84S).

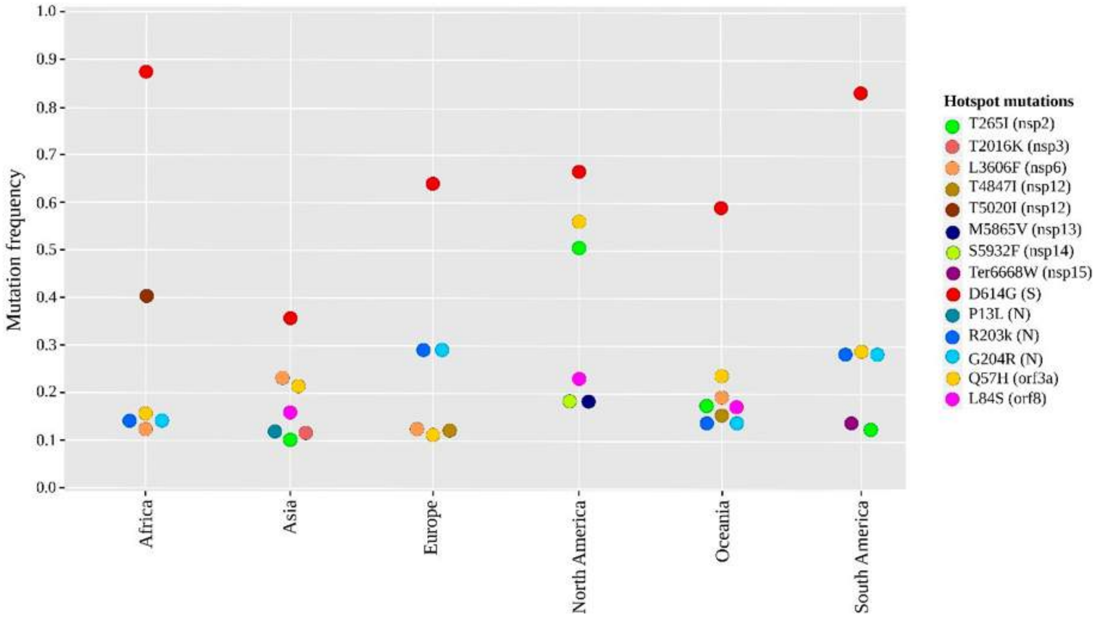

Figure 4.12: **Frequencies of recurrent hotspot mutations per geographic area**.Distribution of fourteen non-synonymous mutations with a frequency ¿0.1 of the genomes subdivided into six geographical areas; Africa (n = 6), Asia (n = 7), Europe (n = 6), North America (n = 6), Oceania (n = 8), South America (n = 6). The locations of mutations in viral proteins with their color codes are indicated in the legend.

Different patterns of these non-synonymous hotspot mutations were observed between the six geographic regions. Only two mutations were common in the six geo-graphical regions: The high-frequency mutation D614G (in S) and the Q57H mutation(in ORF3a). Seven mutations were more frequent in a single geographic region, including two mutations T2016K (in nsp3-Multi-domain) and P13L (in N) in Asia, two mutations M5865V (in nsp13-helicase) and D5932T (in nsp14-ExoN) in North America, one T5020I(in nsp12-RdRp) in Africa, one T4847I (in nsp12-RdRp) in Europe, and one Ter6668W(in nsp15-NendoU) in South America. However, the other five non-synonymous hotspot mutations were variable between the six geographical regions, including two R203K and G204R (in N) that were particularly predominant in Africa, Europe, South America,and Oceania; whereas, L3606F (in nsp6-TM) was common in Africa, Asia, Europe, and Oceania. Thus, L84S (in ORF8) was found in Asia, North America, and Oceania. In addition, T265I (in nsp2) was frequent in Asia, North America, South America, and Oceania.

3.4.3 The Distribution of Hotspot Mutation Patterns of SARS-CoV-2over Time

A different pattern of hotspot mutations over time is clearly distinguished between the six continents from January to May 2020 (**Figure 4.13**). The number of mutations was normalized for each of the six geographic areas for 15 days per the total number of genomes recovered during this period (depending on the date of sample collection provided by GISAID). The L84S mutation (red) was the first mutation observed (early January in Asia) and the most propagated between January–February in North America and Oceania, before starting to drop dramatically after. Remarkably, the D614G (orange) was the most common on six continents. This mutation first appeared on 24 January 2020,in Asia (China), after four days it was observed in Europe (Germany), and then gained its predominance over time, when the outbreak of positive cases was reported in the United States and Canada (Supplementary Table S2). The highest recorded frequency of D614G was in Africa; this mutation was present in most African genomes from late February to May, with a small fluctuation in frequency in mid-March. On the same continent, the frequency of genomes containing the T5020I (lawn green) mutation increased until the end of April before dis-appearing in May.

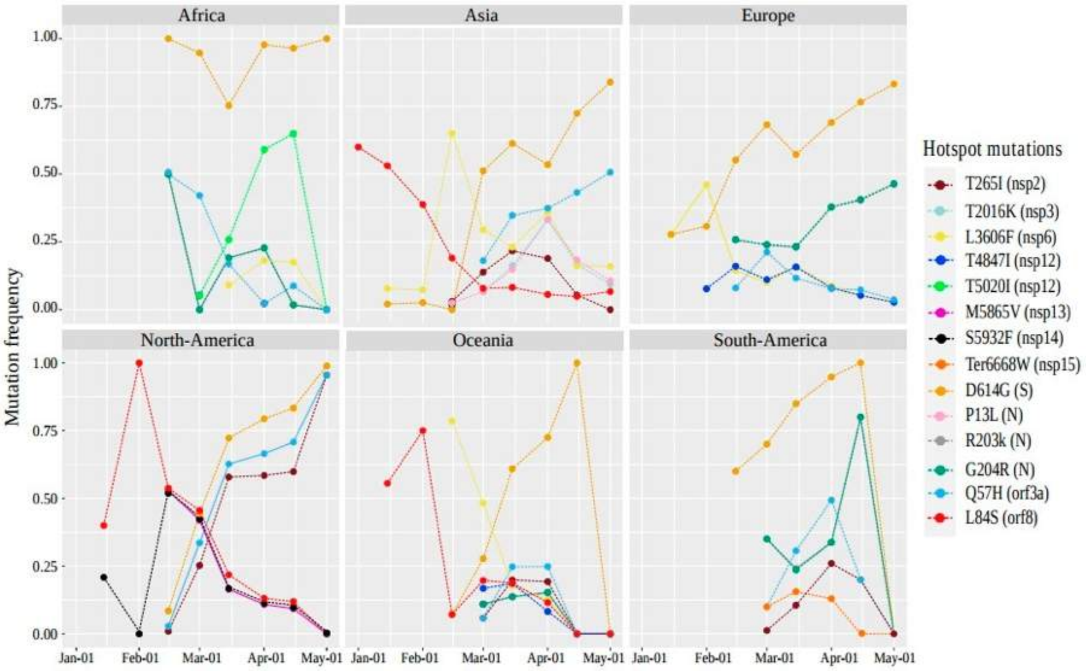

Figure 4.13: **Tracking hotspot mutations over time per geographic area.** Hotspot mutation frequencies were plotted for each of them over a period of 15 days in each geographic area, first by normalizing the number of genomes harboring a given mutation in a period of 15 days, per the total number genomes recovered at this time for each of six continents. The X axis represents the time measured in 15 days and the Y axis represents the frequencies of the genomes harboring the hotspot mutations.

The other three mutations (Q57H-sky blue, R203K-gray, and G203R-green) started with a high frequency (0.5) at the beginning of March and decreased slightly over time.

While in Europe, except for D614G, the two R203K-G203R mutations were the most prevalent, showing continuous growth with the same frequencies overtime. In addition, the L3606F (yellow) mutation showed an increase during February,followed by a decrease to nearly 0.1 frequency in early May.

For North America, three hotspot mutations, D614G, Q57H, and T265I (garnet red), continued to increase with the same trend after their appearance. Unlike the other three mutations (L84S, S5932F-black, and M5865V-dark pink), their frequencies were reduced over time especially from mid-February. Interestingly, a different pattern of hotspot mutations was observed in South America and Oceania between March and April. Focusing on South America,a new stop-loss Ter6668W (dark orange) hotspot mutation (in nsp15-NendoU) was re-ported in North American genomes from March and decreased one month later, while at that date, the frequency of the co-occurring mutations R203K-G203R was increased over time. Overall, the fourteen hotspot mutations were seen between January–March and most of them gained their frequency outside of Asia.

3.4.4 Mutagenesis of D614G and Impact of RBD Mutations on the Binding Ability of Spike to ACE2

As shown the Figure 5, the non-synonymous D614G mutation did not have an impact on the two or three-dimensional structure of the spike glycoprotein. However, D614residue in the wild-type spike is involved in three hydrogen bonds; one with A647 in the same subunit (S1) and two bonds with THR-859 and LYS-854 located at S2 subunit of the adjacent protomer **(Figure 4.14 - A)**.
The substitution of D614 by G in the mutant spike resulted in the loss of the two hydrogen bonds with THR-859 and LYS-854 in the S2 subunit of the adjacent promoter **(Figure 4.14 - B)**. Such modification could result in a weak interaction between S1 and S2 subunits and thus increase the rate of S1/S2 cleavage, which would improve the virus entry to host cells. To evaluate the effect of RBD mutations on the binding affinity of the spike protein to ACE2, the Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) method was employed to calculate the binding affinity of 35 spike mutants to ACE2 (except for the stop-gain mutation). Four mutations potentially enhanced the binding affinity of spike/ACE2 complex by a binding affinity change (G) ¡ 1.0 kcal/mol, while nine were shown to potentially reduce its affinity by a G ¿ 1.0 kcal/mol **(Table 4.2)**. However, the remaining 22 did not significantly affect the binding affinity of spike to ACE2.

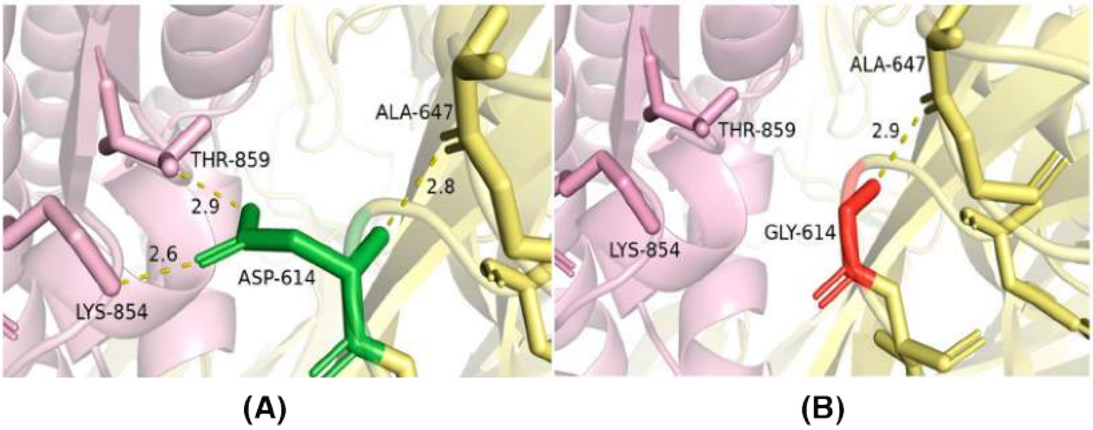

Figure 4.14: Comparison of spike wild-type residue ASP-614 (A) and the mutated GLY-614 (B).
ASP-614 (green sticks) in subunit S1 (yellow) is involved in two hydrogen bonds with THR-859 and LYS-854 from the S2 subunit (pink). The substitution of ASP by GLY at position 614 causes the loss of the two hydrogen bonds between S1 subunit and THR-859 and LYS-854 in the S2 subunit (pink).

Table 4.2: Impact of mutations on the binding affinity between spike protein RBD and ACE2, evaluated by MM-GBSA binding-free energy calculation (GBind).

Mutations	**GBind (kcal/mol)**	**G 1 (kcal/mol)**	**Effect on Spike/ACE2**
V367F	62.47	3.53	Potentially decreased binding affinity
S477N	62.69	3.31	
R408I	62.8	3.2	
V483A	63.85	2.15	
A522S	64.03	1.97	
G339D	64.08	1.92	
N354D	64.39	1.61	
K356N	64.81	1.19	
H519Q	64.84	1.16	
Wild Type	66	0	Wild-type MMGBSA value
N440K	67.88	1.88	Potentially increased binding affinity
N450K	67.88	1.88	
D364Y	68.24	2.24	
S477R	69.86	3.86	

3.4.5 Clustering and Divergence of SARS-CoV-2 Genomes

To compare the mutational profile similarity between the 79 countries, we used the Jaccard distance as a suitable metric for clustering, allowing the overall similarity measure, ranging from 0 (identical) to 1 (no overlap). We first calculated the mutational frequency in each country individually, then the Jaccard method was used to measure the distances between countries (see Materials and Methods). Figure S1 shows the pairwise similarity between countries, scaled from 0 (light yellow) to 1 (red). The clustering between the 79 countries showed two main clusters, each subdivided into several sub-clusters (SCs), and these two clusters included countries of the six continents. In addition, the countries in cluster 2 were closer to each other than those in cluster 1,demonstrating high genetic similarities between strains from these countries. We also observed fifteen SCs with a

Table 4.3: Jaccard distance between countries based on their mutational frequencies. Only the distance less than 0.5 (>50% overlap) between countries is displayed.

Cluster	Sub-Cluster	Countries	Jaccard Distance	Geographic Areas
Cluster 2	SC-3	Nigeria, Serbia, Croatia, Ireland, Peru	0.26	Africa, Europe, South America
Cluster 2	SC-4	Vietnam, Jordan	0.27	Asia
Cluster 2	SC-5	Sri Lanka, Kuwait	0.30	Asia
Cluster 2	SC-6	Greece, Portugal	0.32	Europe
Cluster 2	SC-7	Singapore, Thailand	0.35	Asia
Cluster 2	SC-8	Finland, Poland	0.35	Europe
Cluster 2	SC-9	Slovenia, Jamaica	0.35	Europe, North America
Cluster 2	SC-10	Denmark, Iceland	0.36	Europe
Cluster 2	SC-11	Germany, Russia	0.36	Europe
Cluster 1	SC-12	Hungary, Latvia	0.41	Europe
Cluster 1	SC-13	Chile, Brazil	0.43	South America
Cluster 1	SC-14	Iran, Pakistan	0.43	Asia
Cluster 2	SC-15	Netherland, Belgium, Austria	0.49	Europe

Meanwhile,the intraspecific divergence of SARS-CoV-2 was also assessed in the genomes of eachcountry compared to the genome reference Wuhan-Hu-1/2019. As shown in **(Figure 4.15 - A)**,the overall circulating strains in more than 50 countries seem to have a divergence per-centage of less than 0.1%, which indicates that the majority of SARS-CoV-2 genomeshave developed less than 18 mutations in them. The highest percentage of divergencein Asia, Europe, North America, South America, Oceania, and Africa was observed inHong Kong (0.45%), Serbia (0.42%) Mexico (0.07%), Colombia (0.05%), Guam (0.05%),and Gambia (0.43%), respectively. While the lowest percentage was shown in Portugal(0.01), Canada (0.05%), Bangladesh (0.02%), Peru (0.01%), New Zealand (0.02%), and DRC (0.03%). Moreover, the phylogenetic divergence tree **(Figure 4.15 - B)** shows that the highest rate was among genomes from Asia, followed by Europe, and North America.

In Asia, most strains showed a divergence of 0.0221 to 0.0231. Likewise, European strains clustered between 0.0223 and 0.0231, while North American strains had a divergence of0.0221 to 0.0230. Using the Nextstrain clade nomenclature, we can identify two main clades with different divergence profiles; first and most divergent “A2” clade, although the first strain observed was from China. This clade mainly contained genomes from Europe. The second “B1” clade appeared to be less divergent and to a large extent included Asian strains. Nevertheless, the genomes of Africa, North America, and South America were scattered across the phylogenetic divergence tree without a specific coating. The rate of divergence also varied within clades: A2 included three subclades, the sub-c2 harboring the Q57H mutation, with a divergence of 0.0224 to 0.0229, and mainly included strains from North America and Asia. The sub-c3 containing mostly European genomes shared the R203K mutation: In this subclade, a low rate of divergence was observed

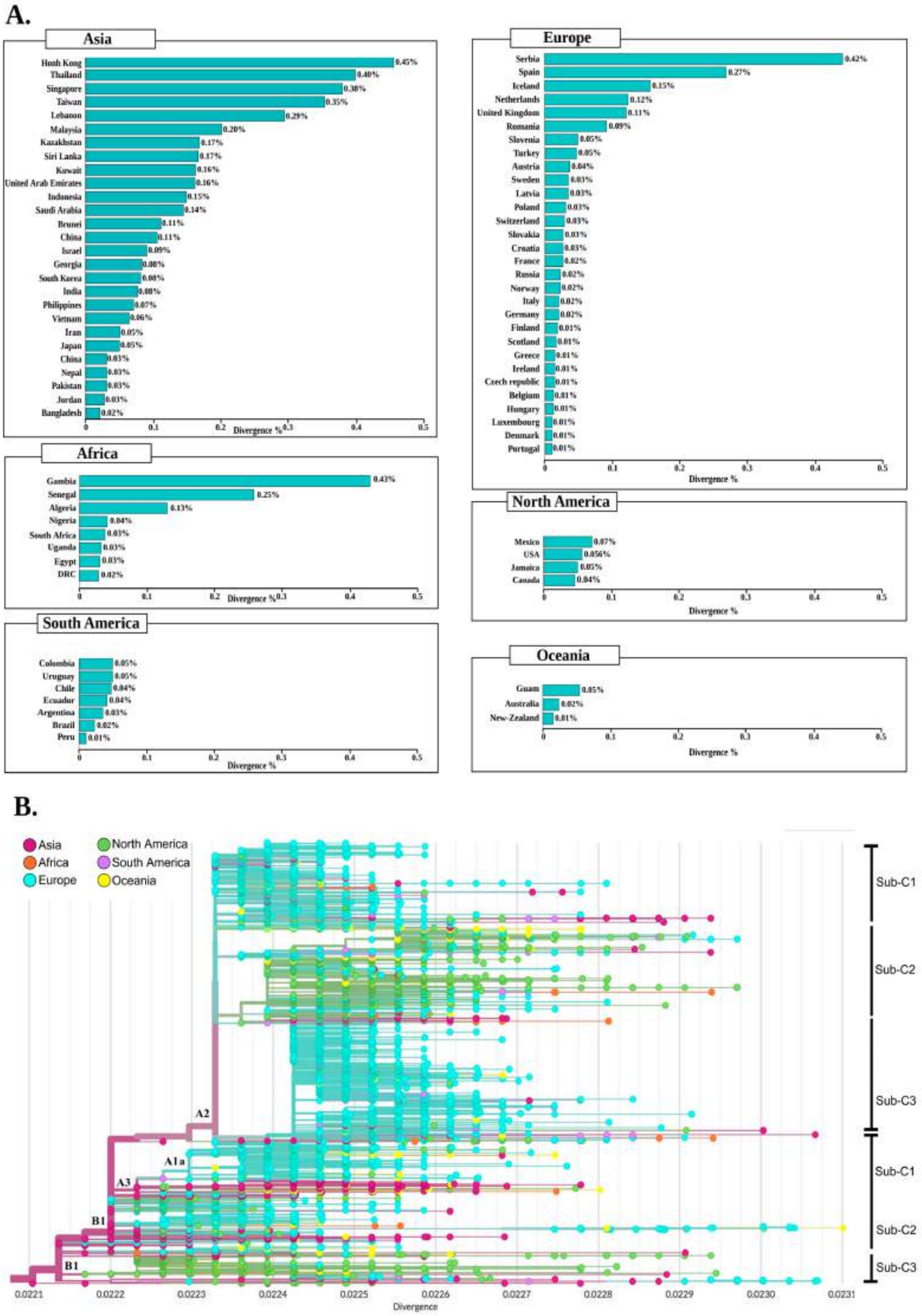

Figure 4.15: Divergence of SARS-CoV-2 genomes from different geographic areas compared with the genome reference Wuhan-Hu-1/2019.
(A) The bar graph illustrating the divergence (measured in percentage) of the SARS-CoV-2 genomes of each country compared to the reference genome Wuhan-Hu-1/2019. The divergence calculation method is detailed in the Materials and Methods section. **(B)** The phylogenetic divergence tree of the 30,983 SARS-CoV-2 genomes grouped into six geographic regions. The length of the branches shows the divergence and the color codes indicate the six geographical areas.

in all continents except Africa, while the greatest divergence was a strain from Taiwan (Asia) (¿0.0223). On the other hand, clade 2 (B1) harbored mainly genomes from North America and Asia, while the high divergence in this clade observed in Eu-rope (France) with 0.0231. The sub-c2 and sub-c3 of this clade appeared to be the most diverse with the lowest divergence in the United Kingdom and the highest in Australia.

3.4.6 Phylogenetics and Spatio Dynamics of SARS-CoV-2

The topology of the maximum likelihood phylogenetic tree **(Figure 4.16 - A)** shows a clear clustering: one cluster containing mainly Asian strains, while the second containing European strains with a specific clade sharing the D614G mutation. For each cluster, we identified different clades: cluster 1 containing two main clades A1a and B1 harboring

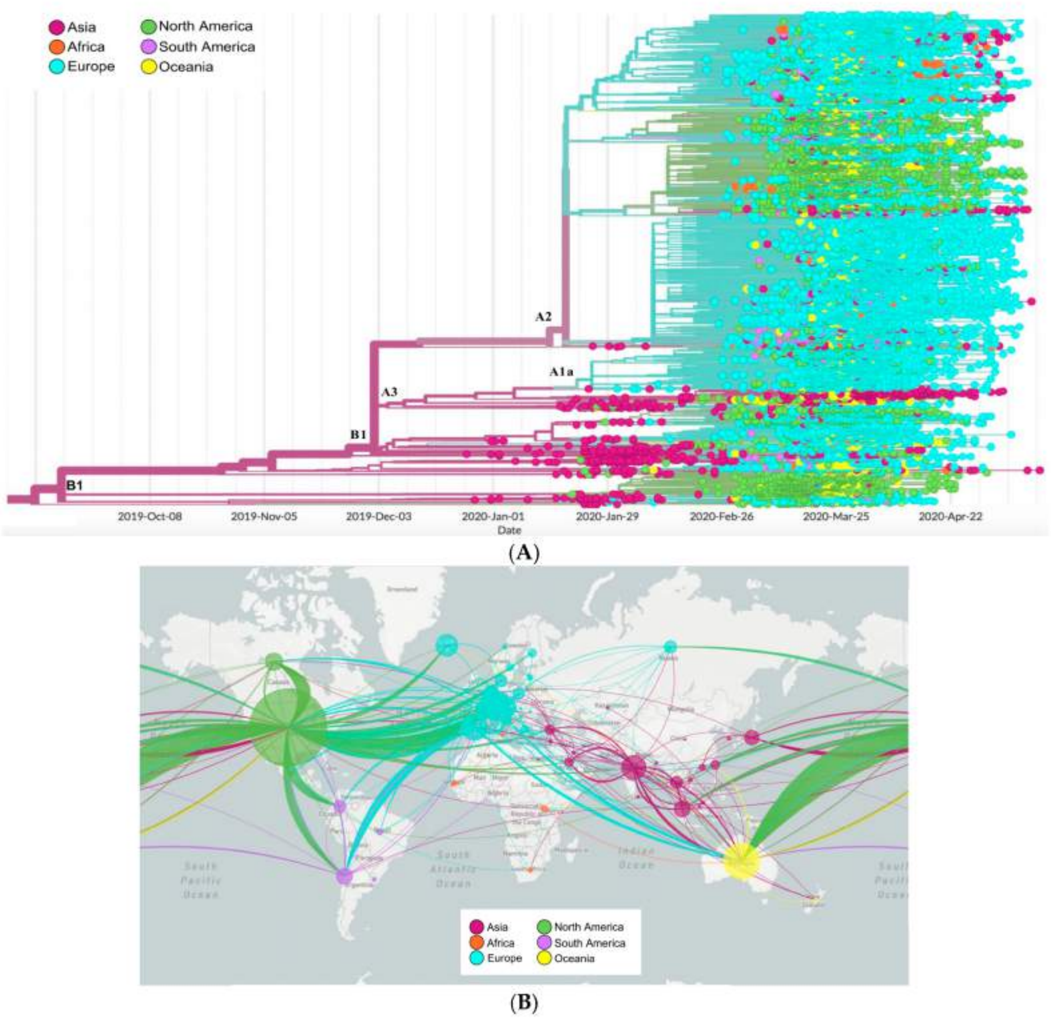

Figure 4.16: Phylogenetic tree and spatial dynamics of SARS-CoV-2.
(A) Phylogenetic analysis of 30,983 SARS-CoV-2 genomes grouped into six geographic areas. The length of the branches represents the distance in time. **(B)** Phylodynamic analysis representing the propagation and evolution of 30,983 SARS-CoV-2 genomes in different geographic areas. The color codes represent the six geographic areas.

mainly strains from Asia, North America, and Asia, Europe, respectively. However, cluster 2 harbored three clades: B2, A2, A2a without a specific pattern. The distribution of African genomes across the phylogenetic tree showed a close relationship with different continents. In the first clade, African genomes (mainly from South West Africa) clustered with Asia and showed the lowest divergence rate. Meanwhile, genomes clustering in the European clade shared the three-pattern mutations mostly common in Europe:G28881A, G28882A, and G28883C.

The map **(Figure 4.16 - B)** shows the spatio-dynamics of SARS-CoV-2 and provides an insight into the viral strain's potential geographical origin based on the sample used and displays a complex and interconnected network of strains.From these samples, strains from Asia appear to have diverged and resulted in other strains in all the investigated regions. European strains seem to have given rise to those in North America, South America, and Africa, with multiple divergent strains within Europe itself. Similarly, with less frequency, strains from South and North America appear to be related to some divergent strains in Europe and Asia.

3.5 Discussion

Due to the rapid spread and mortality rate of the new SARS-CoV-2 pandemic, the development of an effective vaccine against this virus is of a high priority [25]. The availability of the first viral sequence derived during the COVID-19 epidemic, Wuhan-Hu-1, was published on 5 January 2020. From this date, numerous vaccination programs were launched [25][26]. Furthermore, drugs and vaccines should target relatively invariant and highly constrained regions of the SARS-CoV-2 genomes, to avoid drug resistance and vaccine escape [27]. For this, monitoring genomic changes in the virus are essential and play a pivotal role in all of the above efforts, due to the appearance of genetic variants, which could affect the efficacy of vaccines. In this study, we investigated the genetic diversity in 30,983 complete SARS-CoV-2 genomes isolated from 79 countries belonging to the six continents, while evaluating the possibility of developing an effective universal vaccine.

Our results showed three different situations of the identified mutations: (i) The mutations that have developed and are gaining a predominance in the six geographic areas; (ii) mutations which were predominant only in certain geographic regions; and(iii) mutations apparently expanding, but low in frequency in all isolates studied. From this third situation, it is interesting to note that a low rate of recurrent mutations was found across genomes, with only 5.27% of the total mutations have a frequency greater than 0.01, while 94.73% had a frequency of less than 0.01, of which 49.68% were single mutations (specific to a genome). In line with previous reports, our results show strong evidence that, so far, the evolution of SARS-CoV-2 has evolved in a non-deterministic process and that this diversification has mainly been due to random genetic drift which plays a dominant role in the spread of low-frequency mutations [28][29][30][31][32], suggesting that there was no strong selective pressure exerted on SARS-CoV-2 by the human population. Although the hotspot mutations are motivated by positive selection, which could indicate that the substitution of a specific amino acid offers an adaptive advantage under particular conditions [16]. Our study showed that more than half of the hotspot mutations

identified in the SARS-Vo2 genomes gained their predominance outside of Asia; including the hotspot mutation Q57H (in ORF3a), until early March, which had not yet been observed among isolates from China, while it emerged before that date in Europe and also spread in isolates from North America. Likewise, seven other hotspot mutations with high frequency in different geographic areas (except Asia); including the double mutations R203K-G204R and double mutations of the N protein (in Europe,South America, and Oceania), M5865V of nsp13-helicase, and D5932T of nsp14-ExoN (in North America), and T5020I of nsp12-RdRp (in Africa). Hotspot mutations, due to their increased frequency in different geographic areas, are considered an important criterion for defining and characterizing emerging clades [33][14].
As a whole, a low rate of intragenomic divergence of SARS-CoV-2 (¡0.5%) was found between all the countries studied. Compared to different geographic areas, the high rate of divergence in Asian countries could be due to multiple sources of infection with different strains at the start of the epidemic. This could suggest an early introduction and rapid spread of genetically close variants to the original strain in continents with high infection rates, such as Eu-rope and North America, which founded the virus's first transmission networks [34][35].Rapid transmission means a single source leading to multiple infections, thus giving the virus fewer life cycles to change: This is consistent with a previous study describing a continued tendency of the virus to diverge over time [36]. Furthermore, the dynamics of transmission showed that the least divergent African variants were grouped with Asian strains, while the most divergent were grouped with Europe and North America. This distribution points to different sources of infection [37][38]. South America's genomes appear to originate from North America and Europe, showing a close clustering with Europe in low and high divergence strains, which is concordant with a previous study[39]. In contrast, certain strains from Oceania allow poor monitoring of the origin of the infection, but show a close relationship with the genomes of Europe. Overall, the North American and European genomes appear to be responsible for most of the spread of the disease. Besides the divergence, the intragenomic clustering between the 79 countries did not have a clear pattern regarding their geographic distributions, reflecting the effect of migration and globalization as previously reported [40][41].

As the virus spreads more widely around the world, it is important to monitor and assess mutations that could be of potential concern as an early warning system to consider as vaccine studies progress.The S protein is a major target for vaccines and therapeutics, due to its key role in mediating virus entry and its immunogenicity trait [8][42]. Analysis of protein S revealed a high-frequency mutation (D614G) with a continuous trend over time in different geographic areas. This mutation is proximal to the S1 cleavage domain at position 614which involved the change of a large amino acid residue (aspartic acid) to a small hydrophobic residue (glycine) and became widely dominant worldwide within a few months[43][44]. Our results showed that this mutation induces a loss of two hydrogen bonds between the S1 and S2 subunits of neighboring protomers and can, therefore, increase the rate of cleavage of these subunits in the pre-fusion state of spike protein to allow its conformational transition to the post-fusion state associated with membrane fusion upon virus entry [45][46]. Indeed, our structural modeling of this mutation has shown no substantial impact on the secondary or tertiary structure of the spike protein. Therefore, it is unlikely that G614D could affect the immunogenicity of RBD epitopes considered important in neutralizing antibodies [29,36]. Likewise, previous studies [40][41][47] re-ported that the antibodies generated from natural infection with mutant type D614 or G614 could carry out a neutralization cross, indicating that the locus is not critical for antibody-mediated immunity, so the D614G mutation is unlikely to have a major impact on the efficacy of vaccines in development, some of which exclusively target RBD region. To this end, the RBD of the spike protein allows the virus to bind to the ACE2host receptor [48][49]. Mutations in this receptor are a likely pathway to evade antibody recognition, such as described in other viruses [50][51]. In all the genomes analyzed, 36non-synonymous RBD mutations were identified and all these mutations had a low frequency (¡0.01) in the genomes of six continents, which is consistent with several studies that have found that mutations are extremely rare in the RBD region [32][40][52][53].The calculated binding-free energy of mutant RBDs of spike protein complexed with human ACE2 revealed only four RBD mutant types (D364Y, N440K, N450K, S477R)displaying a much lower binding-free energy (G), indicating a significantly higher affinity for the ACE2, which could influence the pathogenicity of SARS-CoV-2. Of these four mutations, Ou et al. [54] previously reported that D364Y potentially enhancesthe binding of viral spike protein to ACE2, possibly due to the improved structural stabilization of the RBD beta-sheet scaffold.

Effective COVID-19 vaccines will be a permanent solution to viral infections, and it is likely that more than one strategy will be successful to this end [55]. RNA interference-based therapy (RNAi) could be an alternative in the fight against SARS-CoV-2 [56], where small interfering RNAs (siRNA,20 to 25 nt in length) could affect the region highly conserved from SARS-CoV-2 and could also act as an inhibitor to suppress genetic disorders in the lungs [57]. The efficiency of siRNA to inhibit gene expression and replication by targeting the leader and spike coding sequence of SARS-CoV has already been demonstrated [58][59][60]. Alny-lam Pharmaceuticals (USA) has designed and synthesized over 350 siRNAs targeting highly conserved regions of circulating SARS-CoV-2 genomes [61]. The main siRNA candidates will be evaluated for their antiviral activity in vitro and in vivo, leading to the selection of a candidate for development. It is interesting to note that the effects of siRNAs can be influenced by mutations. Chen et al. 2020 [62] reported nine poten-tial target siRNA sequences in the SARS-CoV-2 genome. To this end, we analyzed the mutations present in these target sequences in the 30,983 genomes of our study. One to seven SNPs in each of the nine target sequences were found (Supplementary Table S6),hence the importance of monitoring the introduction of any mutations that could com-promise the potential efficacy of siRNAs and candidate vaccines.
SARS-CoV-2 has only recently been discovered in the human population; adaptive processes could take years to occur. Although we cannot predict whether adaptive selection will be observed in this virus in the future, we can conclude that the currently circulating strains constitute a homogeneous viral population. We can therefore be cautiously optimistic that, so far,the genetic diversity of SARS-CoV-2 should not be an obstacle to the development of a universal vaccine candidate. Ongoing surveillance of SARS-CoV-2 genomic changes will be essential to monitor and understand host–pathogen interactions

that may contribute to the development of effective vaccines and therapeutics.

3.6 References

1 Ke Wang, Wei Chen, Yu-Sen Zhou, Jian-Qi Lian, Zheng Zhang, Peng Du, Li Gong,Yang Zhang, Hong-Yong Cui, Jie-Jie Geng, et al. Sars-cov-2 invades host cells viaa novel route: Cd147-spike protein.BioRxiv, 2020.

2 Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song,Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirusfrom patients with pneumonia in china, 2019.New England journal of medicine,2020.

3 Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, AhmedAl-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declaresglobal emergency: A review of the 2019 novel coronavirus (covid-19).Internationaljournal of surgery, 76:71–76, 2020.

4 Jos e M Cuevas, Ron Geller, Raquel Garijo, Jos e L opez-Aldeguer, and RafaelSanju an. Extremely high mutation rate of hiv-1 in vivo.PLoS Biol, 13(9):e1002251,2015.

5 Barry T Rouse and Sharvan Sehrawat. Immunity and immunopathology to viruses:what decides the outcome?Nature Reviews Immunology, 10(7):514–526, 2010.

6 Aiping Wu, Yousong Peng, Baoying Huang, Xiao Ding, Xianyue Wang, Peihua Niu,Jing Meng, Zhaozhong Zhu, Zheng Zhang, Jiangyuan Wang, et al. Genome com-position and divergence of the novel coronavirus (2019-ncov) originating in china.Cell host microbe, 27(3):325–328, 2020.

7 Yasmin A Malik. Properties of coronavirus and sars-cov-2.The Malaysian journalof pathology, 42(1):3–11, 2020.

8 Lanying Du, Yuxian He, Yusen Zhou, Shuwen Liu, Bo-Jian Zheng, and Shibo Jiang.The spike protein of sars-cov—a target for vaccine and therapeutic development.Nature Reviews Microbiology, 7(3):226–236, 2009

9 Oliver C Grant, David Montgomery, Keigo Ito, and Robert J Woods. Analysis of thesars-cov-2 spike protein glycan shield reveals implications for immune recognition.Scientific reports, 10(1):1–11, 2020.

10 Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenzadata–from vision to reality.Eurosurveillance, 22(13):30494, 2017.

11 Heng Li. Minimap2: pairwise alignment for nucleotide sequences.Bioinformatics,34(18):3094–3100, 2018.

12 Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Ga-bor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/mapformat and samtools.Bioinformatics, 25(16):2078–2079, 2009.

13 Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, LuanWang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotatingand predicting the effects of single nucleotide polymorphisms, snpeff: Snps in thegenome of drosophila melanogaster strain w1118; iso-2; iso-3.Fly, 6(2):80–92, 2012.

14 Meriem Laamarti, Tarek Alouane, Souad Kartti, MW Chemao-Elfihri, MohammedHakmi, Abdelomunim Essabbar, Mohamed Laamarti, Haitam Hlali, Houda Bendani, Nassma Boumajdi, et al. Large scale genomic analysis of 3067 sars-cov-2genomes reveals a clonal geo-distribution and a rich genetic variations of hotspotsmutations.Plos one, 15(11):e0240345, 2020.

15 Kazuma Kiyotani, Yujiro Toyoshima, Kensaku Nemoto, and Yusuke Nakamura.Bioinformatic prediction of potential t cell epitopes for sars-cov-2.Journal of humangenetics, 65(7):569–575, 2020.

16 Sujay Chattopadhyay, Scott J Weissman, Vladimir N Minin, Thomas A Russo,Daniel E Dykhuizen, and Evgeni V Sokurenko. High frequency of hotspot mutationsin core genes of escherichia coli due to short-term positive selection.Proceedings ofthe National Academy of Sciences, 106(30):12412–12417, 2009.

17 Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch,Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a vi-sualization system for exploratory research and analysis.Journal of computationalchemistry, 25(13):1605–1612, 2004.

18 Gaoqi Weng, Ercheng Wang, Zhe Wang, Hui Liu, Feng Zhu, Dan Li, and TingjunHou. Hawkdock: a web server to predict and analyze the protein–protein complexbased on computational docking and mm/gbsa.Nucleic acids research, 47(W1):W322–W330, 2019.

19 Michael Levandowsky and David Winter. Distance between sets.Nature, 234(5323):34–35, 1971.

20 Esko Ukkonen. Approximate string-matching with q-grams and maximal matches.Theoretical computer science, 92(1):191–211, 1992.

21 Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.Bioinformatics, 22(13):1658–1659,2006.

22 James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Pot-ter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher.Nextstrain: real-time tracking of pathogen evolution.Bioinformatics, 34(23):4121–4123, 2018.

23 Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh.Iqtree: a fast and effective stochastic algorithm for estimating maximum-likelihoodphylogenies.Molec biology and evolution, 32(1):268–274, 2015.

24 Pavel Sagulenko, Vadim Puller, and Richard A Neher. Treetime: Maximum-likelihood phylodynamic analysis.Virus evolution, 4(1):vex042, 2018.

25 Fatima Amanat and Florian Krammer. Sars-cov-2 vaccines: status report.Immunity, 52(4):583–589, 2020.

26 Jinyong Zhang, Hao Zeng, Jiang Gu, Haibo Li, Lixin Zheng, and Quanming Zou.Progress and prospects on vaccine development against sars-cov-2.Vaccines, 8(2):153, 2020.

27 Yung-Fang Tu, Chian-Shiu Chien, Aliaksandr A Yarmishyn, Yi-Ying Lin, Yung-Hung Luo, Yi-Tsung Lin, Wei-Yi Lai, De-Ming Yang, Shih-Jie Chou, Yi-Ping Yang,et al. A review of sars-cov-2 and the ongoing clinical trials.International journalof molecular sciences, 21(7):2657, 2020.

28 Jason W Rausch, Adam A Capoferri, Mary Grace Katusiime, Sean C Patro, and-Mary F Kearney. Low genetic diversity may be an achilles heel of sars-cov-2.Proceedings of the National Academy of Sciences, 117(40):24614–24616, 2020.

29 Carol A Abidha, Joyce Nyiro, Everlyn Kamau, Osman Abdullahi, David James-Nokes, and Charles N Agoti. Transmission and evolutionary dynamics of human-coronavirus oc43 strains in coastal kenya investigated by partial spike sequenceanalysis, 2015–16.Virus Evolution, 6(1):veaa031, 2020

30 Takahiko Koyama, Dilhan Weeraratne, Jane L Snowdon, and Laxmi Parida. Emergence of drift variants that may affect covid-19 vaccine development and antibodytreatment.Pathogens, 9(5):324, 2020.

31 Alessia Lai, Annalisa Bergna, Sara Caucci, Nicola Clementi, Ilaria Vicenti, FilippoDragoni, Anna Maria Cattelan, Stefano Menzo, Angelo Pan, Annapaola Callegaro,et al. Molecular tracing of sars-cov-2 in italy in the first three months of theepidemic.Viruses, 12(8):798, 2020.

32 M Rafiul Islam, M Nazmul Hoque, M Shaminur Rahman, ASM Rubayet Ul Alam,Masuda Akther, J Akter Puspo, Salma Akter, Munawar Sultana, Keith A Crandall,and M Anwar Hossain. Genome-wide analysis of sars-cov-2 virus strains circulatingworldwide implicates heterogeneity.Scientific reports, 10(1):1–9, 2020.

33 Erik Alm, Eeva K Broberg, Thomas Connor, Emma B Hodcroft, Andrey B Komissarov, Sebastian Maurer-Stroh, Angeliki Melidou, Richard A Neher, Aine O'Toole,Dmitriy Pereyaslov, et al. Geographical and temporal distribution of sars-cov-2clades in the who european region, january to june 2020.Eurosurveillance, 25(32):2001410, 2020.

34 Paola Stefanelli, Giovanni Faggioni, Alessandra Lo Presti, Stefano Fiore, Antonella-Marchi, Eleonora Benedetti, Concetta Fabiani, Anna Anselmo, Andrea Ciammaruconi, Antonella Fortunato, et al. Whole genome and phylogenetic analysis of twosars-cov-2 strains isolated in italy in january and february 2020: additional clueson multiple introductions and further circulation in europe.Eurosurveillance, 25(13):2000305, 2020.

35 Michael Worobey, Jonathan Pekar, Brendan B Larsen, Martha I Nelson, Verity Hill,Jeffrey B Joy, Andrew Rambaut, Marc A Suchard, Joel O Wertheim, and PhilippeLemey. The emergence of sars-cov-2 in europe and north america.Science, 370(6516):564–570, 2020.

36 Javaid Ahmad Sheikh, Jasdeep Singh, Hina Singh, Salma Jamal, Mohd Khubaib,Sunil Kohli, Ulrich Dobrindt, Syed Asad Rahman, Nasreen Zafar Ehtesham, andSeyed Ehtesham Hasnain. Emerging genetic diversity among clinical isolates ofsars-cov-2: Lessons for today.Infection, Genetics and Evolution, 84:104330, 2020.

37 Leolin Katsidzira, Lenon Gwaunza, and James G Hakim. The severe acute respiratory syndrome coronavirus 2 (sars-cov-2) epidemic in zimbabwe: Quo vadis?Clinical Infectious Diseases, 71(16):2180–2183, 2020.

38 Marguerite Massinga Loemb e, Akhona Tshangela, Stephanie J Salyer, Jay KVarma, Ahmed E Ogwell Ouma, and John N Nkengasong. Covid-19 in africa:the spread and response.Nature Medicine, 26(7):999–1003, 2020.

39 Julio A Poterico and Orson Mestanza. Genetic variants and source of introductionof sars-cov-2 in south america.Journal of medical virology, 92(10):2139–2145, 2020.

40 Bette Korber, Will Fischer, S Gnana Gnanakaran, Heyjin Yoon, James Theiler,Werner Abfalterer, Brian Foley, Elena E Giorgi, Tanmoy Bhattacharya, Matthew DParker, et al. Spike mutation pipeline reveals the emergence of a more transmissibleform of sars-cov-2.BioRxiv, 2020.

41 Jie Hu, Chang Long He, Qingzhu Gao, Gui Ji Zhang, Xiao Xia Cao, Quan Xin Long,Hai Jun Deng, Lu Yi Huang, Juan Chen, Kai Wang, et al. The d614g mutation ofsars-cov-2 spike protein enhances viral infectivity.BioRxiv, 2020.

42 T Thanh Le, Zacharias Andreadakis, Arun Kumar, R G omez Rom an, Stig Tollefsen, Melanie Saville, Stephen Mayhew, et al. The covid-19 vaccine developmentlandscape.Nat Rev Drug Discov, 19(5):305–306, 2020.

43 Sandra Isabel, Luc ıa Gra na-Miraglia, Jahir M Gutierrez, Cedoljub Bundalovic-Torma, Helen E Groves, Marc R Isabel, AliReza Eshaghi, Samir N Patel,Jonathan B Gubbay, Tomi Poutanen, et al. Evolutionary and structural analysesof sars-cov-2 d614g spike protein mutation now documented worldwide.Scientificreports, 10(1):1–9, 2020.

44 Xianding Deng, Wei Gu, Scot Federman, Louis Du Plessis, Oliver G Pybus, Nuno RFaria, Candace Wang, Guixia Yu, Brian Bushnell, Chao-Yang Pan, et al. Genomicsurveillance reveals multiple introductions of sars-cov-2 into northern california.Science, 369(6503):582–587, 2020.

45 Leyan Tang, Allison Schulkins, Chun-Nan Chen, Kurt Deshayes, and John S Kenney. The sars-cov-2 spike protein d614g mutation shows increasing dominance andmay confer a structural advantage to the furin cleavage domain. 2020.

46 Xiaoli Xiong, Kun Qu, Katarzyna A Ciazynska, Myra Hosmillo, Andrew P Carter,Soraya Ebrahimi, Zunlong Ke, Sjors HW Scheres, Laura Bergamaschi, Guinevere LGrice, et al. A thermostable, closed sars-cov-2 spike protein trimer.Nature Struc-tural Molecular Biology, 27(10):934–941, 2020.

47 Nathan D Grubaugh, William P Hanage, and Angela L Rasmussen. Making senseof mutation: what d614g means for the covid-19 pandemic remains unclear.Cell,182(4):794–795, 2020.

48 Wanbo Tai, Lei He, Xiujuan Zhang, Jing Pu, Denis Voronin, Shibo Jiang, YusenZhou, and Lanying Du. Characterization of the receptor-binding domain (rbd)of 2019 novel coronavirus: implication for development of rbd protein as a viralattachment inhibitor and vaccine.Cellular molecular immunology, 17(6):613–620, 2020.

49 Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, QibinGeng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition bysars-cov-2.Nature, 581(7807):221–224, 2020.

50 Alan HM Wong, Aidan CA Tomlinson, Dongxia Zhou, Malathy Satkunarajah,Kevin Chen, Chetna Sharon, Marc Desforges, Pierre J Talbot, and James M Rini.Receptor-binding loops in alphacoronavirus adaptation and evolution.Nature com-munications, 8(1):1–10, 2017.

51 Barry Rockx, Eric Donaldson, Matthew Frieman, Timothy Sheahan, Davide Corti,Antonio Lanzavecchia, and Ralph S Baric. Escape from human monoclonal anti-body neutralization affects in vitro and in vivo fitness of severe acute respiratorysyndrome coronavirus.The Journal of infectious diseases, 201(6):946–955, 2010.

52 Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthenedsars-cov-2 infectivity.Journal of molecular biology, 432(19):5212–5226, 2020.

53 Xiaojun Li, Elena E Giorgi, Manukumar Honnayakanahalli Marichannegowda,Brian Foley, Chuan Xiao, Xiang-Peng Kong, Yue Chen, S Gnanakaran, Bette Ko-rber, and Feng Gao. Emergence of sars-cov-2 through recombination and strongpurifying selection.Science Advances, 6(27):eabb9153, 2020.

54 Junxian Ou, Zhonghua Zhou, Ruixue Dai, Jing Zhang, Wendong Lan, Shan Zhao,Jianguo Wu, Donald Seto, Lilian Cui, Gong Zhang, et al. Emergence of rbd muta-tions in circulating sars-cov-2 strains enhancing the structural stability and humanace2 receptor affinity of the spike protein.BioRxiv, 2020.

55 Hasan Uluda g, Kylie Parent, Hamidreza Montazeri Aliabadi, and Azita Haddadi.Prospects for rnai therapy of covid-19.Frontiers in bioengineering and biotechnology, 8:916, 2020.

56 Cynthia Liu, Qiongqiong Zhou, Yingzhu Li, Linda V Garner, Steve P Watkins,Linda J Carter, Jeffrey Smoot, Anne C Gregg, Angela D Daniels, Susan Jervey,et al. Research and development on therapeutic agents and vaccines for covid-19and related human coronavirus diseases, 2020.

57 Sanhita Ghosh, Sayeed Mohammad Firdous, and Anirban Nath. sirna could be apotential therapy for covid-19.EXCLI journal, 19:528, 2020.

58 SHI Yi, YANG De Hua, Jie XIONG, JIA Jie, Bing HUANG, and You Xin JIN.Inhibition of genes expression of sars coronavirus by synthetic small interfering rnas.Cell research, 15(3):193–200, 2005.[59] T Li, Y Zhang, L Fu, C Yu, X Li, Y Li, X Zhang, Z Rong, Y Wang, H Ning,et al. sirna targeting the leader sequence of sars-cov inhibits virus replication.Gene therapy, 12(9):751–761, 2005.

60 Chang-Jer Wu, Hui-Wen Huang, Chiu-Yi Liu, Cheng-Fong Hong, and Yi-Lin Chan.Inhibition of sars-cov replication by sirna.Antiviral research, 65(1):45–48, 2005.

61 John Hodgson. The pandemic pipeline.Nature biotechnology, 38(5):523–532, 2020.

62 Wei Chen, Pengmian Feng, Kewei Liu, Meng Wu, and Hao Lin. Computationalidentification of small interfering rna targets in sars-cov-2.Virologica Sinica, 35(3):359–361, 2020.

4 Discussion

The ongoing rapid transmission and global spread of SARS-CoV-2 have raised critical questions about the evolution and adaptation of the viral population driven by mutations, deletions and/or recombination as it spreads across the world, encountering diverse host immune systems and various counter-measures6. The results showed 782 variants sites, of which 512 (65.47%) had a non-synonymous effect. The genetic differences among SARS-CoV-2 strains sampled from diverse locations could therefore be linked with their geographical distributions.

This rapidly evolving virus is capable of adapting swiftly to diverse environments. Moreover, identifying the conformational changes in mutated protein structures and untranslated cisacting elements is of significance for studying the virulence, pathogenicity and transmissibility of SARS-CoV-2.

In line with previous reports, our results show strong evidence that, so far, the evolution of SARS-CoV-2 has evolved in a non-deterministic process and that this diversification has mainly been due to random genetic drift which plays a dominant role in the spread of low-frequency mutations, suggesting that there was no strong selective pressure exerted on SARS-CoV-2 by the human population. Although we cannot predict whether adaptive selection will be observed in this virus in the future, we can conclude that the currently circulating strains constitute a homogeneous viral population.

We can therefore be cautiously optimistic that, so far, the genetic diversity of SARS-CoV-2 should not be an obstacle to the development of a universal vaccine candidate. Ongoing surveillance of SARS-CoV-2 genomic changes will be essential to monitor and understand host-pathogen interactions that may contribute to the development of effective vaccines and therapeutics.

Chapter 5

General discussion

Respiratory tract infections (RTIs) is the top cause of morbidity and mortality from infectious diseases worldwide [109]. Until the end of December 2019, only three pathogens were featured on the WHO Blueprint priority list for research and development: severe acute respiratory syndrome (SARS), coronavirus (SARS-CoV), Middle East Respiratory Syndrome (MERS), Coronavirus (MERS-CoV), and *Mycobacterium tuberculosis* [110]. In January 2020, SARS-CoV-2, the cause of COVID-19, was added to the priority list. Thus, the work carried out during this thesis allowed us to study several aspects of respiratory tract infectious diseases.

The objective of this thesis is to describe the diversity, resistance and transmission mechanism of *Mycobacterium tuberculosis* and SARS-CoV-2. This work highlights the potential of using genomics tools in the control of outbreaks by providing answers to the following questions:

1 Where is the infectious agent coming from?

For the development of optimal strategies for infectious diseases control and infection prevention, it is of critical importance to have an accurate tracing of in-country and cross-border transmission and a rapid identification of emerging mutants' clones. In the case of TB, genotyping is the primary step in its tracing. Generally based on the classical method (MIRU and VNTR), it is mainly used for MDR and XDR-tb surveillance and transmission as shown in different studies in Morocco[111][112][113]. However, the intrinsic lack of discriminatory power of MIRU and VNTR makes these technologies suboptimal for supporting contact tracing. This has been shown by a recent study investigating MDR-TB outbreak among African refugees, which found MIRU-VNTR to be poor at tracking the outbreak compared to WGS [114].

Thus, in this study, we used the WGS for tracing the origin of 2 infectious agents : SARS-CoV-2 and MTB. This is the first study on the transmission and the diversity of TB strains in Morocco using WGS. The analysis of the nine strains showed the domination of lineage 4, mainly LAM9 and H4, which is in agreement with findings from other studies [111][115]. The strains were predicted to have a European origin, which could be due to the French or Spanish colonization of Morocco. In fact, the same pattern was found a few years ago in countries with a similar colonization history like Tunisia and Algeria. In the case of SARS-CoV-2 [116][117], a strong pathogen trading pattern was observed between Morocco and European countries during COVID-19 pandemic. SNP analysis showed that most strains in Morocco or Africa in general originated from countries in Europe.

2 How easy is it for an emerging pathogen to evolve Adaptively, and how fast can it do so?

The answer is that it will depend on the likelihood and speed of acquiring mutations. The continuous and rapid evolution of pathogens is responsive. As we tighten our mitigation, we induce a selective pressure, either by prophylactic steps or pharmaceutical treatments which complicates the disease control.

Mutations are simply errors in replication, introduced during the duplication of the genetic material. Since the majority of mutations are harmful to the pathogen, they are generally easily eliminated by natural selection [118]. However, Pathogens with a high mutation rate, a small genome size (where mutations have significant functional consequences), and a large population have the greatest adaptive ability, such as RNA viruses [119].

In this study, we have considered ways of how SARS-CoV-2 and TB change genetically in response to selection pressure using different pipelines and sequencing technologies. Generally, TB evolves in response to therapeutic treatments as selection pressure; the use, overuse and misuse of antimicrobial drugs for TB treatment induce antibacterial resistance.

In this study, TB sequencing generated reads length between 32 – 300 bp with Illumina Miseq, which allows high accuracy mutation detection suitable for TB resistance analysis. Better profiling of resistance allows better treatment and, therefore, a better outcome. In our analysis, a resistance-associated mutation was identified in the nine strains. Four strains were classified as MDR strains showing resistance to isoniazid and rifampicin associated with hot spot mutations (in rpob and katg/inhA) detectable with Haine or gen expert. Other Known resistance-associated mutations were also identified using WGS, such as mutations responsible for pyrazinamide or amikacin resistance.

However, with the Illumina range of reads length, it is challenging to extract repetitive region such as PE/PEE/PERS genes or insertion sequence, which could be used to study the diversity of TB [120]. Generally, The PPE genes demonstrate high SNP density than the rest of the genomes, as shown in this study.

The overall diversity analysis highlighted that sensitive strains are less diverse than the resistant ones(Lam9 and H4), suggesting a fitness mechanism displayed by those strains to adapt. Longer read sequencing technology may overcome this issue, thus Nanopore Minion with priming walking technique was used for SARS-CoV-2 strains and resulted in a 99% in both precision and sensitivity and showed a homogenous diversity of Moroccan strains. Despite the nanopore sequencing technology's cost and time effectiveness, it performs poorly in indels detection; an observation that was reported previously [121].

wherase in TB the therapeutic treatment is what pushes the pathogen to higher drugpersistence through natural selection, the enforcement of social distancing in SARS-CoV-2 pushed the reproduction rate below the pandemic threshold; the pathogen becomes naturally pressured towards higher transmissibility. In most cases, natural selection determines the destiny of a newly occurring mutation. The diversity for SARS-CoV-2 world wild showed a substitution rate of 8.101 × 104 substitution/site/year, which was in the same order of magnitude as SARS-CoV-2 and MERS-CoV [122][123]. This diversity helped the virus develop several key specific sites and haplotypes, increasing its infectivity or pathogenicity.

As described in chapter 3, the beginning of the spike mutation D614G emerged independently and simultaneously in March. It swept across all three continents and was suggested to be a result of natural selection. However, the analysis of 3067 SARS-CoV-2 genomes demonstrated the appearance of the D614G mutation in different parts of China since late January. These findings raised the possibility that this mutation could be due to chance founder events where viruses carrying D614G just happened to start the majority of early transmission events in different locations.

As previously shown in studies on human respiratory cells and on animal models, viruses harboring the D614G mutation increased infectivity and transmission by improving the interaction with the host cell receptor, resulting in higher levels of the virus in the host and an increased rate of transmission [124].

Since our study (chapter 5), new variants have emerged as a consequence of selective sweeps, a very common phenomenon in pathogens. In late September 2020, a new variant has emerged in the UK (Lineage B.1.1.7),,, which had accumulated 17 new mutations before its detection (14 nonsynonymous mutations, six synonymous, and three deletions) [125], one of which is in the spike protein receptor-binding domain (N501Y). Unlike D614G mutation, which could plausibly have benefited from early chance events, Lineage B.1.1.7 expanded when SARS-CoV-2 cases were widespread and have seemingly achieved dominance by outcompeting an existing population of circulating variants [126], which suggests a significant amount of prior evolution, possibly in a chronically infected host. The host genetics majorly influences selective sweeps, as was the case in South Africa with the emergence of new variant 501Y.V2 in the spike protein [8]. Preliminary studies suggest the variant is associated with a higher viral load , which is similar to the UK stains are fast transmitted. The hypothesis has suggested that the high rate of TB cases and HIV history in South Africa has influenced the selection pressure imposed by the host cell on the virus, making it more unstable.

Mutants arising through antigenic drift may result in a novel strain capable of infecting other organisms. That was the case of the Y453F mutation that started in summer 2020 in the Netherlands, and which increased the binding affinity to ACE2 in mink [8] and enhanced human to mink, mink to mink, and mink to human transmission.

3 What's the impact of mutations on the infectious Agent?

Some SNPs, as previously showed, were associated with increasing severity, Transmissibility or resistance [127][128][129]. It has been suggested that some of these variants may have improved the interaction with the host cell receptor, enabled the infectious agent to propagate better, to escape the immune system, or evade therapy. This would result in higher levels of the agent in the host and an increased rate of transmission.

This study observed phenotypic ofloxacin resistance among MTBC clinical strains lacking the common mutations gyrA D94G and gyrB A90V gyrB N538D and E540V. Early detection of less common FLQ resistance mutations could allow patients to be placed on appropriate therapy early and decrease the chances of treatment failure and acquired drug resistance. The analysis of newly identified mutations using docking suggested that the Pro439Ala mutation in gyrB decreases the affinity to ofloxacin and levofloxacin second-line treatment drugs resulting in medium-resistant as confirmed with a phenotypic test.

In the case of COVID-19, we identified four mutations in the spike protein (N440K N450K D364Y S477R) with higher binding affinity compared to the wild type, potentially enhancing the binding of viral spike protein to ACE2. Impact of COVID-19 on TB It is anticipated that strict lockout procedures will trigger a large rise in TB cases and deaths [130]. Steps in the elimination of tuberculosis may be derailed for at least five years [131]. Lower and middle-income countries are the most affected by tuberculosis, whereas high-income countries have a low TB rate [22]. Europe is the second epicentre of COVID-19, which may help understand why COVID-19 has mobilized more global capital and human resources in a year than tuberculosis has in decades[132].

4 Superposing of two pandamics

Poverty is known to be a major risk factor in developing TB. A recent prediction of the world bank estimates that more than 88 million people will fall under the extreme poverty line by the end of 2021, which will have a long-term effect on the TB burden [133]. Transcriptomic analysis has recently taken place to determine if the severity of COVID-19 was related to the symptomatic and asymptomatic spectrum of tb [6]. Results showed an increase in the severity of Covi-19 in patients with active TB (ATB). They suggested that this may be due to the abundance of circulating myeloid subpopulations. In both diseases, the development of IFN-gamma and the response signature of IFN I and IFN III increased significantly [134]. Also, the use of immunosuppressive drugs in critically-ill COVID-19 patients increases the risk of infection or tuberculosis' reactivation in the long term, which can pose the greatest threat to ending the TB epidemic [135].

The occurrence of death due to COVID-19 was found to be 21-fold lower in countries with a national BCG vaccination policy than countries without such a policy [136]. In many correlation studies, BCG vaccination was associated with low incidence rates of COVID-19 and was suggested as an immune stimulant in countries with BCG routine. While studies have indicated that the BCG effect lasts only ten years [137]. Furthermore, in order to boost the central T cell memory population that can last for decades, another shot of the BCG vaccine is required at least three months after the first one [138]. According to a modelling analysis by STOP-TB partnership11 in collaboration with John Hopkins University and USAID, a 3-month lockdown and a long 10-month restoration of services [131] will lead to a 6.3 million additional cases of TB and an additional 1.4 million TB deaths, bringing the total to 1.66 (1.3 – 2.1) million TB deaths in 2021, near the global level of TB mortality of the year 2015 [131].

5 Impact of COVID-19 on TB treatment and diagnostic services:

TB diagnosis and treatment facilities were temporarily designated for COVID-19 diagnosis and treatment. Meanwhile, patients with suspected or reported tuberculosis were unable to visit the hospital for fear of contracting COVID-19. This can postpone diagnosis for suspected tuberculosis cases and raise the risk of TB infection in households.

Overwhelming health care systems by COVID-19 cases is likely to impact TB treatment and diagnostic services in several ways:

- Diversion of resources, both human and financial, to deal with the outbreak;
- Most news outlets and national attention has been focused on pandemic management without taking into consideration the TB programs;
- Exhausting health care personnel, which can result in an increase in diagnosis errors, risk of illeness or even death.
- Discouraging people from visiting TB services because of COVID-19 spread at health care facilities,
- Late diagnosis and inappropriate treatment of TB, that may increase the risk of developing drug-resistance.

6 Impact of COVID-19 on the prevention and control of TB:

Among the main prevention measures of COVID-19 to reduce community transmission of this virus is the requirement for people to stay at home during the lockdown. Prolonged contact at home is one of the risks factors associated with tuberculosis transmission [139]. TB has a long incubation period, which means that a rise in cases may occur in the following years.

HIV/global epidemics in Africa can serve as an example: after the worldwide HIV/-global TB pandemic, epidemics occurred in many countries such as South Africa [140], public health must be maintained at the highest possible level of alert.

The worldwide pandemic of COVID-19 may affect the global strategy of ending TB by 2035 in several ways:

- Study and data on tuberculosis were not carried out in 2020, such as lectures, workshops and conferences. The World Tuberculosis Day, which takes place on 24 March each year and helps raise funding, awareness to support tb control, was repealed in several countries.

- COVID-19 has had a detrimental impact on vaccination services, such as the BCG vaccine used to combat pediatric tuberculosis

- Disturbance of TB preventive therapy, given to high-risk groups in order to prevent latent TB activation.

- Disruptions of funding are expected to affect routine public health programs, either by the divergence of resources to pandemic control or broader economic effects of the pandemic and strained national budgets.

7 Impact of COVID-19 on late reactivation of TB:

The impact of COVID-19 on the immune system could be associated with a higher risk of developing active TB disease [141]. Pneumonic or respiratory failure caused by COVID-19 may promote lung complications, such as chronic inflammation and respiratory system damage, increasing the risk of developing tb [141]. Association between some diseases such as HIV/AIDS and TB development has been previously established [142]; the active TB can be aquired directly following TB exposure or through reactivation of latent TB infection.

8 Strategies to reduce the Impact of COVID-19 on TB Control

To reduce household transmission of TB, the simple prevention and control measures recommended by the WHO can be implemented at home [143]. One way is the use of virtual care and digital health technologies to overcome TB diagnosis and treatment delays due to COVID-19. Moreover, decentralizing TB care to community health workers and assisting private health sectors and academic research institutions in providing TB testing and treatment could all be effective.

9 Impact of the COVID-19 on scientific research and lessons to learn

We estimate that trillions of dollars have been invested in responding to the COVID-19 pandemic [144], supporting clinical and public health work, and quickening the rate of scientific progress. In certain ways, the urgency and Covidization are understandable that research is needed desperately to combat this pandemic. Besides, Crisis-led urgent approaches have a successful track record when it comes to being funded on the heels of a crisis; that was the case during the Ebola and AIDS outbreaks in West Africa which resulted in the emergence of numerous therapies and diagnostics over a short period. However, the sense of urgency can be at the expense of other more dangerous diseases by delaying or cancelling funds which can have a massive impact on the after COVID-19 period.

As Scientists, we must consider the future too and the implications of COVID-19 oriented strategies. Well-intentioned scientists who possess specialized knowledge in fields not related to COVID-19 spend a tremendous amount of energy turning their work, and likely making monumental errors as they jump into unrelated fields while lacking expertise and insight. We have seen floods of highly questionable and proven-unfounded research as well as a lowering of scientific standards. For others, it may be time to return to their original research topics, in which the magnitude of their accomplishments will be more significant. There is a significant risk of pivoting to a new research area only to discover the target vanished and that the solution itself is no longer needed.

In general, Findings should be reproducible; therefore, some degree of duplication is needed. However, many of the published duplications were unnecessary, Leading to enormous redundancy and waste of time. A quick search in the PubMed database showed more than 1300 published papers mentioning SARS-Cov-2 and ACE2. Hundreds of these scientific papers focus on blocking the entry of viruses into cells with a clear overlap.

In Morocco, the COVID-19 had drawn attention to the already overburdened public health systems and shed light on the state of scientific research, institution, logistics. Effort and attention were oriented to face the pandemic raising many issues in the scientific community:

9.1 Collaboration over competition

By the start of the pandemic, the institutions with access to genomic tools struggled to get the first sequencing results due to difficult access to the viral strains. These circumstances highlighted Morocco's major problem, which is a lack of collaboration between institutions. Competition is good in a way that it pushes scientists to do better but should not be encouraged over collaborations, especially with the lack of expertise in the genomic field.

9.2 Long term vision

The majority of institutions in Morocco have opted for having new sequencing equipment rather than using another platform. A huge amount of funds has been used on equipment that could have helped fund other research areas such as drug or vaccine development.

What is currently needed to compete with the rest of the world is a more forward-looking and imaginative vision of how this pathology could affect life after the end of this pandemic by exploring etiology, pathogenesis, pathophysiology, and vaccines.

9.3 Other deadly diseases DO exist:

I think that scientists should not get distracted from other health-related issues that are as equally, if not even more, life-threatening. Virologist especially should shift their sight to the hundreds of viruses colonizing other animals worldwide and adapt curiosity-based research plans to evaluate the future potential risk of a spillover in order to have better control in the future. For that scientist need stable fundings to ensure surveillance and vaccine or drug development.

A year ago, we considered the emergence of MDR bacterial strains the greatest pandemic threat, but the COVID-19 prove us wrong for now.

However, the threat is real. In the USA, over 2.8M MDR cases occur each year, 1.25% die. The situation is much worse in third-world countries. The fight against multidrug resistance is concerning as our most effective weapons are losing their efficacy. The pipeline for new antibiotics is running dry as new antibiotics are time-consuming and unrewarding for university researchers.

However, the main reason that makes antibiotic development so unattractive is the Revenues. As it was clearly explained in a review by Renwick, "Sales volumes are limited by the short treatment duration inherent in antibiotic therapy, and local antimicrobial stewardship programs are increasingly restricting the use of antibiotics. A truly novel antibiotic would likely be reserved for rare infections caused by the most highly resistant strains of bacteria" [145].

Although we don't know whether another zoonotic RNA virus can evolve and thrive in our bodies, we do know that multidrug-resistant bacteria will spread. As a result, if governments do not make significant investments in RD to combat pandemic threats, the current COVID-19 episode could be the preface to a much worse global era.

9.4 More conventional management:

The case of COVID-19 has proven that the research management system is weak and is often left at the mercy of inter mediated companies and bureaucracy. I think efforts should be made to accelerate delivery and facilitate the process. Maybe it's time for universities to have a department for scientific research management and not go through all the layers of bureaucracy involving people with no scientific background, facilitating access to external collaboration and funds. Meanwhile, the scientist's ability to get funding has predicated on the competition for decades, it is blocked by the long process and imaginary limitations for Morocco.

9.5 Investing in science:

COVID-19 revealed the potential black hole in the finances of universities and active bodies carrying out scientific research. Usually, grant application review takes months if not years, decision during the pandemic, even in Morocco, was within weeks, while the management of the finances could have been better. As we all agree, the fastest the funds are delivered to the university account, the quicker the research team can be immobilized. I believe that this pandemic was also a great example of the potential effect of rapid funding on science and helped develop many homegrown solutions that can boost diagnosis for many other diseases in Morocco. Furthermore, governments must continue to invest in physical infrastructure such as laboratories to experiment and innovate. I also believe that in order to increase resources, researchers should be aware of and apply for international funds for research and development. Obviously, we should not rely completely on foreign aids, especially with the COVID-19situation where funds for Africa, in general, will be reduced.

Here is a non-complete list of "neglected" research areas, where increased research budgets could have given us a head-start:

1. Vaccination
2. Pharmaceutical targets in zoonotic viruses
3. In-house diagnosis test
4. Clear and efficient pandemic plans and models
5. Antibiotics Development
6. Infrastructural requirements

9.6 Investing in young scientists:

I believe that Investing in science also means investing in future scientists. It means encouraging young people to seek careers in science and giving them the right educational and professional opportunities. The priority should be given to quality over quantity; ensuring that students get adequate training and exposure to advanced science will have a positive impact on their creativity and productivity. Training young scientists also requires investments in further professional development opportunities in the early and mid-career stages to provide enough incentives and hope.

10 The positive impact of COVID-19

On the bright side, COVID-19 has proven how easy and fast sharing of biological data can be; the use of preprints has further improved transparency and encourage discourse between scientists and the public. I think that is bringing back the meaning of science free and accessible and will help encourage science for sharing rather than for papers. Even Scientific journals have reduced their replaying time. I think that These prove that the research community can respond quickly and that the layers of bureaucracy normally shrouding clinical research can be reduced.

Lastly, many laboratories have been equipped with equipment that will be used in other research areas. More importantly, In morocco, the genomics field has finally been installed as a strong tool for tracking and controlling infectious disease. I think with a good mindset, we can finally go back to the competition and have a clear scientific impact on the local community and the world as a whole, because One thing is certain: pandemic pathogens will hit again!

Chapter 6

Conslusion

This thesis presents *Mtb* and SARS-CoV-2 genomic data analysis to characterise the variation and downstream effects, identify the transmission dynamic, and classify the lineages. Chapter 2 and 3 focus on the sequencing and genomic analysis of *Mtb* and Sars-CoV-2 and the applicability of NGS to predict drug resistance, diversity , transmission and predict the effect of mutation on the pathogen evolution. Chapter 4 integrates large scale SARS-CoV-2 data to look for specific variation to geographic locations to analyze the transmission dynamic and variations effect. I hope that this data will enable researchers to answer questions concerning the diversity of *Mtb* and SARS-CoV-2. The results from this work will contribute to our understanding of these complex pathogens. Several fruitful avenues of research could emerge from this work:

- In vitro verification of the Docking results
- In order to have a better idea of the population structure of *Mtb* in Morocco, increasing the sampling size is necessary.
- Characterization of host-pathogen interactions to study the impact of the Moroccan genotype on the infection
- A deep analysis of tb and sars-Cov-2 in Morocco has not been explored; Transcriptional sequencing would allow us to evaluate the pathogenicity level of the circulating strain
- Identifying the Co-infection pattern of TB and SARS-CoV-2

The future of infectious disease genomic analysis :

- WGS has been proven to be a robust tool for infectious disease control. National TB programs should start to adopt WGS as a diagnostic tool.
- Data centralization: sequence data generated in clinics should be sent to a centralised database along with metadata such as location and date of collection. This could help develop control measures and routes of transmission and help TB programs identify high-risk areas and where to focus efforts.

- Sputum samples do not allow the proper identification of the other pathogen present. Deep sequencing without culture requirement would enable the sampling of multiple pathogens at once, such as *M. tuberculosis* and HIV, SARS-coV-2, which are well known to co-infect.

- With the installation of Genomics in Morocco, other infectious diseases should be under control, such as leishmaniose, *M.Bovis* ...

Bibliography

[1] Anthony S Fauci and David M Morens. The perpetual challenge of infectious diseases. *New England Journal of Medicine*, 366(5):454–461, 2012.

[2] Matteo Fumagalli, Uberto Pozzoli, Rachele Cagliani, Giacomo P Comi, Nereo Bresolin, Mario Clerici, and Manuela Sironi. Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. *PLoS Genet*, 6(2):e1000849, 2010.

[3] David M Morens, Gregory K Folkers, and Anthony S Fauci. Emerging infections: a perpetual challenge. *The Lancet infectious diseases*, 8(11):710–719, 2008.

[4] Mika J Mäkelä, Tuomo Puhakka, Olli Ruuskanen, Maija Leinonen, Pekka Saikku, Marko Kimpimäki, Soile Blomqvist, Timo Hyypiä, and Pertti Arstila. Viruses and bacteria in the etiology of the common cold. *Journal of clinical microbiology*, 36(2):539–542, 1998.

[5] Steven A Dosh, John M Hickner, ARCH G III MAINOUS, and Mark H Ebell. Predictors of antibiotic prescribing for nonspecific upper respiratory infections, acute bronchitis, and acute sinusitis. *Journal of Family Practice*, 49(5):407–407, 2000.

[6] Yongyu Liu, Lijun Bi, Yu Chen, Yaguo Wang, Joy Fleming, Yanhong Yu, Ye Gu, Chang Liu, Lichao Fan, Xiaodan Wang, et al. Active or latent tuberculosis increases susceptibility to covid-19 and disease severity. *MedRxiv*, 2020.

[7] World Health Organization. Global tuberculosis report 2019. 2019.

[8] CJ Cambier, Stanley Falkow, and Lalita Ramakrishnan. Host evasion and exploitation schemes of mycobacterium tuberculosis. *Cell*, 159(7):1497–1509, 2014.

[9] WILLIAM FIRTH Wells, HERBERT L Ratcliffe, Cretyl Crumb, et al. On the mechanics of droplet nuclei infection. ii. quantitative experimental air-borne tuberculosis in rabbits. *American journal of hygiene*, 47(1):11–28, 1948.

[10] Lalita Ramakrishnan. Revisiting the role of the granuloma in tuberculosis. *Nature Reviews Immunology*, 12(5):352–366, 2012.

[11] World Health Organization et al. Covid-19 weekly epidemiological update, 22 december 2020. 2020.

[12] Thomas M Daniel, Joseph H Bates, and Katharine A Downes. History of tuberculosis. *Tuberculosis: pathogenesis, protection, and control*, pages 13–24, 1994.

[13] World Health Organization et al. Global tuberculosis report 2020: executive summary. 2020.

[14] Jia-Ru Chang, Yih-Yuan Chen, Tsi-Shu Huang, Wei-Feng Huang, Shu-Chen Kuo, Fan-Chen Tseng, Ih-Jen Su, Chien-Hsing Lin, Yao-Shen Chen, Jun-Ren Sun, et al. Clonal expansion of both modern and ancient genotypes of mycobacterium tuberculosis in southern taiwan. *PloS one*, 7(8):e43018, 2012.

[15] I Sutherland. Recent studies in the epidemiology of tuberculosis, based on the risk of being infected with tubercle bacilli. *Advances in tuberculosis research. Fortschritte der Tuberkuloseforschung. Progres de l'exploration de la tuberculose*, 19:1–63, 1976.

[16] Pam Sonnenberg, Judith R Glynn, Katherine Fielding, Jill Murray, Peter Godfrey-Faussett, and Stuart Shearer. How soon after infection with hiv does the risk of tuberculosis start to increase? a retrospective cohort study in south african gold miners. *Journal of Infectious Diseases*, 191(2):150–158, 2005.

[17] Vadim Pokrovsky. Tuberculosis and hiv/aids: the alien and the predator. *Lancet (London, England)*, 390(10102):1618–1619, 2017.

[18] Riccardo Alagna, Giorgio Besozzi, Luigi Ruffo Codecasa, Andrea Gori, Giovanni Battista Migliori, Mario Raviglione, and Daniela Maria Cirillo. Celebrating tb day at the time of covid-19. *The European Respiratory Journal.*

[19] Claudia Stochino, Simone Villa, Patrizia Zucchi, Pierpaolo Parravicini, Andrea Gori, and Mario Carlo Raviglione. Clinical characteristics of covid-19 and active tuberculosis co-infection in an italian reference hospital. *European Respiratory Journal*, 56(1), 2020.

[20] Marina Tadolini, Luigi Ruffo Codecasa, José-María García-García, François-Xavier Blanc, Sergey Borisov, Jan-Willem Alffenaar, Claire Andréjak, Pierre Bachez, Pierre-Alexandre Bart, Evgeny Belilovski, et al. Active tuberculosis, sequelae and covid-19 co-infection: first cohort of 49 cases. *European Respiratory Journal*, 56(1), 2020.

[21] Ilaria Motta, Rosella Centis, Lia D'Ambrosio, J-M García-García, Delia Goletti, Gina Gualano, Filippo Lipani, Fabrizio Palmieri, Adrián Sánchez-Montalvá, Emanuele Pontali, et al. Tuberculosis, covid-19 and migrants: preliminary analysis of deaths occurring in 69 patients from two cohorts. *Pulmonology*, 26(4):233–240, 2020.

[22] World Health Organization et al. End tb strategy. 2020.

[23] Andrew D Kerkhoff, David A Barr, Charlotte Schutz, Rosie Burton, Mark P Nicol, Stephen D Lawn, and Graeme Meintjes. Disseminated tuberculosis among hospitalised hiv patients in south africa: a common condition that can be rapidly diagnosed using urine-based assays. *Scientific reports*, 7(1):1–11, 2017.

[24] Ndivhuho A Makhado, Edith Matabane, Mauro Faccin, Claire Pinçon, Agathe Jouet, Fairouz Boutachkourt, Léonie Goeminne, Cyril Gaudin, Gugu Maphalala, Patrick Beckert, et al. Outbreak of multidrug-resistant tuberculosis in south africa undetected by who-endorsed commercial tests: an observational study. *The Lancet Infectious Diseases*, 18(12):1350–1359, 2018.

[25] Keertan Dheda, Tawanda Gumbo, Neel R Gandhi, Megan Murray, Grant Theron, Zarir Udwadia, GB Migliori, and Robin Warren. Global control of tuberculosis: from extensively drug-resistant to untreatable tuberculosis. *The lancet Respiratory medicine*, 2(4):321–338, 2014.

[26] World Health Organization. Global tuberculosis report 2018. 2018.

[27] Direction de l'Epidémiologie et de Lutte Contre les Maladies. Situation epidémiologique de la tuberculose au maroc. *Ministère de la Santé.*

[28] Ministère de la Santé. Plan national d'accélération de la réduction de l'incidence de la tuberculose 2013–2016.

[29] Hind Karimi, Amal Oudghiri, Latifa En-Nanei, Mohammed El Mzibri, Amin Laglaoui, Imane Chaoui, and Mohammed Abid. Frequency of genomic mutations mediating resistance of mycobacterium tuberculosis isolates to rifampicin in northern morocco. *Revista do Instituto de Medicina Tropical de São Paulo*, 62, 2020.

[30] Amal Oudghiri, Hind Karimi, Fouad Chetioui, Fathiah Zakham, Jamal Eddine Bourkadi, My Driss Elmessaoudi, Amin Laglaoui, Imane Chaoui, and Mohammed El Mzibri. Molecular characterization of mutations associated with resistance to second-line tuberculosis drug among multidrug-resistant tuberculosis patients from high prevalence tuberculosis city in morocco. *BMC infectious diseases*, 18(1):1–8, 2018.

[31] Imane Chaoui, Amal Oudghiri, and Mohammed El Mzibri. Characterization of gyra and gyrb mutations associated with fluoroquinolone resistance in mycobacterium tuberculosis isolates from morocco. *Journal of global antimicrobial resistance*, 12:171–174, 2018.

[32] Wifak Ennassiri, Sanae Jaouhari, Wafa Cherki, Reda Charof, Abdelkarim Filali-Maltouf, and Ouafae Lahlou. Extensively drug-resistant tuberculosis (xdr-tb) in morocco. *Journal of global antimicrobial resistance*, 11:75–80, 2017.

[33] Hanaa Mouchrik, Abdelmajid Soulaymani, Mhammed Jabri, Hinde Hami, Abdelrhani Mokhtari, et al. Pulmonary tuberculosis in morocco: A two year retrospective study. *Journal of Tuberculosis Research*, 6(01):104, 2018.

[34] Mina Sadeq and Jamal Eddine Bourkadi. Spatiotemporal distribution and predictors of tuberculosis incidence in morocco. *Infectious diseases of poverty*, 7(1):1–13, 2018.

[35] Mena Cimino, Lorenzo Alamo, and Leiria Salazar. Permeabilization of the mycobacterial envelope for protein cytolocalization studies by immunofluorescence microscopy. *BMC microbiology*, 6(1):1–3, 2006.

[36] Forbes. Mycobacterial taxonomy. *J Clin Microbiol*, 55(2):380–383, 2017.

[37] Roland Brosch, Stephen V Gordon, M Marmiesse, P Brodin, C Buchrieser, K Eiglmeier, T Garnier, C Gutierrez, G Hewinson, K Kremer, et al. A new evolutionary scenario for the mycobacterium tuberculosis complex. *Proceedings of the national academy of Sciences*, 99(6):3684–3689, 2002.

[38] Joel D Ernst, Giraldina Trevejo-Nuñez, Niaz Banaiee, et al. Genomics and the evolution, pathogenesis, and diagnosis of tuberculosis. *The Journal of clinical investigation*, 117(7):1738–1745, 2007.

[39] Anastasia Koch and Valerie Mizrahi. Mycobacterium tuberculosis. *Trends in microbiology*, 26(6):555–556, 2018.

[40] James J Dunn, Jeffrey R Starke, and Paula A Revell. Laboratory diagnosis of mycobacterium tuberculosis infection and disease in children. *Journal of clinical microbiology*, 54(6):1434–1441, 2016.

[41] Erik C Hett, Michael C Chao, Lynn L Deng, and Eric J Rubin. A mycobacterial enzyme essential for cell division synergizes with resuscitation-promoting factor. *PLoS Pathog*, 4(2):e1000001, 2008.

[42] Pablo J Bifani, Barun Mathema, Natalia E Kurepina, and Barry N Kreiswirth. Global dissemination of the mycobacterium tuberculosis w-beijing family strains. *Trends in microbiology*, 10(1):45–52, 2002.

[43] Jennifer A Philips and Joel D Ernst. Tuberculosis pathogenesis and immunity. *Annual Review of Pathology: Mechanisms of Disease*, 7:353–384, 2012.

[44] Noton K Dutta and Petros C Karakousis. Latent tuberculosis infection: myths, models, and molecular mechanisms. *Microbiology and Molecular Biology Reviews*, 78(3):343–371, 2014.

[45] Mireia Coscolla and Sebastien Gagneux. Consequences of genomic diversity in mycobacterium tuberculosis. In *Seminars in immunology*, volume 26, pages 431–444. Elsevier, 2014.

[46] Judith R Glynn, Jennifer Whiteley, Pablo J Bifani, Kristin Kremer, and Dick van Soolingen. Worldwide occurrence of beijing/w strains of mycobacterium tuberculosis: a systematic review. *Emerging infectious diseases*, 8(8):843, 2002.

[47] Ruth Hershberg, Mikhail Lipatov, Peter M Small, Hadar Sheffer, Stefan Niemann, Susanne Homolka, Jared C Roach, Kristin Kremer, Dmitri A Petrov, Marcus W Feldman, et al. High functional diversity in mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biol*, 6(12):e311, 2008.

[48] Inaki Comas, Susanne Homolka, Stefan Niemann, and Sebastien Gagneux. Genotyping of genetically monomorphic bacteria: Dna sequencing in mycobacterium tuberculosis highlights the limitations of current methodologies. *PloS one*, 4(11):e7815, 2009.

[49] Florian Gehre, Jacob Otu, Lindsay Kendall, Audrey Forson, Awewura Kwara, Samuel Kudzawu, Aderemi O Kehinde, Oludele Adebiyi, Kayode Salako, Ignatius Baldeh, et al. The emerging threat of pre-extensively drug-resistant tuberculosis in west africa: preparing for large-scale tuberculosis research and drug resistance surveillance. *BMC medicine*, 14(1):1–12, 2016.

[50] Erasto V Mbugi, Bugwesa Z Katale, Elizabeth M Streicher, Julius D Keyyu, Sharon L Kendall, Hazel M Dockrell, Anita L Michel, Mark M Rweyemamu, Robin M Warren, Mecky I Matee, et al. Mapping of mycobacterium tuberculosis complex genetic diversity profiles in tanzania and other african countries. *PloS one*, 11(5):e0154571, 2016.

[51] Jose A Caminero, María J Pena, Maria I Campos-Herrero, José C Rodríguez, Isabel Garcia, Pedro Cabrera, Carmen Lafoz, Sofia Samper, Howard Takiff, Octavio Afonso, et al. Epidemiological evidence of the spread of a mycobacterium tuberculosis strain of the beijing genotype on gran canaria island. *American journal of respiratory and critical care medicine*, 164(7):1165–1170, 2001.

[52] Thomas R Frieden, Lisa Fine Sherman, Khin Lay Maw, Paula I Fujiwara, Jack T Crawford, Beth Nivin, Victoria Sharp, Dial Hewlett, Karen Brudney, David Alland, et al. A multi-institutional outbreak of highly drug-resistant tuberculosis: epidemiology and clinical outcomes. *Jama*, 276(15):1229–1235, 1996.

[53] Ida Parwati, Reinout van Crevel, and Dick van Soolingen. Possible underlying mechanisms for successful emergence of the mycobacterium tuberculosis beijing genotype strains. *The Lancet infectious diseases*, 10(2):103–111, 2010.

[54] Judith R Glynn, Kristin Kremer, Martien W Borgdorff, Mar Pujades Rodriguez, and Dick van Soolingen. Beijing/w genotype mycobacterium tuberculosis and drug resistance. 2006.

[55] Kristin Kremer, Marieke J van der Werf, Betty KY Au, Dang D Anh, Kai M Kam, H Rogier Van Doorn, Martien W Borgdorff, and Dick van Soolingen. Vaccine-induced immunity circumvented by typical mycobacterium tuberculosis beijing strains. *Emerging infectious diseases*, 15(2):335, 2009.

[56] Leo Kang-Yang Lim, Li Hwei Sng, Wah Win, Cynthia Bin-Eng Chee, Li Yang Hsu, Estelle Mak, Arul Earnest, Marcus Eng-Hock Ong, Jeffery Cutter, and Yee Tang Wang. Molecular epidemiology of mycobacterium tuberculosis complex in singapore, 2006-2012. *PLoS One*, 8(12):e84487, 2013.

[57] Bright Varghese, Philip Supply, Mohammed Shoukri, Caroline Allix-Beguec, Ziad Memish, Naila Abuljadayel, Raafat Al-Hakeem, Fahad AlRabiah, and Sahal Al-Hajoj.

Tuberculosis transmission among immigrants and autochthonous populations of the eastern province of saudi arabia. *PloS one*, 8(10):e77635, 2013.

[58] Dick Van Soolingen, Lishi Qian, PE De Haas, James T Douglas, Hamadou Traore, Francoise Portaels, Huang Zi Qing, D Enkhsaikan, P Nymadawa, and JD Van Embden. Predominance of a single genotype of mycobacterium tuberculosis in countries of east asia. *Journal of clinical microbiology*, 33(12):3234–3238, 1995.

[59] Biljo V Joseph, Smitha Soman, Indulakshmi Radhakrishnan, Véronique Hill, D Dhanasooraj, R Ajay Kumar, Nalin Rastogi, and Sathish Mundayoor. Molecular epidemiology of mycobacterium tuberculosis isolates from kerala, india using is6110-rflp, spoligotyping and miru-vntrs. *Infection, genetics and evolution*, 16:157–164, 2013.

[60] Jian Zhang, Seiha Heng, Stéphanie Le Moullec, Guislaine Refregier, Brigitte Gicquel, Christophe Sola, and Bertrand Guillard. A first assessment of the genetic diversity of mycobacterium tuberculosis complex in cambodia. *BMC infectious diseases*, 11(1):1–7, 2011.

[61] Sabai Phyu, Ruth Stavrum, Thandar Lwin, Øyvind S Svendsen, Ti Ti, and Harleen MS Grewal. Predominance of mycobacterium tuberculosis eai and beijing lineages in yangon, myanmar. *Journal of clinical microbiology*, 47(2):335–344, 2009.

[62] Marc Choisy, Duy Hung Nguyen, Thanh Hoa Thi Tran, Kim Lien Thi Pham, Phuong Thao Thi Dinh, Jules Philippe, Thai Son Nguyen, Minh Ly Ho, Sang Van Tran, Anne-Laure Bañuls, et al. High prevalence of beijing and eai4-vnm genotypes among m. tuberculosis isolates in northern vietnam: sampling effect, rural and urban disparities. *PLoS One*, 7(9):e45553, 2012.

[63] Jia-Ru Chang, Yih-Yuan Chen, Tsi-Shu Huang, Wei-Feng Huang, Shu-Chen Kuo, Fan-Chen Tseng, Ih-Jen Su, Chien-Hsing Lin, Yao-Shen Chen, Jun-Ren Sun, et al. Clonal expansion of both modern and ancient genotypes of mycobacterium tuberculosis in southern taiwan. *PloS one*, 7(8):e43018, 2012.

[64] Yih-Yuan Chen, Jia-Ru Chang, Wei-Feng Huang, Shu-Ching Hsu, Shu-Chen Kuo, Jun-Ren Sun, and Horng-Yunn Dou. The pattern of cytokine production in vitro induced by ancient and modern beijing mycobacterium tuberculosis strains. *PLoS One*, 9(4):e94296, 2014.

[65] Horng-Yunn Dou, Shu-Chen Huang, and Ih-Jen Su. Prevalence of mycobacterium tuberculosis in taiwan: a model for strain evolution linked to population migration. *International journal of evolutionary biology*, 2011, 2011.

[66] Daniel J Bretl, Chrystalla Demetriadou, and Thomas C Zahrt. Adaptation to environmental stimuli within the host: two-component signal transduction systems of mycobacterium tuberculosis. *Microbiology and Molecular Biology Reviews*, 75(4):566–582, 2011.

[67] Alejandra N Martinez, Smriti Mehra, and Deepak Kaushal. Role of interleukin 6 in innate immunity to mycobacterium tuberculosis infection. *The Journal of infectious diseases*, 207(8):1253–1261, 2013.

[68] Charitha Mendis, Vasanthi Thevanesam, Athula Kumara, Susiji Wickramasinghe, Dushantha Madegedara, Chandika Gamage, Stephen V Gordon, Yasuhiko Suzuki, Champa Ratnatunga, and Chie Nakajima. Insight into genetic diversity of mycobacterium tuberculosis in kandy, sri lanka reveals predominance of the euro-american lineage. *International Journal of Infectious Diseases*, 87:84–91, 2019.

[69] Ola B Brynildsrud, Caitlin S Pepperell, Philip Suffys, Louis Grandjean, Johana Monteserin, Nadia Debech, Jon Bohlin, Kristian Alfsnes, John O-H Pettersson, Ingerid Kirkeleite, et al. Global expansion of mycobacterium tuberculosis lineage 4 shaped by colonial migration and local adaptation. *Science advances*, 4(10):eaat5869, 2018.

[70] Sebastien Gagneux. Ecology and evolution of mycobacterium tuberculosis. *Nature Reviews Microbiology*, 16(4):202, 2018.

[71] Damien Portevin, Sébastien Gagneux, Iñaki Comas, and Douglas Young. Human macrophage responses to clinical isolates from the mycobacterium tuberculosis complex discriminate between ancient and modern lineages. *PLoS pathog*, 7(3):e1001307, 2011.

[72] Boatema Ofori-Anyinam, Gregory Dolganov, Tran Van, J Lucian Davis, Nicholas D Walter, Benjamin J Garcia, Marty Voskuil, Kristina Fissette, Maren Diels, Michele Driesen, et al. Significant under expression of the dosr regulon in m. tuberculosis complex lineage 6 in sputum. *Tuberculosis*, 104:58–64, 2017.

[73] Bouke C de Jong, Philip C Hill, Roger H Brookes, Sebastien Gagneux, David J Jeffries, Jacob K Otu, Simon A Donkor, Annette Fox, Keith PWJ McAdam, Peter M Small, et al. Mycobacterium africanum elicits an attenuated t cell response to early secreted antigenic target, 6 kda, in patients with tuberculosis and their household contacts. *The Journal of infectious diseases*, 193(9):1279–1286, 2006.

[74] Jesús Gonzalo-Asensio, Wladimir Malaga, Alexandre Pawlik, Catherine Astarie-Dequeker, Charlotte Passemar, Flavie Moreau, Françoise Laval, Mamadou Daffé, Carlos Martin, Roland Brosch, et al. Evolutionary history of tuberculosis shaped by conserved mutations in the phopr virulence regulator. *Proceedings of the National Academy of Sciences*, 111(31):11491–11496, 2014.

[75] Adwoa Asante-Poku, Isaac Darko Otchere, Stephen Osei-Wusu, Esther Sarpong, Akosua Baddoo, Audrey Forson, Clement Laryea, Sonia Borrell, Frank Bonsu, Jan Hattendorf, et al. Molecular epidemiology of mycobacterium africanum in ghana. *BMC infectious diseases*, 16(1):1–8, 2016.

[76] Adwoa Asante-Poku, Dorothy Yeboah-Manu, Isaac Darko Otchere, Samuel Y Aboagye, David Stucki, Jan Hattendorf, Sonia Borrell, Julia Feldmann, Emelia Danso,

and Sebastien Gagneux. Mycobacterium africanum is associated with patient ethnicity in ghana. *PLoS Negl Trop Dis*, 9(1):e3370, 2015.

[77] Andrea Gudan, Branka Artuković, Željko Cvetnić, Silvio Špičić, Ana Beck, Marko Hohšteter, Tomo Naglić, Ingeborg Bata, and Željko Grabarević. Disseminated tuberculosis in hyrax (procavia capensis) caused by mycobacterium africanum. *Journal of Zoo and Wildlife Medicine*, 39(3):386–391, 2008.

[78] Rebuma Firdessa, Stefan Berg, Elena Hailu, Esther Schelling, Balako Gumi, Girume Erenso, Endalamaw Gadisa, Teklu Kiros, Meseret Habtamu, Jemal Hussein, et al. Mycobacterial lineages causing pulmonary and extrapulmonary tuberculosis, ethiopia. *Emerging infectious diseases*, 19(3):460, 2013.

[79] Yann Blouin, Yolande Hauck, Charles Soler, Michel Fabre, Rithy Vong, Céline Dehan, Géraldine Cazajous, Pierre-Laurent Massoure, Philippe Kraemer, Akinbowale Jenkins, et al. Significance of the identification in the horn of africa of an exceptionally deep branching mycobacterium tuberculosis clade. *PloS one*, 7(12):e52841, 2012.

[80] Solomon A Yimer, Gunnstein Norheim, Amine Namouchi, Ephrem D Zegeye, Wibeke Kinander, Tone Tønjum, Shiferaw Bekele, Turid Mannsåker, Gunnar Bjune, Abraham Aseffa, et al. Mycobacterium tuberculosis lineage 7 strains are associated with prolonged patient delay in seeking treatment for pulmonary tuberculosis in amhara region, ethiopia. *Journal of clinical microbiology*, 53(4):1301–1309, 2015.

[81] Jean Claude Semuto Ngabonziza, Chloé Loiseau, Michael Marceau, Agathe Jouet, Fabrizio Menardo, Oren Tzfadia, Rudy Antoine, Esdras Belamo Niyigena, Wim Mulders, Kristina Fissette, et al. A sister lineage of the mycobacterium tuberculosis complex discovered in the african great lakes region. *Nature communications*, 11(1):1–11, 2020.

[82] Mireia Coscolla, Sebastien Gagneux, Fabrizio Menardo, Chloé Loiseau, Paula Ruiz-Rodriguez, Sonia Borrell, Isaac Darko Otchere, Adwoa Asante-Poku, Prince Asare, Leonor Sánchez-Busó, et al. Phylogenomics of mycobacterium africanum reveals a new lineage and a complex evolutionary history. *Microbial genomics*, 7(2):000477, 2021.

[83] Ministère de la Santé. *Prise en charge de la tuberculose chez l'enfant, l'adolescent et l'adulte.* PhD thesis, 2020.

[84] Cesare Saltini. Chemotherapy and diagnosis of tuberculosis. *Respiratory medicine*, 100(12):2085–2097, 2006.

[85] L Hadj-Kacem, H Hadj-Kacem, N Chakroun-Fki, A Bahloul, MN Mhiri, T Rebai, H Ayadi, and L Ammar-Keskes. Screening of y chromosome microdeletions in tunisian infertile men. *Archives of andrology*, 52(3):169–174, 2006.

[86] Ajit Lalvani. Diagnosing tuberculosis infection in the 21st century: new tools to tackle an old enemy. *Chest*, 131(6):1898–1906, 2007.

[87] Ibrahim O Al-Orainey. Diagnosis of latent tuberculosis: Can we do better? *Annals of thoracic medicine*, 4(1):5, 2009.

[88] Danica Helb, Martin Jones, Elizabeth Story, Catharina Boehme, Ellen Wallace, Ken Ho, JoAnn Kop, Michelle R Owens, Richard Rodgers, Padmapriya Banada, et al. Rapid detection of mycobacterium tuberculosis and rifampin resistance by use of on-demand, near-patient technology. *Journal of clinical microbiology*, 48(1):229–237, 2010.

[89] Joacim Rocklöv, Henrik Sjödin, and Annelies Wilder-Smith. Covid-19 outbreak on the diamond princess cruise ship: estimating the epidemic potential and effectiveness of public health countermeasures. *Journal of travel medicine*, 27(3):taaa030, 2020.

[90] Tao Zhang, Qunfu Wu, and Zhigang Zhang. Probable pangolin origin of sars-cov-2 associated with the covid-19 outbreak. *Current biology*, 30(7):1346–1351, 2020.

[91] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. The proximal origin of sars-cov-2. *Nature medicine*, 26(4):450–452, 2020.

[92] Nisreen A Alwan. Surveillance is underestimating the burden of the covid-19 pandemic. *The Lancet*, 396(10252):e24, 2020.

[93] Vitaly Volpert. Mathematical modelling in the era of coronavirus (six month in a new reality), 2020.

[94] Yujiro Toyoshima, Kensaku Nemoto, Saki Matsumoto, Yusuke Nakamura, and Kazuma Kiyotani. Sars-cov-2 genomic variations associated with mortality rate of covid-19. *Journal of human genetics*, 65(12):1075–1082, 2020.

[95] Qianqian Li, Jiajing Wu, Jianhui Nie, Li Zhang, Huan Hao, Shuo Liu, Chenyan Zhao, Qi Zhang, Huan Liu, Lingling Nie, et al. The impact of mutations in sars-cov-2 spike on viral infectivity and antigenicity. *Cell*, 182(5):1284–1294, 2020.

[96] Barnaby E Young, Siew-Wai Fong, Yi-Hao Chan, Tze-Minn Mak, Li Wei Ang, Danielle E Anderson, Cheryl Yi-Pin Lee, Siti Naqiah Amrun, Bernett Lee, Yun Shan Goh, et al. Effects of a major deletion in the sars-cov-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *The Lancet*, 396(10251):603–611, 2020.

[97] David Harrington, Beatrix Kele, Spiro Pereira, Xose Couto-Parada, Anna Riddell, Suzanne Forbes, Hamish Dobbie, and Teresa Cutino-Moguel. Confirmed reinfection with sars-cov-2 variant voc-202012/01. *Clinical Infectious Diseases*, 2021.

[98] Nicholas G Davies, Rosanna C Barnard, Christopher I Jarvis, Adam J Kucharski, James Munday, Carl AB Pearson, Timothy W Russell, Damien C Tully, Sam Abbott, Amy Gimma, et al. Estimated transmissibility and severity of novel sars-cov-2 variant of concern 202012/01 in england. *medRxiv*, 2020.

[99] P Conti, Al Caraffa, CE Gallenga, SK Kritas, I Frydas, A Younes, P Di Emidio, G Tetè, F Pregliasco, and G Ronconi. The british variant of the new coronavirus-19 (sars-cov-2) should not create a vaccine problem. *Journal of biological regulators and homeostatic agents*, 35(1), 2021.

[100] Philippe Gautret, Jean-Christophe Lagier, Philippe Parola, Line Meddeb, Morgane Mailhe, Barbara Doudier, Johan Courjon, Valérie Giordanengo, Vera Esteves Vieira, Hervé Tissot Dupont, et al. Hydroxychloroquine and azithromycin as a treatment of covid-19: results of an open-label non-randomized clinical trial. *International journal of antimicrobial agents*, 56(1):105949, 2020.

[101] Helena F Florindo, Ron Kleiner, Daniella Vaskovich-Koubi, Rita C Acúrcio, Barbara Carreira, Eilam Yeini, Galia Tiram, Yulia Liubomirski, and Ronit Satchi-Fainaro. Immune-mediated approaches against covid-19. *Nature nanotechnology*, 15(8):630–645, 2020.

[102] Shibo Jiang, Christopher Hillyer, and Lanying Du. Neutralizing antibodies against sars-cov-2 and other human coronaviruses. *Trends in immunology*, 41(5):355–359, 2020.

[103] Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, et al. On the origin and continuing evolution of sars-cov-2. *National Science Review*, 7(6):1012–1023, 2020.

[104] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[105] Puja Mehta, Daniel F McAuley, Michael Brown, Emilie Sanchez, Rachel S Tattersall, and Jessica J Manson. Covid-19: consider cytokine storm syndromes and immunosuppression. *The lancet*, 395(10229):1033–1034, 2020.

[106] Luke J Alderwick, Virginie Molle, Laurent Kremer, Alain J Cozzone, Timothy R Dafforn, Gurdyal S Besra, and Klaus Fütterer. Molecular structure of embr, a response element of ser/thr kinase signaling in mycobacteriumtuberculosis. *Proceedings of the National Academy of Sciences*, 103(8):2558–2563, 2006.

[107] LJ Alderwick, GS Lloyd, H Ghadbane, JW May, A Bhatt, et al. The c-terminal domain of the arabinosyltransferase mycobacterium tuberculosis. 2011.

[108] Christine E Cade, Adrienne C Dlouhy, Katalin F Medzihradszky, Saida Patricia Salas-Castillo, and Reza A Ghiladi. Isoniazid-resistance conferring mutations in mycobacterium tuberculosis katg: Catalase, peroxidase, and inh-nadh adduct formation activities. *Protein Science*, 19(3):458–474, 2010.

[109] Alimuddin Zumla and Michael S Niederman. The explosive epidemic outbreak of novel coronavirus disease 2019 (covid-19) and the persistent threat of respiratory tract

infectious diseases to global health security. *Current opinion in pulmonary medicine*, 2020.

[110] World Health Organization et al. List of blueprint priority diseases. *R&D Blueprint. World Health Organization*, 2020.

[111] Imane Chaoui, Thierry Zozio, Ouafae Lahlou, Radia Sabouni, Mohammed Abid, Rajae El Aouad, Mohammed Akrim, Said Amzazi, Nalin Rastogi, and Mohammed El Mzibri. Contribution of spoligotyping and miru-vntrs to characterize prevalent mycobacterium tuberculosis genotypes infecting tuberculosis patients in morocco. *Infection, Genetics and Evolution*, 21:463–471, 2014.

[112] Nada Bouklata, Philip Supply, Sanae Jaouhari, Reda Charof, Fouad Seghrouchni, Khalid Sadki, Youness El Achhab, Chakib Nejjari, Abdelkarim Filali-Maltouf, Ouafae Lahlou, et al. Molecular typing of mycobacterium tuberculosis complex by 24-locus based miru-vntr typing in conjunction with spoligotyping to assess genetic diversity of strains circulating in morocco. *PloS one*, 10(8):e0135695, 2015.

[113] Mohammad Asgharzadeh, Hossein Samadi Kafil, Ali Amoghi Roudsary, and Gholam Reza Hanifi. Tuberculosis transmission in northwest of iran: using miru-vntr, etr-vntr and is6110-rflp methods. *Infection, Genetics and Evolution*, 11(1):124–131, 2011.

[114] Timothy M Walker, Matthias Merker, Astrid M Knoblauch, Peter Helbling, Otto D Schoch, Marieke J Van Der Werf, Katharina Kranzer, Lena Fiebig, Stefan Kröger, Walter Haas, et al. A cluster of multidrug-resistant mycobacterium tuberculosis among patients arriving in europe from the horn of africa: a molecular epidemiological study. *The Lancet Infectious Diseases*, 18(4):431–440, 2018.

[115] Violet N Chihota, Antoinette Niehaus, Elizabeth M Streicher, Xia Wang, Samantha L Sampson, Peter Mason, Gunilla Källenius, Sayoki G Mfinanga, Marnomorney Pillay, Marisa Klopper, et al. Geospatial distribution of mycobacterium tuberculosis genotypes in africa. *PLoS One*, 13(8):e0200632, 2018.

[116] Malika Ifticene, Saïd Kaïdi, Mesbah-Mounir Khechiba, Djamel Yala, and Fadila Boulahbal. Genetic diversity of mycobacterium tuberculosis strains isolated in algeria: Results of spoligotyping. *International journal of mycobacteriology*, 4(4):290–295, 2015.

[117] Amine Namouchi, Anis Karboul, Besma Mhenni, Neila Khabouchi, Raja Haltiti, Ridha Ben Hassine, Bechir Louzir, Abdellatif Chabbou, and Helmi Mardassi. Genetic profiling of mycobacterium tuberculosis in tunisia: predominance and evidence for the establishment of a few genotypes. *Journal of Medical Microbiology*, 57(7):864–872, 2008.

[118] Kai Kupferschmidt. The pandemic virus is slowly mutating. but does it matter?, 2020.

[119] Santiago F Elena and Rafael Sanjuán. Adaptive value of high mutation rates of rna viruses: separating causes from consequences. *Journal of virology*, 79(18):11555–11558, 2005.

[120] Giovanni Delogu, Michael J Brennan, and Riccardo Manganelli. Pe and ppe genes: a tale of conservation and diversity. *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control*, pages 191–207, 2017.

[121] Rowena A Bull, Thiruni N Adikari, James M Ferguson, Jillian M Hammond, Igor Stevanovski, Alicia G Beukers, Zin Naing, Malinna Yeang, Andrey Verich, Hasindu Gamaarachchi, et al. Analytical validity of nanopore sequencing for rapid sars-cov-2 genome analysis. *Nature communications*, 11(1):1–8, 2020.

[122] Zhongming Zhao, Haipeng Li, Xiaozhuang Wu, Yixi Zhong, Keqin Zhang, Ya-Ping Zhang, Eric Boerwinkle, and Yun-Xin Fu. Moderate mutation rate in the sars coronavirus genome and its implications. *BMC evolutionary biology*, 4(1):1–9, 2004.

[123] Matthew Cotten, Simon J Watson, Paul Kellam, Abdullah A Al-Rabeeah, Hatem Q Makhdoom, Abdullah Assiri, Jaffar A Al-Tawfiq, Rafat F Alhakeem, Hossam Madani, Fahad A AlRabiah, et al. Transmission and evolution of the middle east respiratory syndrome coronavirus in saudi arabia: a descriptive genomic study. *The Lancet*, 382(9909):1993–2002, 2013.

[124] Yixuan J Hou, Shiho Chiba, Peter Halfmann, Camille Ehre, Makoto Kuroda, Kenneth H Dinnon, Sarah R Leist, Alexandra Schäfer, Noriko Nakajima, Kenta Takahashi, et al. Sars-cov-2 d614g variant exhibits efficient replication ex vivo and transmission in vivo. *Science*, 370(6523):1464–1468, 2020.

[125] Farid Rahimi and Amin Talebi Bezmin Abadi. Implications of the emergence of a new variant of sars-cov-2, vui-202012/01. *Archives of Medical Research*, 2021.

[126] Adam S Lauring and Emma B Hodcroft. Genetic variants of sars-cov-2—what do they mean? *JAMA*, 2021.

[127] Alifiya S Motiwala, Yang Dai, Edward C Jones-López, Soo-Hee Hwang, Jong Seok Lee, Sang Nae Cho, Laura E Via, Clifton E Barry III, and David Alland. Mutations in extensively drug-resistant mycobacterium tuberculosis that do not code for known drug-resistance mechanisms. *The Journal of infectious diseases*, 201(6):881–888, 2010.

[128] Prerna Arora, Stefan Pöhlmann, and Markus Hoffmann. Mutation d614g increases sars-cov-2 transmission. *Signal Transduction and Targeted Therapy*, 6(1):1–2, 2021.

[129] Ádám Nagy, Sándor Pongor, and Balázs Győrffy. Different mutations in sars-cov-2 associate with severe and mild outcome. *International journal of antimicrobial agents*, 57(2):106272, 2021.

[130] TB Stop. Partnership. the tb response is heavily impacted by the covid-19 pandemic, 2020.

[131] Alimuddin Zumla, BJ Marais, TD McHugh, M Maeurer, Adam Zumla, N Kapata, F Ntoumi, P Chanda-Kapata, S Mfinanga, R Centis, et al. Covid-19 and tuberculosis—threats and opportunities. *The International Journal of Tuberculosis and Lung Disease*, 24(8):757–760, 2020.

[132] Daniel J Carter, Philippe Glaziou, Knut Lönnroth, Andrew Siroka, Katherine Floyd, Diana Weil, Mario Raviglione, Rein MGJ Houben, and Delia Boccia. The impact of social protection and poverty elimination on global tuberculosis incidence: a statistical modelling analysis of sustainable development goal 1. *The Lancet Global Health*, 6(5):e514–e522, 2018.

[133] world bank. Covid-19 to add as many as 150 million extreme poor by 2021. 2020.

[134] Jérôme Hadjadj, Nader Yatim, Laura Barnabei, Aurélien Corneau, Jeremy Boussier, Nikaïa Smith, Hélène Péré, Bruno Charbit, Vincent Bondet, Camille Chenevier-Gobeaux, et al. Impaired type i interferon activity and inflammatory responses in severe covid-19 patients. *Science*, 369(6504):718–724, 2020.

[135] Lucas M Kimmig, David Wu, Matthew Gold, Natasha N Pettit, David Pitrak, Jeffrey Mueller, Aliya N Husain, Ece A Mutlu, and Gökhan M Mutlu. Il-6 inhibition in critically ill covid-19 patients is associated with increased secondary infections. *Frontiers in medicine*, 7, 2020.

[136] Luis E Escobar, Alvaro Molina-Cruz, and Carolina Barillas-Mury. Bcg vaccine protection from severe coronavirus disease 2019 (covid-19). *Proceedings of the National Academy of Sciences*, 117(30):17720–17726, 2020.

[137] JAC Sterne, LC Rodrigues, and IN Guedes. Does the efficacy of bcg decline with time since vaccination? *The international journal of tuberculosis and lung disease*, 2(3):200–207, 1998.

[138] Subhadra Nandakumar, Sunil Kannanganat, Karen M Dobos, Megan Lucas, John S Spencer, Rama Rao Amara, Bonnie B Plikaytis, James E Posey, and Suraj B Sable. Boosting bcg-primed responses with a subunit apa vaccine during the waning phase improves immunity and imparts protection against mycobacterium tuberculosis. *Scientific reports*, 6(1):1–14, 2016.

[139] D Guwatudde, M Nakakeeto, EC Jones-Lopez, A Maganda, A Chiunda, RD Mugerwa, JJ Ellner, G Bukenya, and Christopher C Whalen. Tuberculosis in household contacts of infectious cases in kampala, uganda. *American journal of epidemiology*, 158(9):887–898, 2003.

[140] Salim S Abdool Karim, Gavin J Churchyard, Quarraisha Abdool Karim, and Stephen D Lawn. Hiv infection and tuberculosis in south africa: an urgent need to escalate the public health response. *the Lancet*, 374(9693):921–933, 2009.

[141] Ya Gao, Ming Liu, Yamin Chen, Shuzhen Shi, Jie Geng, and Jinhui Tian. Association between tuberculosis and covid-19 severity and mortality: A rapid systematic review and meta-analysis. *Journal of medical virology*, 93(1):194–196, 2021.

[142] Stephen D Lawn and Gavin Churchyard. Epidemiology of hiv-associated tuberculosis running head: Epidemiology of tb/hiv. *Current Opinion in HIV and AIDS*, 4(4):325, 2009.

[143] Yohhei Hamada, Philippe Glaziou, Charalambos Sismanidis, and Haileyesus Getahun. Prevention of tuberculosis in household members: estimates of children eligible for treatment. *Bulletin of the World Health Organization*, 97(8):534, 2019.

[144] Robin Naidoo and Brendan Fisher. Reset sustainable development goals for a pandemic world, 2020.

[145] Mathew Renwick and Elias Mossialos. What are the economic barriers of antibiotic r&d and how can we overcome them? *Expert opinion on drug discovery*, 13(10):889–892, 2018.

1 Appendix Chapter 3

Table 6.1: GenBank accession numbers for orf1ab and spike genes downloaded from NCBI and used for selective pressure analysis

GenBank Accession No					
orf1ab					
QIS61084	QIC53222	QII57287	QII57197	QIT06925	QIS60736
QIS61108	QII57317	QII57177	QHU36863	QIT06937	QIS60760
QIS61132	QIE07450	QIK50426	QHZ00378	QIT06949	QIS60772
QIS61144	QIC53203	QHU36843	QID21047	QIS60298	QIS60784
QIS61156	QID21067	QIA98605	QHR63249	QIS60328	QIS60796
QIS61204	QIK02943	QIE07480	QHW06048	QIS60338	QIS60808
QIS61216	QHS34545	QHU79193	QII57337	QIS60286	QIS60832
QIS61252	QHR63289	QII57237	QHU36823	QIS60544	QIS60856
QIS61300	QHR84448	QIJ96512	QHZ87581	QIS60568	QIS60868
QIS61312	QIE07460	QII57167	QID98793	QIS60592	QIS60880
QIS61360	QIA20042	QIK02953	QHQ71962	QIS60616	QIS60904
QIS61384	QHW06058	QIH55220	QHQ71972	QIS60640	QIS60916
QIS61420	QIK50447	QII57267	QIT06877	QIS60676	QIS60988
QIS61480	QIE07470	QII57187	QIT06889	QIS60712	QIS61000
QIS61492	QHQ82463	QHN73809	QIT06913	QIS60724	QIS61072
QII57165					
Spike					
QIS30425	QIS61254	QIS60582	QHZ00379	QHZ00379	QIC53213
QIS30495	QIS61338	QIS60774	QHZ87582	QHZ87582	QIC53204
QIS30615	QIS61422	QIS60858	QIA98583	QIA98583	QIK02944
QIS30625	QIS30275	QIS60906	QIS60489	QIS60489	QHR84449
QIS30675	QIS30295	QIS60930	QIS60288	QIS60288	QIK02964

Table 6.1: GenBank accession numbers for orf1ab and spike genes downloaded from NCBI and used for selective pressure analysis

GenBank Accession No					
QIQ49882	QIS30335	QIS60978	QIS60546	QIS60546	QIA20044
QIO04367	QIS30375	QIS61230	QIS60570	QIS60570	QHW06059
QIJ96493	QID21058	QII57278	QIK50427		

Table 6.2: Normalized data for singleton mutations identified in this study

Origin	**Genomes**	**N.M**	**O.M**	**Mut.Uq/G**	**O.M/G**	**Mut.uq/G**
Malaysia	10	29	38	2,9	3,8	0,763157895
China	334	220	329	0,658682635	0,98502994	0,668693009
USA	783	130	301	0,166028097	0,384418902	0,431893688
France	140	19	77	0,135714286	0,55	0,246753247
England	300	23	116	0,076666667	0,386666667	0,198275862
Spain	44	5	29	0,113636364	0,659090909	0,172413793
Finland	40	6	37	0,15	0,925	0,162162162
Iceland	334	12	78	0,035928144	0,233532934	0,153846154
Vietnam	10	2	15	0,2	1,5	0,133333333
Taiwan	21	4	33	0,19047619	1,571428571	0,121212121
Canada	123	6	68	0,048780488	0,552845528	0,088235294
Switzerland	52	2	27	0,038461538	0,519230769	0,074074074
Italy	24	1	16	0,041666667	0,666666667	0,0625
Australia	72	3	60	0,041666667	0,833333333	0,05
Germany	29	1	24	0,034482759	0,827586207	0,041666667
Japan	85	1	24	0,011764706	0,282352941	0,041666667
NewZealand	8	1	31	0,125	3,875	0,032258065
Netherlands	173	1	51	0,005780347	0,294797688	0,019607843

N.M : Number of mutations
O.M: Overall mutations
Mut.Uq/G: Mutation uniq / genome
O.M/G: Overall mutations/genome
Mut.uq/G: Mutation uniq /overall Mutations

Some Supplemental material could not be included within the manuscript due to their large size. they can be found directly using the link bellow.

`https://drive.google.com/drive/folders/1qx4eZquf3hemFFkCjlZi_LQSkys2g03S?usp=sharing`

2 Appendix Chapter 4

Table 6.3: Mutations with a non-synonymous effect found in the receptor-binding domain (RBD) of the spike protein

[] Gene	R.L	A.A	N.C	AA.C	Mutation type	Impact	G.A	N.M/ G.A	Mut freq/ G.A
S	C	T	c.989C>T	p.Pro330Leu	missense_variant	MODERATE	Europe	1	5,03145E-05
S	G	A	c.1016G>A	p.Gly339Asp	missense_variant	MODERATE	Europe	1	5,03145E-05
S	G	T	c.1030G>T	p.Ala344Ser	missense_variant	MODERATE	Asia	2	0,000947418
S	A	G	c.1060A>G	p.Asn354Asp	missense_variant	MODERATE	Asia	2	0,000947418
S	G	T	c.1068G>T	p.Lys356Asn	missense_variant	MODERATE	Europe	1	5,03145E-05
S	G	A	c.1076G>A	p.Ser359Asn	missense_variant	MODERATE	Asia	1	0,000473709
S	G	T	c.1090G>T	p.Asp364Tyr	missense_variant	MODERATE	Asia	1	0,000473709
S	G	T	c.1099G>T	p.Val367Phe	missense_variant	MODERATE	Asia	5	0,002368546
S	G	T	c.1099G>T	p.Val367Phe	missense_variant	MODERATE	Europe	11	0,000553459
S	G	T	c.1099G>T	p.Val367Phe	missense_variant	MODERATE	North America	2	0,00029304
S	T	C	c.1129T>C	p.Phe377Leu	missense_variant	MODERATE	Europe	3	0,000150943
S	C	T	c.1151C>T	p.Pro384Leu	missense_variant	MODERATE	Africa	1	0,004672897
S	A	G	c.1153A>G	p.Thr385Ala	missense_variant	MODERATE	North America	2	0,00029304
S	T	G	c.1195T>G	p.Ser399Ala	missense_variant	MODERATE	Asia	2	0,000947418
S	A	G	c.1204A>G	p.Ile402Val	missense_variant	MODERATE	Africa	1	0,004672897
S	T	G	c.1220T>G	p.Val407Gly	missense_variant	MODERATE	Asia	1	0,000473709
S	G	T	c.1223G>T	p.Arg408Ile	missense_variant	MODERATE	Europe	1	5,03145E-05
S	A	G	c.1228A>G	p.Ile410Val	missense_variant	MODERATE	Asia	1	0,000473709
S	A	G	c.1302A>G	p.Ile434Met	missense_variant	MODERATE	South America	1	0,002717391
S	G	T	c.1303G>T	p.Ala435Ser	missense_variant	MODERATE	Europe	1	5,03145E-05
S	C	A	c.1317C>A	p.Asn439Lys	missense_variant	MODERATE	Europe	80	0,004025157
S	T	G	c.1320T>G	p.Asn440Lys	missense_variant	MODERATE	Europe	1	5,03145E-05
S	G	T	c.1337G>T	p.Gly446Val	missense_variant	MODERATE	Oceania	3	0,001886792

Table 6.3: Mutations with a non-synonymous effect found in the receptor-binding domain (RBD) of the spike protein

[] Gene	R.L	A.A	N.C	AA.C	Mutation type	Impact	G.A	N.M/ G.A	Mut freq/ G.A
S	T	G	c.1350T>G	p.Asn450Lys	missense_variant	MODERATE	North America	1	0,00014652
S	C	A	c.1376C>A	p.Ser459Tyr	missense_variant	MODERATE	Africa	1	0,004672897
S	A	C	c.1379A>C	p.Asn460Thr	missense_variant	MODERATE	North America	1	0,00014652
S	G	A	c.1426G>A	p.Gly476Ser	missense_variant	MODERATE	North America	9	0,001318681
S	G	A	c.1430G>A	p.Ser477Asn	missense_variant	MODERATE	Europe	22	0,001106918
S	C	A	c.1431C>A	p.Ser477Arg	missense_variant	MODERATE	Europe	1	5,03145E-05
S	C	T	c.1433C>T	p.Thr478Ile	missense_variant	MODERATE	Europe	47	0,00236478
S	T	C	c.1448T>C	p.Val483Ala	missense_variant	MODERATE	North America	29	0,004249084
S	C	T	c.1472C>T	p.Pro491Leu	missense_variant	MODERATE	Asia	1	0,000473709
S	C	T	c.1472C>T	p.Pro491Leu	missense_variant	MODERATE	Oceania	1	0,000628931
S	G	T	c.1546G>T	p.Glu516*	stop_gained	HIGH	Asia	1	0,000473709
S	T	A	c.1557T>A	p.His519Gln	missense_variant	MODERATE	Asia	1	0,000473709
S	G	T	c.1558G>T	p.Ala520Ser	missense_variant	MODERATE	North America	8	0,001172161
S	C	G	c.1562C>G	p.Pro521Arg	missense_variant	MODERATE	Europe	2	0,000100629
S	G	T	c.1564G>T	p.Ala522Ser	missense_variant	MODERATE	Oceania	2	0,001257862
S	G	C	c.1587G>C	p.Lys529Asn	missense_variant	MODERATE	Asia	1	0,000473709

R.L: Reference Allele
N.C: Nucleotidic Change
A.A: Alternative Allele
AA.C: Amino acid change
G.A: Geographical Area
N.M/ G.A: Number of mutation per geographic area
Mu freq/ G.A: Mutation frequency per geographic area

3 Appendix Chapter 6

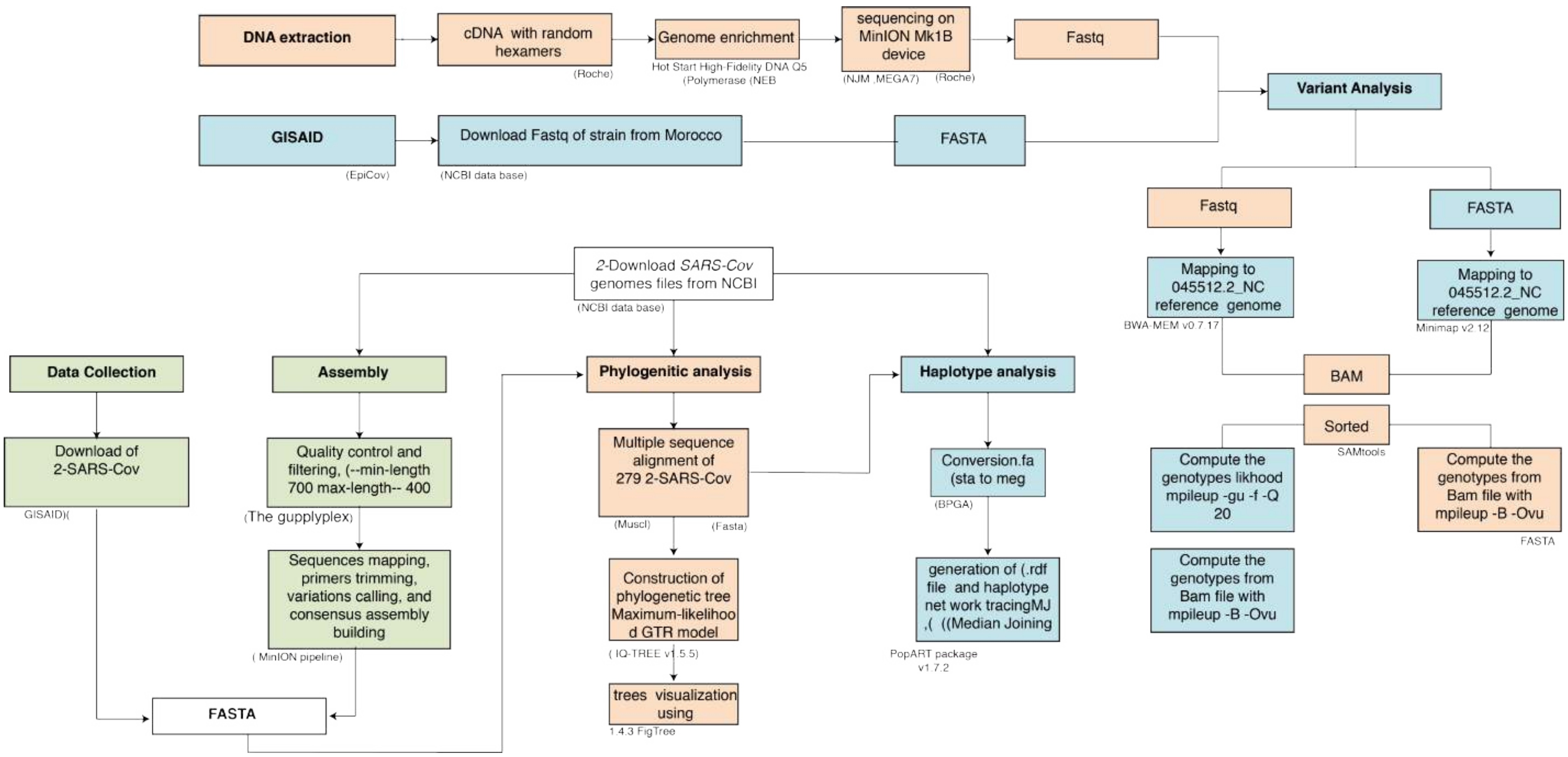
DNA extraction
cDNA with random hexamers
(Roche)
Genome enrichment
Hot Start High-Fidelity DNA Q5 (Polymerase (NEB
sequencing on MinION Mk1B device
(NJM .MEGA7)
(Roche)
Fastq
GISAID
(EpiCov)
Download Fastq of strain from Morocco
(NCBI data base)
FASTA
Variant Analysis
Fastq
Mapping to 045512.2_NC reference genome
BWA-MEM v0.7.17
FASTA
Mapping to 045512.2_NC reference genome
Minimap v2.12
BAM
Sorted
SAMtools
Compute the genotypes likhood mpileup -gu -f -Q 20
Compute the genotypes from Bam file with mpileup -B -Ovu
Compute the genotypes from Bam file with mpileup -B -Ovu
FASTA
2-Download *SARS-Cov* genomes files from NCBI
(NCBI data base)
Data Collection
Download of 2-SARS-Cov
GISAID)(
Assembly
Quality control and filtering, (--min-length 700 max-length-- 400
(The gupplyplex)
Sequences mapping, primers trimming, variations calling, and consensus assembly building
(MinION pipeline)
FASTA
Phylogenitic analysis
Multiple sequence alignment of 279 2-SARS-Cov
(Muscl)
(Fasta)
Construction of phylogenetic tree Maximum-likelihood GTR model
(IQ-TREE v1.5.5)
trees visualization using
1.4.3 FigTree
Haplotype analysis
Conversion.fa (sta to meg
(BPGA)
generation of (.rdf file and haplotype net work tracingMJ ,(((Median Joining
PopART package v1.7.2

Figure 6.1: Genomic analysis workflow for large pathogens data

4 Appendix Publications

1. **LAAMARTI, M.**, ALOUANE, T., ...
   IIBRAHIMI, A. (2020). Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geodistribution and a rich genetic variations of hotspots mutations. PLoS ONE 15(11): e0240345. https://doi.org/10.1371/journal.pone.0240345.

2. **LAAMARTI, M.**,, M. W. Chemao-Elfihri, ..., Azeddine Ibrahimi. Genome Sequences of Six SARS-CoV-2 Strains Isolated in Morocco, Obtained Using Oxford Nanopore MinION Technology. Microbiology Resource Announcements Aug 2020, 9 (32) e00767-20; DOI: 10.1128/MRA.00767-20

3. **LAAMARTI, M.**, ALOUANE, T., ... Azeddine Ibrahimi. Genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes: moving toward a universal vaccine for the "confined virus"?. Pathogens. doi: https://doi.org/10.1101/2020.06.20.163188

4. **LAAMARTI, M.**, ELMRIMAR N, ALOUANE T, EL OUNASS M, IBRAHIMI A. Comparative genomic analysis of *Mycobacterium tuberculosis* isolated from Morocco. Forthcoming.

5. **LAAMARTI, M.**, Amina M, Loubna T, Jamal Eddine E, Maltouf F.Genomic analysis of two **Bacillus safensis** gives insights into the features associated with their adaptation to desert environment and reveal potential plant growth promoting traits in the specie core genome. Genome Biology and Evolution (Under review). Forthcoming.

6. **LAAMARTI, M.** , ELMRIMAR N, ALOUANE T, EL OUNASS M, IBRAHIMI A. C. Draft genome sequences of two Mycobacterium tuerbculosis strains isolated from Morocco. Microbiology Resource Announcements Oct 2020.

7. **LAAMARTI, M.**,M W Chemao-Elfihri, Amina M, S Kartti, A Essabbar, T Alouane, L Temsamani, J Eljamali, M Ouadghiri, N El Hajjami, L Sbabou, L Belyamani, A, A Filali-Maltouf. (2020).Do the Moroccan SARS-CoV-2 genetic diversity hamper the use of the developed universal vaccines in Morocco?. Biorxiv. doi:https://doi.org/10.1101/2020.06.30.181123

8. Ezzaki, A., **LAAMARTI, M.**, Uwingabiye, J., & Sekhsokh, Y. (2018). CURRENT ANTIBIOTIC SUSCEPTIBILITY PROFILE OF ESCHERICHIA COLI STRAINS FROM COMMUNICTY ACQUIRED URINARY TRACT INFECTIONS IN MOROCCO.

9. Lahlou, L., El Mrimar, N., **LAAMARTI, M.**, ... & IBRAHIMI, A. (2017). Whole-genome shotgun sequences of three multidrug-resistant Mycobacterium tuberculosis strains isolated from Morocco. Genome Announc., 5(46), e01275-17.

10. Lahlou L, El Mrimar N, Alouane T, **LAAMARTI, M.** et al. Annotated Whole-Genome Shotgun Sequence of Multidrug-Resistant Mycobacterium tuberculosis MTB13_M Isolated from Morocco. Genome Announc.

11. Alouane, T., Uwingabiye, J., Lemnouer, A., Lahlou, L., **LAAMARTI, M.**, Kartti, S., ... & IBRAHIMI, A. (2017). First whole-genome sequences of two multidrug-resistant Acinetobacter baumannii strains isolated from a Moroccan hospital floor. Genome Announc.

12. Eljaoudi, R., Elomri, N., **LAAMARTI, M.**, Cherrah, Y...,
Ibrahimi, A. (2017). Antioxidants status in type 2 diabetic patients in Morocco. Turkish journal of medical sciences.

13. LM W Chemao-Elfihri, Amina M, **LAAMARTI, M.**, S Kartti, A Essabbar, T Alouane, L Temsamani, J Eljamali, M Ouadghiri, N El Hajjami, L Sbabou, L Belyamani, A, A Filali-Maltouf. (2020).Draft Genome Sequence of Stenotrophomonas maltophilia MDMC339, Isolated from Soil of Merzouga Desert in Morocco. Microbiol Resour Announc. 2020 Aug 6;9(32):e00634-20. doi: 10.1128/MRA.00634-20