الجمهورية الجزائرية الديمقراطية الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة امحمد بوقرة بومرداس
Université M'Hamed BOUGARA de Boumerdès

Faculté des Sciences
Département de Biologie

Mémoire de fin d'étude

En vue de l'obtention du diplôme de Master II
Filière : Biotechnologie
Spécialité : Biotechnologie et pathologies moléculaires

Thème

Molecular sequencing and phylogenic analysis of Algerian strains of SARS -CoV-2.

Réalisé par :

- **HADJKACI Toufik**

Soutenu devant le de Jury :

Présidente	AKMOUSSI-TOUMI Siham	MCB	UMBB
Examinateur	SADAOUI-SMADHI Nesrine	MCB	UMBB
Promotrice	AISSAT Faiza	MAA	UMBB
Co-promotrice	KHEMILI-TALBI Souad	Pr	UMBB

Année universitaire 2020/2021

Acknowlegment

First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-Magnificent; the Ever-Thankful, for His help and bless. I am totally sure that this work would have never become truth, without His guidance.

I would like to thank Ms. Khemili Talbi.S and Ms. Aissat.F for their consistent support and guidance during the running of this project.

Furthermore, I would like to thank the jurors for the time they gave for this work

Finally, many thanks to all participants that took part in the study and enabled this research to be possible.

Dedication

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents, Hacene and HENNACHE Yamina whose words of encouragement and push for tenacity ring in my ears.

My sisters and brothers who never left my side and are very special; I will always appreciate all they have done.

MY DEAREST CHABNI Kamilia, who leads me through the hardest times with light of hope and support.

Toufik

ABSTRACT

COVID-19 was quickly designated as a global pandemic by the World Health Organization as there have been about 98.0 million confirmed cases and about 2.0 million confirmed deaths, as of January 2021. In this study, we have gone through different characteristics of the virus, and the tools used on data analysis and representation, for a better understanding of sars-cov-2.

NGS showed a major role on identifying the nucleotide sequence of it ARN; the results were shared on the genomic platforms for scientific research purposes. Based on the data available on Gisaid, we carried out a phylogenic analysis of Algerian sars-cov2, by which we were able to diagram our data , and understand more about the origin of sars-cov-2 in Algeria.

RESUME

COVID-19 a été rapidement désigné comme une pandémie mondiale par l'Organisation mondiale de la santé, car il y avait eu environ 98,0 millions de cas confirmés et environ 2,0 millions de décès confirmés, en janvier 2021. Dans cette étude, nous avons examiné différentes caractéristiques du virus, et les outils utilisés sur l'analyse et la représentation des données, pour une meilleure compréhension du sras-cov-2.

Le NGS a montré un rôle majeur dans l'identification de la séquence nucléotidique de son ARN ; les résultats ont été partagés sur les plateformes génomiques à des fins de recherche scientifique. Sur la base des données disponibles sur Gisaid, nous avons réalisé une analyse phylogénique du sras-cov2 algérien, par laquelle nous avons pu schématiser nos données, et comprendre plus sur l'origine de sars-cov 2 en algérie.

TABLE OF CONTENTS

II. NEXT-GENERATION SEQUENCING (NGS) A Tool for SARS-CoV-2 DIAGNOSIS, AND REVEALS THE PROGRESSION OF COVID 19

III. SARS-COV-2 PHELOGENETIC ANALYSIS

LISTE OF TABLES **Page**

LISTE OF FIGURES **Page**

ABREVIANTION

2G: second generation

ARDS: acute respiratory distress syndrome

BAL: bronchoalveolaire lavage

Bp: base pair

BGI: Beijing Genomics Institute

BALF: bronchoalveolar lavage fluid

CT: computerized tomography

CLpro: chymotrypsin-like pro

cat: cathepsine

CD: Cluster of differentiation

cDNA: complementary Deoxyribonucleic Acid

CDC: Centers for Disease Control and Prevention

DNA: Deoxyribonucleic Acid

dNTP: deoxynucleoside triphosphate

Epicov: covid epidemiology

FM: Fitch Margoliash

GM-CSF: Granulocyte macrophage colony stimulating factor

G-csf: Granulocyte-Colony-Stimulating Factor

GB: gigabyte

IL: interleukine

ISFET: Ion Sensitive Field Effect Transistor

MRP: matrix representation with parsimony

MERS: middle east respiratory syndrome

MCP: Monocyte Chemoattractant Protein

MIP: macrophage inflammatory protein

MODS: multi-organ dysfunction

ME: minimum evolution

NGS: next generation sequencing

NSP: non structural proteins

Nt: nucleotid

ORF: open reading frame

Ppl: polypeptides

PLpro: papaine-like pro

qPCR: quantitative polymerase chain reaction

RNA: Ribonucleic acid

RT-PCR: reverse transcription polymerase chain reaction

R0: pronounced “R naught,” is a mathematical term that indicates how contagious an infectious disease is. It’s also referred to as the reproduction number

rRNA: ribosomal RNA

Sars: Severe acute respiratory syndrome

TMPS: Transmembrane protease,serine

Th: lymphocytes T helper

WHO: world health organization

WPGMA: weighted pair group method with arithmetic mean

GENERAL INTRODUCTION

In late 2019, the first cases of pneumonia of unknown etiology were identified in Wuhan, Hubei Province, People's Republic of China. Chinese authorities identified a new type of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which rapidly spread across the globe, causing the coronavirus disease 2019 (COVID-19) pandemic.

On 11 March the World Health Organization (WHO) declared the COVID-19 a pandemic. From the moment of the first cases to 23 September 2020, more than 31 million people were confirmed with the virus, and more than 971,000 deaths have occurred due to the disease.

Like other RNA viruses, SARS-CoV-2, while adapting to their new human hosts, is prone to genetic evolution with the development of mutations over time, resulting in mutant variants that may have different characteristics than its ancestral strains. Several variants of SARS-CoV-2 have been described during the course of this pandemic, among which only a few are considered variants of concern (VOCs) by the WHO, given their impact on global public health. Based on the recent epidemiological update by the WHO, as of June 22, 2021, four SARS-CoV-2 VOCs have been identified since the beginning of the pandemic:

- Alpha (B.1.1.7): first variant of concern described in the United Kingdom (UK) in late December 2020
- Beta (B.1.351): first reported in South Africa in December 2020
- Gamma(P.1): first reported in Brazil in early January 2021
- Delta (B.1.617.2): first reported in India in December 2020
  https://www.ncbi.nlm.nih.gov/books/NBK554776/

Next-generation sequencing (NGS) is a technology used by countless laboratories across the world for investigating the genetic makeup of all forms of living beings, but its utilization in infectious disease diagnostics is relatively scarce at the present moment. Information gleaned from NGS, whereby the pathogen's genome sequence is determined, yields a much greater trove of knowledge than the data produced by standard testing procedures, including information for the development of therapeutics and vaccines, the monitoring of changes in the virus as it circulates through the population, and deeper insights into patterns of transmission across time and geography (John and *al.*.2021).

Methods for phylogenetic tree construction are mostly based on a single gene in coronavirus genome or one artificial gene composed of the full-length genomic sequence. Notably, critical

limitations exist in these approaches. Gene selection is a major problem that phylogenetic methods based on a single gene need to tackle.

Compared with other phylogenetic analysis methods, the supertree method showed more resolution power for phylogenetic analysis of coronaviruses. In particular, the MRP pseudo-sequence supertree analysis firmly disputes bat coronavirus RaTG13 be the last common ancestor of SARS-CoV-2, which was implied by other phylogenetic tree analysis based on viral genome sequences. Furthermore, the discovery of evolution and mutation in SARS-CoV-2 was achieved by MRP pseudo-sequence supertree analysis. Taken together, the MRP pseudo-sequence supertree provided more information on the SARS-CoV-2 evolution inference relative to the normal phylogenetic tree based on full-length genomic sequences.

In this study, we gonna try to enlighten the different NGS methods already used to sequence the sars-cov-2 genome, and by using the available data on different platforms, concerning the Algerian sars-cov-2 sequences, we gonna try to make a phylogenetic analysis of the virus, and trace the origins of the algerian sequences.

CHAPTER ONE

Covid-19: towards a new pandemic

I. COVID-19: TOWARDS A NEW PANDEMIC

1. Generalities and definition

1.1 Generality

An epidemic of viral pneumonia of unknown ethology emerged in the city of Wuhan (Hubei province, China) in December 2019, around a live animal market. On January 9, 2020, the discovery of a new coronavirus (first called 2019-nCoV and then officially SARS Cov2, different from the SARS-CoV viruses (SARS epidemic in 2003) and MERS-CoV (epidemic evolving since 2012 in the Middle East) has been officially announced by the Chinese health authorities and the World Health Organization (WHO) This new virus is the causative agent of this new infectious respiratory disease called CoVID-19 (for CoronaVirus Disease).

The virus seems to be usually present rather in bats. In connection with interspecies encounters in this Wuhan market, there has very probably been transfer of the virus to other animal species (snake initially whose involvement has been refuted, or more likely the pangolin in the latest phylogenetic projections). It was during this passage between species that there was probably a mutation allowing the virus to infect humans (Institut Pasteur, 2021).

1.2 Coronavirus definition

Coronaviruses are a group of related RNA viruses that cause diseases in mammals and birds. In humans and birds, they cause respiratory tract infections that can range from mild to lethal. Mild illnesses in humans include some cases of the common cold (which is also caused by other viruses, predominantly rhinoviruses), while more lethal varieties can cause SARS, MERS, and COVID-19. In cows and pigs they cause diarrhea, while in mice they cause hepatitis and encephalomyelitis.

The name “coronavirus”, from Latin meaning “crown virus”, is due to the appearance of virions under an electron microscope, with a fringe of large bulbous projections that resemble a solar corona.

Bats and birds, as warm-blooded flyers, have been the ideal hosts for coronaviruses ensuring the evolution and spread of the coronavirus. Coronaviruses are normally specific to an animal taxon as host, mammals or birds depending on their species; but they can sometimes change host as a result of a mutation.

More recently, three types of coronavirus have been identified which are responsible for serious pneumonia:

1- SARS-CoV, a pathogen of severe acute respiratory syndrome (SARS) in 2002-2004.

2- MERS-CoV, that of Middle East respiratory syndrome from 2012.

3- SARS-CoV-2, that of the 2019 coronavirus disease (Covid-19) which appeared in China in 2019 and responsible for a severe pandemic in 2020-2021 (Van Doremalen and *al*., 2020).

1.3 General characteristics

1.3.1 Description

- Size: 125nm (larger than influenza, SARS and MERS viruses).
- Single-stranded RNA viruses 29,903 base pairs.
- Survival in open air: 3 hours maximum after aerosolization (droplets fall in maximum 20 to 30 minutes), up to 72 hours on cardboard and stainless steel but with a significant reduction in viral load.

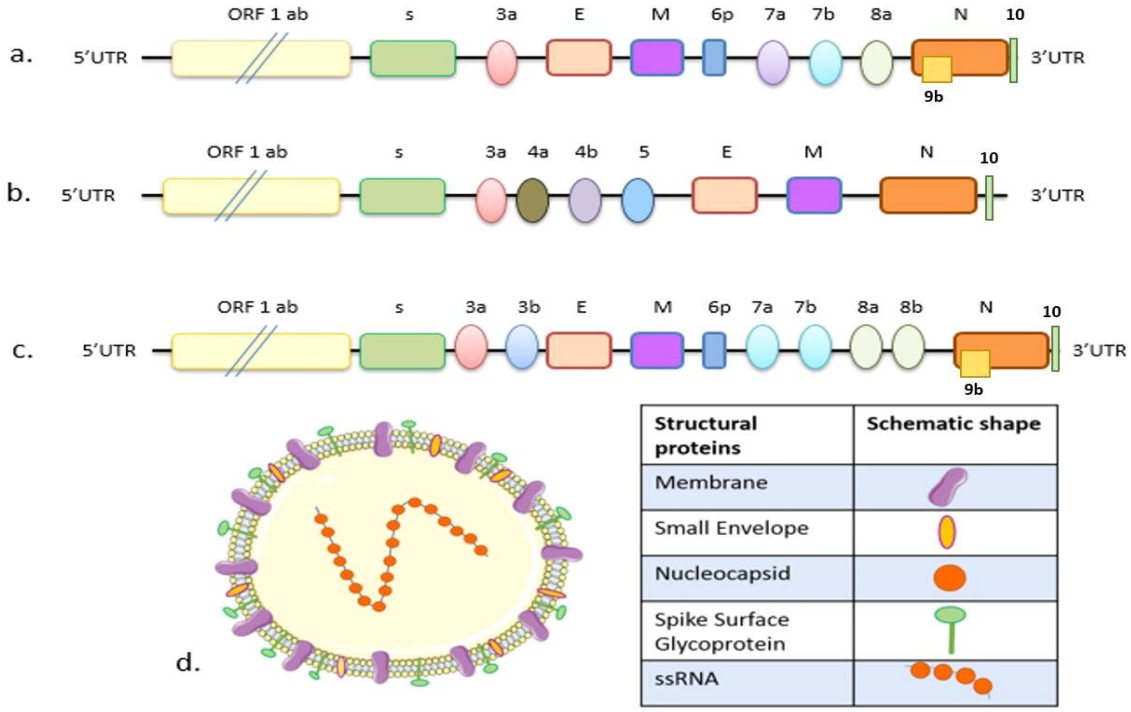

Figure 01. The schematic genomic structure of Coronavirus (Mohamadian and *al*., 2021)

1.3.2 Classification

- Type: Virus
- Realm: Riboviria

- Order: Nidovirales
- Suborder: Cornidovirineae
- Family: Coronaviridae
- Subfamily: Orthocoronavirinae
- Genus: Betacoronavirus
- Subgenus: Sarbecovirus (like SARS-CoV-1)

1.3.3 Viral shedding:

- LBA: 93% -> the most efficient
- Sputum: 72%
- Blood: 1% of samples; in favor of viremia in some cases. But no viral replication, so no risk (no indication for additional precautions).
- Stools: 29% (seems to occur more at the end of the illness and in patients in the process of recovery).
- Urine: none of the samples (Wang and *al.*, 2020 ; Tariq and *al.*, 2020).

1.3.4 TRANSMISSION – EPIDEMIOLOGY:

- R0> 2 or even 3 (WHO estimates between 1.4 and 2.5; remains highly debated). There is certainly a disparity between patients regarding their individual R0 (some individuals are "super" contaminants).
- Short inter-generational interval: 4.5 days
- Attack rate> 10% (significantly higher than that of seasonal influenza in a naive population).
- Average length of hospitalization: 11 days ± 4 days (15 days for severe forms)
- Case fatality rate estimated at less than 1% under optimal management conditions and in the absence of comorbidities. (Province of HUB EI 2.9%, Iran 8.7%) (Bidar and *al.*, 2020).

2. Primary reservoirs and hosts of coronaviruses

The source of origination and transmission are important to be determined in order to develop preventive strategies to contain the infection. In the case of SARS-CoV, the researchers initially focused on raccoon dogs and palm civets as a key reservoir of infection.

However, only the samples isolated from the civets at the food market showed positive results for viral RNA detection, suggesting that the civet palm might be secondary hosts.

In 2001 the samples were isolated from the healthy persons of Hongkong and the molecular assessment showed 2.5% frequency rate of antibodies against SARS-coronavirus. These indications suggested that SARS-coronavirus may be circulating in humans before causing the outbreak in 2003. Later on, Rhinolophus bats were also found to have anti-SARS-CoV antibodies suggesting the bats as a source of viral replication. The Middle East respiratory syndrome (MERS) coronavirus first emerged in 2012 in Saudi Arabia.

MERS-coronavirus also pertains to beta-coronavirus and having camels as a zoonotic source or primary host. In a recent study, MERS-coronavirus was also detected in Pipistrellus and Perimyotisbats, proffering that bats are the key host and transmitting medium of the virus. Initially, a group of researchers suggested snakes be the possible host, however, after genomic similarity findings of novel coronavirus with SARS-like bat viruses supported the statement that not snakes but only bats could be the key reservoirs. Further analysis of homologous recombination revealed that receptor binding spike glycoprotein of novel coronavirus is developed from a SARS-CoV (CoVZXC21or CoVZC45) and a yet unknown Beta-CoV.

Nonetheless, to eradicate the virus, more work is required to be done in the aspects of the identification of the intermediate zoonotic source that caused the transmission of the virus to human.

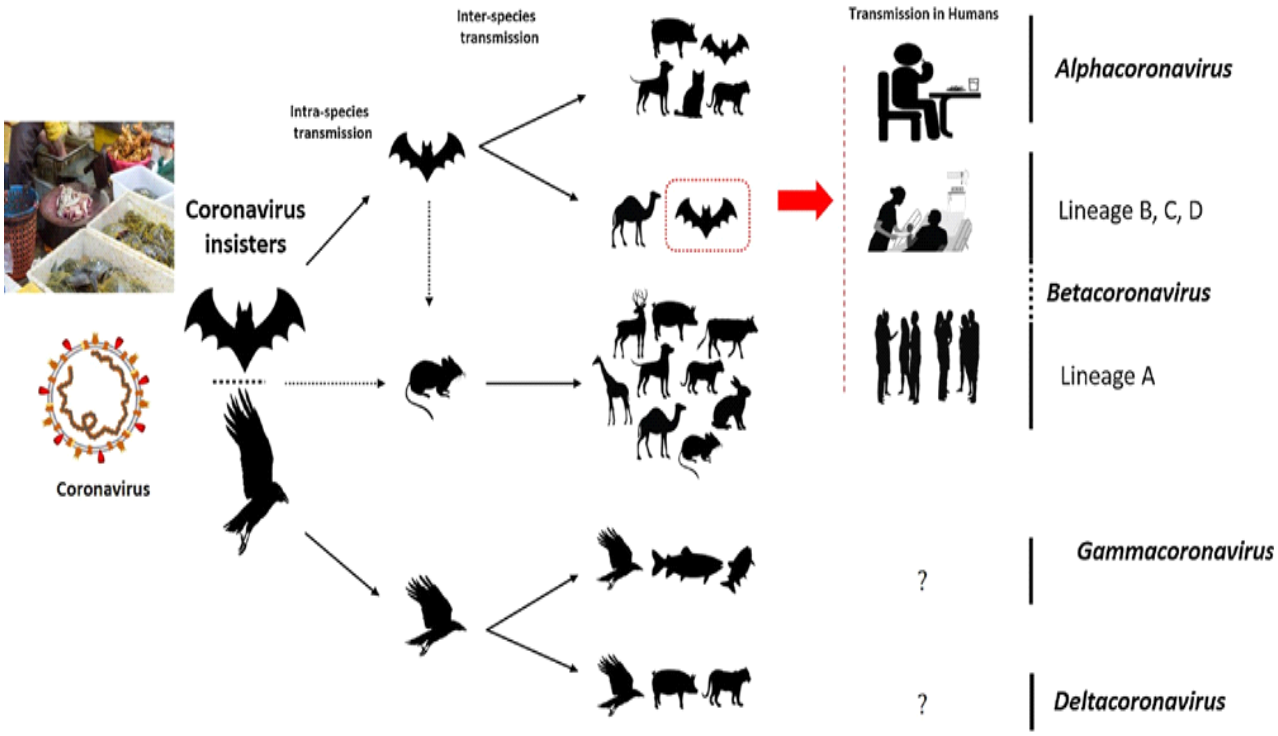

Figure 02. The key reservoirs and mode of transmission of coronaviruses (suspected reservoirs of SARS-CoV-2 are red encircled) (Shereen and *al*., 2020).

As seen in the figure 02, only a and b coronaviruses have the ability to infect humans, the consumption of infected animal as a source of food is the major cause of animal to human transmission of the virus and due to close contact with an infected person, the virus is further transmitted to healthy persons. Dotted black arrow shows the possibility of viral transfer from bat whereas the solid black arrow represents the confirmed transfer.

3. SARS-CoV-2 Virus structure and integration

SARS-CoV-2 belongs to the Beta coronavirus genus and is a member of the Corona virinae family. The virus particles are spherical or pleomorphic in shape, with a diameter of about 60–140 nm. Coronaviruses have one of the largest single-strand RNA genomes with 27–32 kilobases (kb). Some of the coronaviruses encode for the hemagglutinin-esteraseprotein, 3a/b protein,and 4a/b protein on their surface. The genome organization of SARS-CoV-2 is similar to other coronaviruses, which is composed of mainly the open reading frames (ORFs). Roughly 67% of the genome encodes by the ORF1a/b and it encodes for 16 non-structural polyproteins (nsp1-16), while the remaining 33% encodes for accessory proteins and structural proteins. ORF1a and ORF1b contain a frameshift which produces two polypeptides, pp1a and pp1ab. Papain-like protease (PLpro) or chymotrypsin-like protease (3CLpro), process these two polypeptides into 16 nsps. SARS-CoV-2 encodes for at least four major structural proteins that includes spike protein (S), membrane protein (M), an envelope protein (E), and nucleocapsid protein (N). These structural proteins are encoded by S, M, E, N genes at ORFs 10 and 11 on the one-third of the genome near the 30-end (figure 03).

These mature structural proteins are responsible for viral maintenance and replication. Most of the probes and primers used to detect the SARS-CoV-2 are constructed against the genetic targets of ORF1ab and the N gene region (Chilamakuri and Agarwal, 2021).

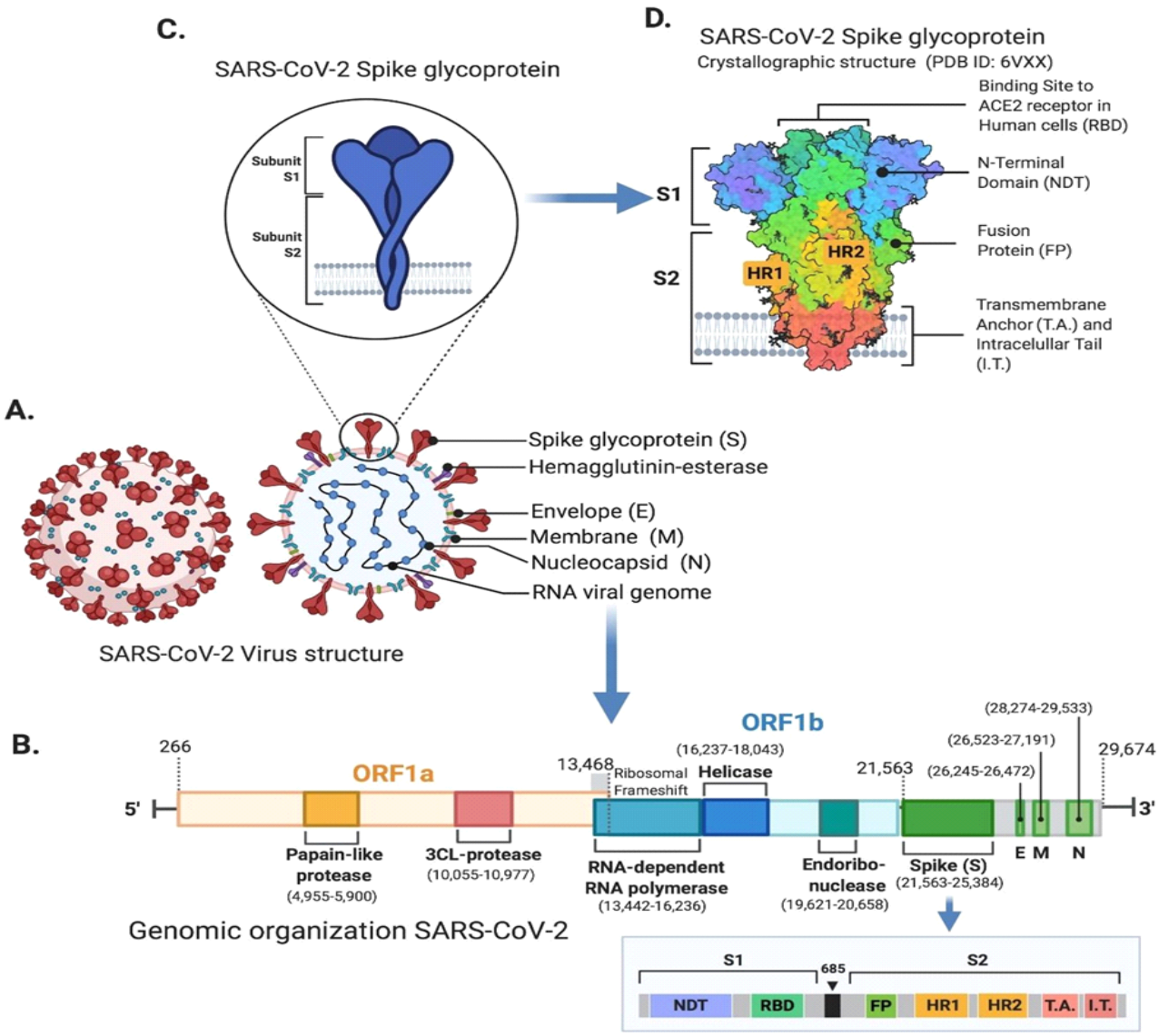

Figure 03. Structure and genomic organization of SARS-CoV-2 (Chilamakuri and Agarwal, 2021).

The figure 03 is a Schematic representation of SARS-CoV-2 virus structure and the positions of spike glycoprotein, hemagglutinin-esterase, envelope, membrane, nucleocapsid, and RNA viral genome. (**B**) Genomic organization of SARS-CoV-2 representing ORF1a, ORF1B which encode for nonstructural proteins such as papain-like protease, 3CL-protease, RNA-dependent RNA polymerase, helicase, and endoribonuclease. Genes coding for spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins are also displayed. Ribosomal frameshift location between ORF1 and ORF2 is shown at the junction of ORF1/2. Genomic positions are shown with dashed lines followed by nucleotide position number in RNA viral genome. The box highlights the genomic organization of spike (S) gene showing distinct S1 and S2 subunits coding segments. (**C**) Schematic magnified representation of SARS-CoV-2 spike glycoprotein showing S1 and S2 subunits. (**D**) Crystallographic structure of SARS-CoV-2 spike glycoprotein adapted from PDB ID:6VXX. Receptor binding domain (RBD) representing ACE2 receptor

binding site in human cells, N-terminal domain (NTD), fusion protein (FP), transmembrane anchor (T.A.), and intracellular tail (I.T.) protein domains are displayed (Chilamakuri and Agarwal, 2021).

Once the virus enters into a host cell, the synthesis of structural and accessory proteins begins with transcription and translation processes. The synthesis of the new viral RNA genome occurs with the help of RNA-dependent RNA polymerase, which utilizes the negative stand template. The binding affinity of SARS-CoV-2 for the angiotensin-converting enzyme 2 (ACE2) receptor ishigherth another SARSs, which in turn facilitates the rapid transmission of SARS-CoV-2.

The M protein is the most abundant structural glycoprotein and is responsible for the transport of nutrients across the cell membrane while giving shape to the virus particle. The S or spike protein is a type I membrane glycoprotein which constitutes virus peplomers. The N protein aids in binding the viral RNA genome while maintaining RNA stability. The E protein plays an important role in viral release as well as assembly during pathogenesis. The analysis of the whole genome sequence of SARS-CoV-2 shows that it shares 85-95% sequence similarity with SARS-CoV, indicating that SARS-CoV-2 is more compatible with SARS-CoV (Chilamakuri and Agarwal, 2021).

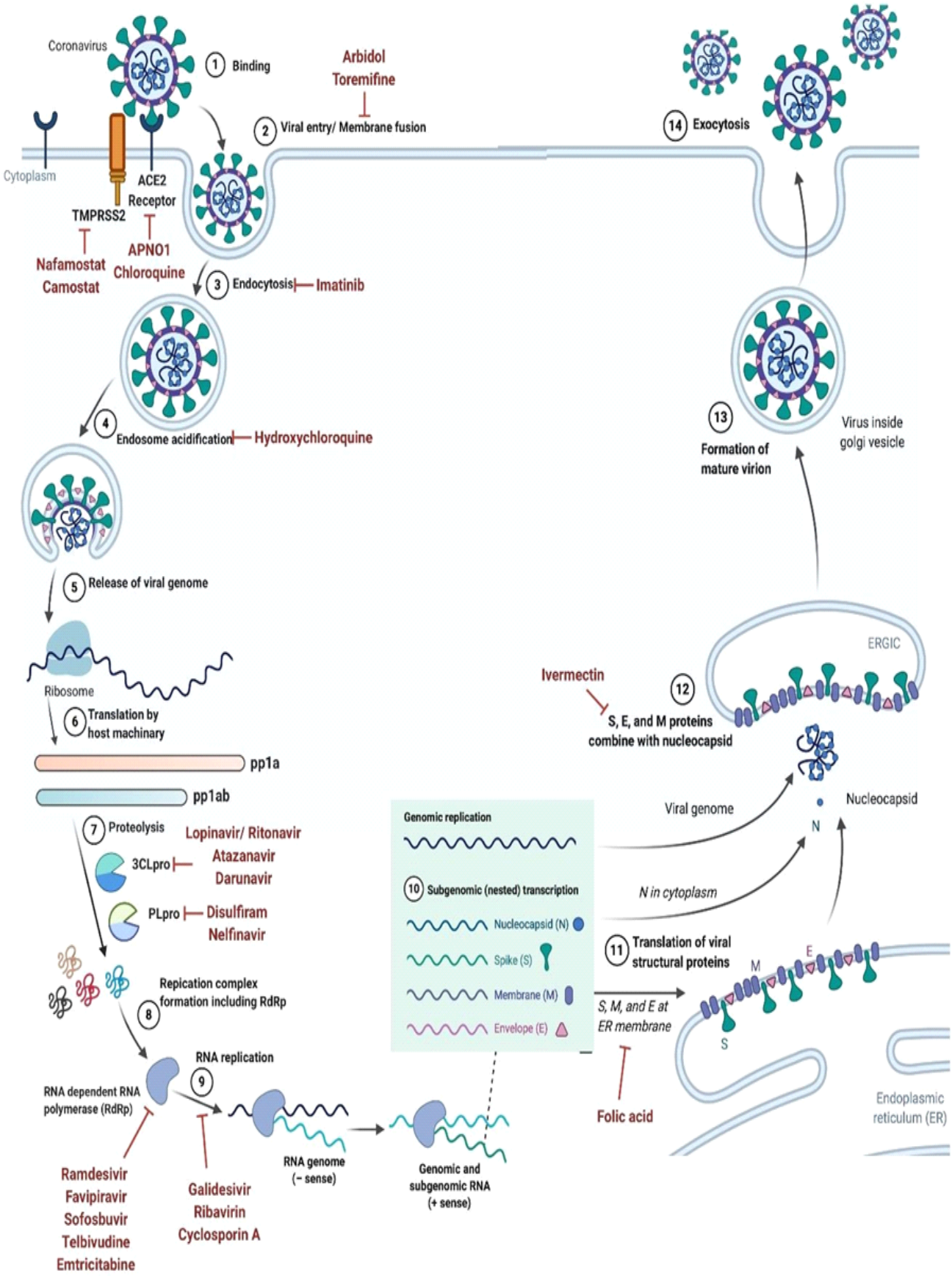

Figure 04. Schematic representation of SARS-CoV-2 virus life cycle (Chilamakuri and Agarwal, 2021).

4. SARS- CoV-2 Virus Receptor Mechanism

As mentioned above, spike (S) protein, which is located on the surface of SARS- CoV-2 is vital for infection and pathogenesis. The entry of SARS-CoV-2 into a host cell is mediated by the S protein, which ultimately gives coronaviruses a crown-like appearance as they form spikes on their surface .The S protein consists of three subunits, the ectodomain, a single-pass transmembrane anchor, and an intracellular C-terminaltail. The ectodomain can be further divided into a receptor-binding S1 subunit and a membrane-fusion S2 subunit. The SARS-CoV-2 virus enters the host cell through the interaction of the receptor-binding S1 subunit with the ACE2 receptor on the host cell surface. Human ACE2 receptors are expressed in almost all tissues, and they are most abundant in the lungs, kidneys, brain stem, adipose tissue, heart, vasculature, stomach, liver, as well as the nasal and oral mucosa. The S2 subunit fuses the host and viral membranes, while facilitating the entry of the viral genome into host cells. This process requires S protein priming by host cell proteases, which leads to S protein cleavage at the S1–S2 boundary. Recent reports have showed thatSARS-CoV-2 uses ACE2 forentry, while utilizing the transmembrane protease,serine2 (TMPRSS2), and endosomal cysteine proteases cathepsin B and L (CatB/L), for S protein priming . Going forward, knowledge about the receptor recognition and interaction mechanisms will be critical in identifying effective therapeutic targets (Chilamakuri and Agarwal, 2021).

5. Genomic variations in SARS- CoV-2

The genome of the SARS-CoV-2 has been reported over 80% identical to the previous human coronavirus (SARS-like bat CoV). The Structural proteins are encoded by the four structural genes, including spike (S), envelope (E), membrane (M) and nucleocapsid (N) genes. The orf1ab is the largest gene in SARS-CoV-2 which encodes the pp1ab protein and 15 nsps. The orf1a gene encodes for pp1a protein which also contains 10 nsps.

According to the evolutionary tree, SARS-CoV-2 lies close to the group of SARS-coronaviruses. Recent studies have indicated notable variations in SARS-CoV and SARS-CoV-2 such as the absence of 8a protein and fluctuation in the number of amino acids in 8b and 3c protein in SARS-CoV-2. It is also reported that Spike glycoprotein of the Wuhan coronavirusis modified via homologous recombination. The spike glycoprotein of SARS-CoV-2 is the mixture of bat SARS-CoV and a not known Beta-CoV. In a fluorescent study, it was confirmed that the SARS-CoV-2 also uses the same ACE2 (angiotensin-converting enzyme 2) cell receptor and mechanism for the entry to host cell which is previously used by the SARS-CoV (Shereen and *al*., 2020).

The single N501T mutation in SARS-CoV-2's Spike protein may have significantly enhanced its binding affinity for ACE2.

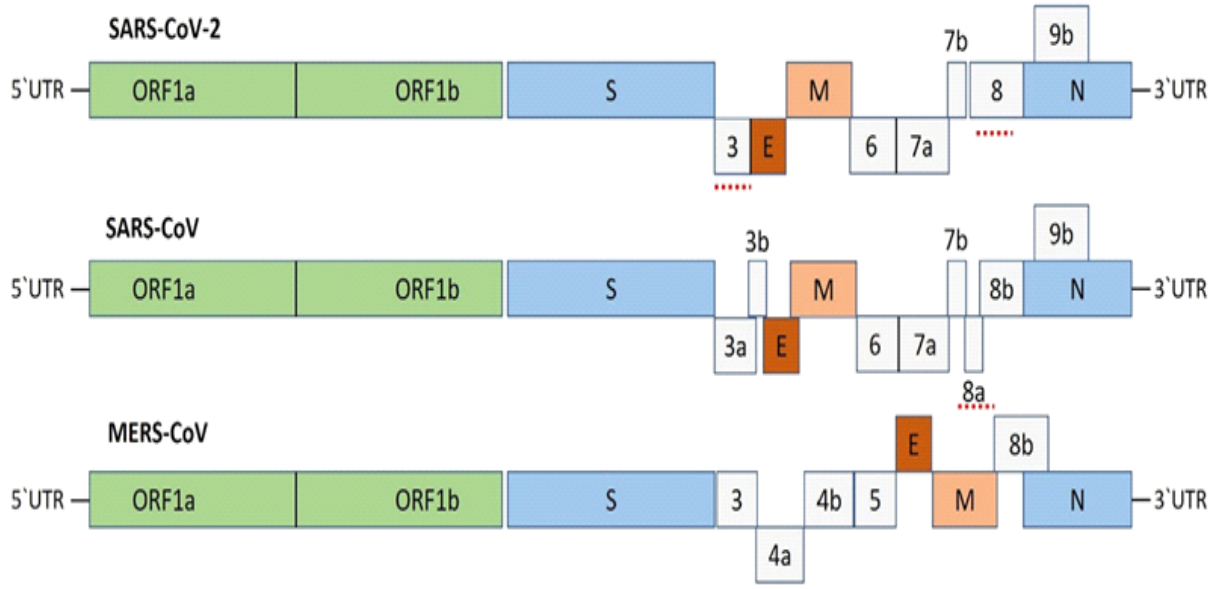

Figure 05. Betacoronaviruses genome organization (Shereen and *al*., 2020)

The Betacoronavirus for human (SARS-CoV-2, SARS-CoV and MERS-CoV) genome comprises of the 5′-untranslated region (5′-UTR), open reading frame (orf) 1a/b (green box) encoding non-structural proteins (nsp) for replication, structural proteins including spike (blue box), envelop (maroon box), membrane (pink box), and nucleocapsid (cyan box) proteins, accessory proteins (light gray boxes) such as orf 3, 6, 7a, 7b, 8 and 9b in the SARS-CoV-2 genome, and the 3′-untranslated region (3′-UTR). The doted underlined in red are the protein which shows key variation between SARS-CoV-2 and SARS-CoV. The length of nsps and orfs are not drawn in scale (Shereen and *al*., 2020)

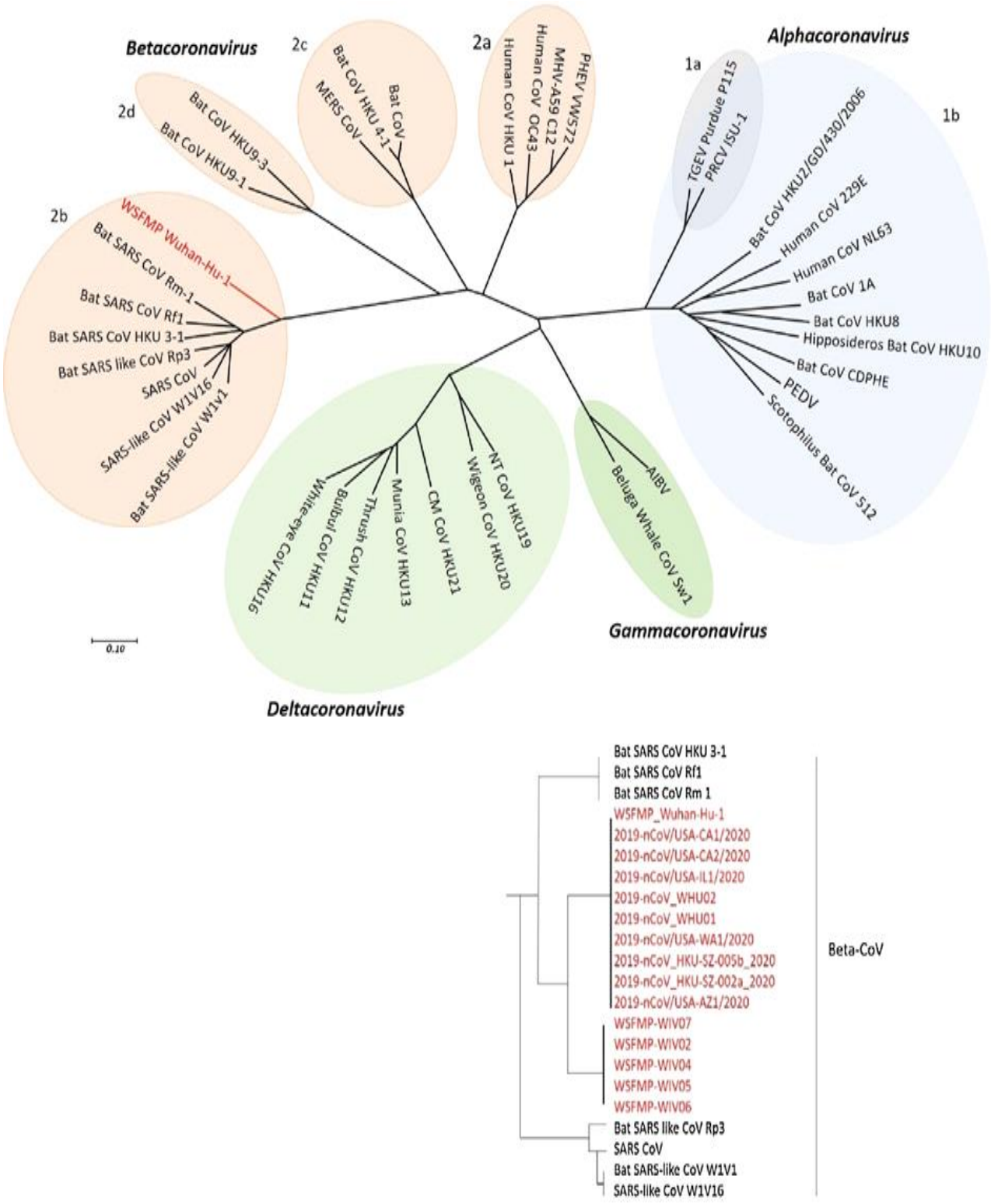

Figure 06. Phylogenetic tree of coronaviruses (content in red is the latest addition of newly emerged SARS-CoV-2 and WSFMP Wuhan-Hu-1 is used as a reference in the tree) (Shereen and *al*., 2020)

As we see at the (Figure06) phylogenetic tree showing the relationship of Wuhan-Hu-1 (denoted as red) to selected coronavirus is based on nucleotide sequences of the complete genome. The viruses are grouped into four generation (prototype shown): Alpha-coronavirus

(sky blue), Beta-coronavirus (pink), Gamma-coronavirus (green) and Delta-coronavirus (light blue). Subgroup clusters are labeled as 1a and 1b for the Alpha-coronavirus and 2a, 2b, 2c, and 2d for the Betacoronavirus. This tree is based on the published trees of Coronavirinae and reconstructed with sequences of the complete RNA- dependent RNA polymerase- coding region of the representative novel coronaviruses (maximum likelihood method using MEGA 7.2 software). severe acute respiratory syndrome coronavirus (SARS- CoV); SARS- related coronavirus (SARSr- CoV); the Middle East respiratory syndrome coronavirus (MERS- CoV); porcine enteric diarrhea virus (PEDV); Wuhan seafood market pneumonia (Wuhan-Hu-1). Bat CoV RaTG13 Showed high sequence identity to SARS-CoV-2.

6. Sars-cov-2 pathophysiology

6.1 Disease pathophysiology

Although much has been discovered regarding the transmission and presentation, less is known about the pathophysiology of COVID-19 (Parasher, 2020).

An overview of the disease pathophysiology has been shown in figure below (figure 07)

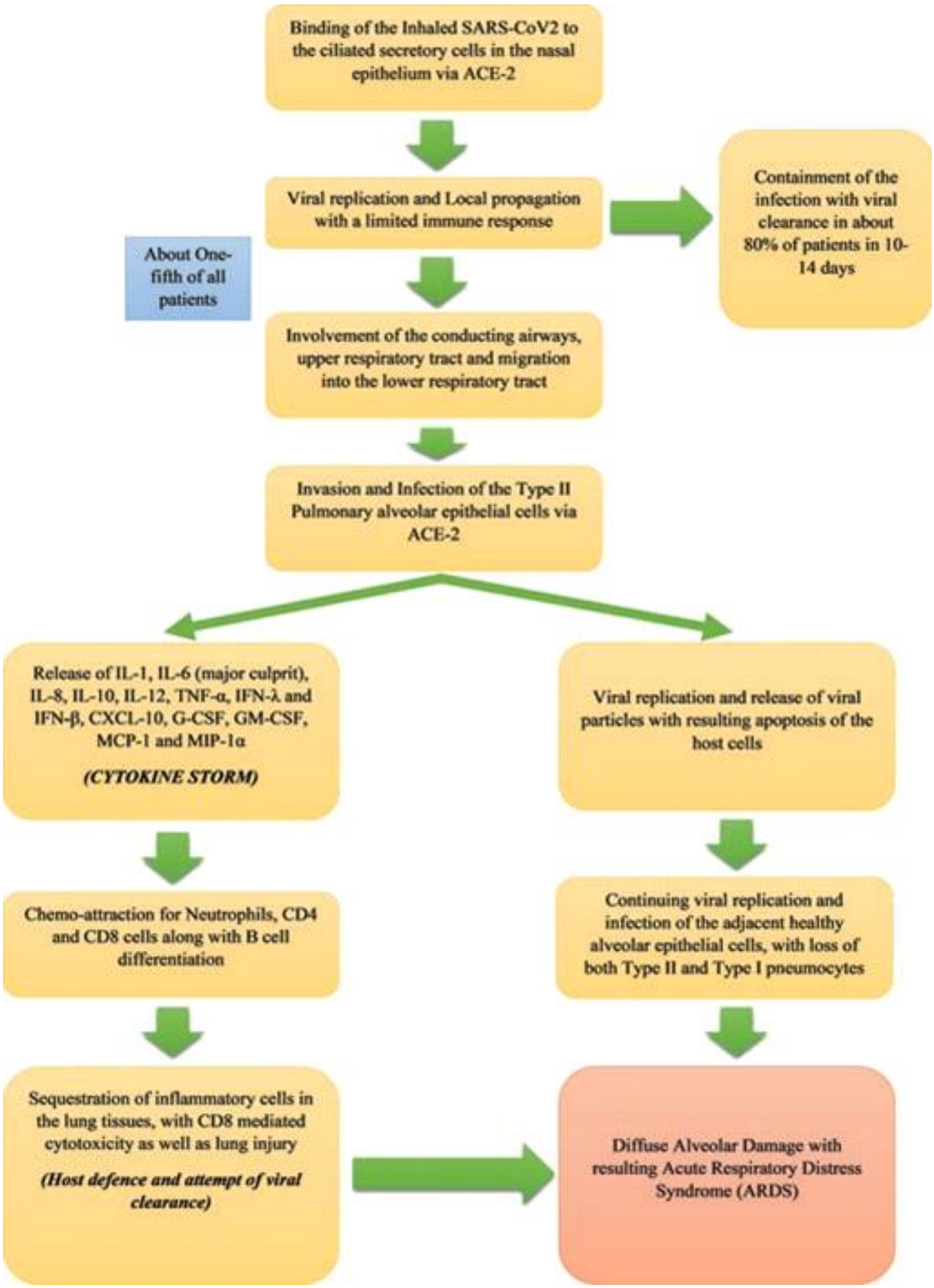

Figure 07. Pathophysiology of COVID-19 (Parasher, 2020).

6.2 Asymptomatic phase

The SARS-CoV-2 which is received via respiratory aerosols binds to the nasal epithelial cells in the upper respiratory tract. The main host receptor for viral entry into cells is the ACE-2, which is seen to be highly expressed in adult nasal epithelial cells.24 25 The virus undergoes local replication and propagation, along with the infection of ciliated cells in the conducting airways.26 This stage lasts a couple of days and the immune response generated during this phase is a limited one. In spite of having a low viral load at this time, the individuals are highly infectious, and the virus can be detected via nasal swab testing (Parasher, 2020).

6.3 Invasion and infection of the upper respiratory tract:

In this stage, there is migration of the virus from the nasal epithe- lium to the upper respiratory tract via the conducting airways. Due to the involvement of the upper airways, the disease manifests with symptoms of fever, malaise and dry cough. There is a greater immune response during this phase involving the release of C-X-C motif chemokine ligand 10 (CXCL-10) and interferons (IFN-βandIFN-λ) from the virus-infected cells.27 The majority of patients do not progress beyond this phase as the mounted immune response is sufficient to contain the spread of infection (Parasher, 2020).

6.4 Involvement of the lower respiratory tract and progression to acute respiratory distress syndrome (ARDS)

About one-fifth of all infected patients progress to this stage of disease and develop severe symptoms. The virus invades and enters the type 2 alveolar epithelial cells via the host receptor ACE-2 and starts to undergo replication to produce more viral Nucleocapsids. The virus-laden pneumocytes now release many different cytokines and inflammatory markers such as interleukins (IL-1, IL-6, IL-8, IL-120 and IL-12), tumour necrosis factor-α (TNF-α), IFN-λ and IFN-β, CXCL10, monocyte chemoattractant protein-1 (MCP-1) and macrophage inflammatory protein-1α (MIP-1α). This ‘cytokine storm’ acts as a chemo-attractant for neutrophils, CD4 helper T cells and CD8 cytotoxic T cells, which then begin to get sequestered in the lung tissue. These cells are responsible for fighting off the virus, but in doing so are responsible for the subsequent inflammation and lung injury. The host cell undergoes apoptosis with the release of new viral particles, whichthen infect the adjacent type 2 alveolar epithelial cells in the same manner. Due to the persistent injury caused by the sequestered inflammatory cells and viral replication leading to loss of both type 1 and type 2 pneumocytes, there is diffuse alveolar damage eventually culminating in an acute respiratory distress syndrome (Parasher, 2020).

7. Potential explanation for the difference between children and adults in COVID-19

The first possibility is that the expression level of ACE2 may differ between adults and children. A previous study showed that ACE2 was more abundantly expressed on well-differentiated ciliated epithelial cells. Human lung and epithelial cells continue to develop following the birth. ACE2 expression may be lower in pediatric population. From the lung gene expression analysis portal, ACE2 expression in mice increased around at birth. Its expression reduced till around P10, then increased. Because infants were susceptible to severe disease among children, this pattern may be in line with patients' clinical picture. In addition, gender may also affect ACE2 expression. ACE2 gene is located on the X-chromosome. Circulating ACE2 levels are higher in men than in women. This may be in part responsible for the difference in severity and mortality between men and women both in the adult and the pediatric population.

The second possibility is that children have a qualitatively different response to the SARS-CoV-2 virus to adults. With ageing, continuous antigen stimulation and thymic involution lead to a shift in T cell subset distribution from naïve T cells to central memory T cells, effector T cells and effector memory T cells. This process is accompanied by the loss of expression of co-stimulatory molecules such as CD27 and CD28, with increased susceptibility to infections. Whether the appearance of pathological T cells in adult patients with severe COVID-19 diseases is due to the compensation for this fundamental ageing process or not is unclear. At the early stage after birth, CD4+ T cells are impaired in production of Th1 associated pro-inflammatory cytokines and skewed toward Th2. CD8+ T cells reduced expression of cytotoxic and inflammatory mediators. Less killing ability by T cells at early stage after birth may explain susceptibility to SARS-CoV-2 in infants. The study comparing aged and young macaques infected with SARS-CoV showed that aged macaques had more robust pro-inflammatory responses with worse lung pathology. A similar result was reported using aged and young mice infected with SARS-CoV. Severe COVID-19 infection is characterized by a massive pro-inflammatory response or cytokine storm that results in ARDS and multi-organ dysfunction (MODS). It has been also suggested that inflammatory responses in adults and children are much different. Ageing is associated with increasing pro-inflammatory cytokines that govern neutrophil functions and have been correlated with the severity of ARDS. So far there is no animal model for SARS-CoV-2, but we expect to see a preclinical model in the future.

The third possibility is that the simultaneous presence of other viruses in the mucosa lungs and airways, common in young children, can let SARS-CoV-2 virus compete with them and limit its growth. At this point, we do not have study testing various viruses along with SARS-CoV-2 to determine this possibility.

8. Diagnosis and imaging

8.1 Molecular tests (RT-PCR)

Samples are collected from the upper respiratory tract via naso-pharyngeal and oropharyngeal swabs and from the lower respira- tory tract via expectorated sputum and bronchoalveolar lavage (only for mechanically ventilated patients). After being stored at 4°C, the samples are sent to the laboratory where amplification of the viral genetic material is done through a reverse-transcription process. This involves the synthesis of a double-stranded DNA molecule from the existing viral RNA by either reverse- transcription PCR (RT-PCR) or a real-time RT-PCR. Finally, the conserved portions of the SARS-CoV-2 genetic code are identified on the amplified genetic material.

The test is recommended to be repeated for verification in cases of a positive test and again to confirm viral clearance in COVID- 19 positive cases. The sensitivity of these tests is not very high, that is, approximately 53.3% of COVID-19-confirmed patients had positive oropharyngeal swabs, and about 71% of patients came out to be RT-PCR positive with sputum samples. The RT-PCR results usually show positivity after 2–8 days (Parasher, 2020).

8.2. Serology

Till date, no effective antibody test has been developed. A center for disease control and prevention (CDC) research on a test developed by the US Vaccine Research Centre at the National Institutes of Health is on-going, which seems to have a specificity higher than 99% with a sensitivity of 96%.

8.3 Blood tests

- A normal or decreased white blood cell count (and lymphopenia) can be observed in many cases, which is also; considered to be indicative of a worse prognosis.

- Increased levels of lactate dehydrogenase, C reactive protein, creatine kinase (CK MB and CK MM), aspartate aminotransferase and alanine amino-transferase can be seen.

- Increased D-dimer levels and an elevated neutrophil-tolymphocyte ratio are seen in some patients.

- Coagulation abnormalities can be observed in severe cases, as indicated by increase in prothrombin time and international normalised ratio (Parasher, 2020).

8.4 Chest X-ray

Chest X-ray is usually inconclusive in the early stages of the disease and might not show any significant changes. As the infection progresses, bilateral multifocal alveolar opacities are observed, which may also be associated with pleural effusion (Parasher, 2020).

8.5 Computerized tomography:

High-resolution CT (HRCT) is extremely sensitive and the method of choice for diagnosing COVID-19 pneumonia, even in initial stages of the illness. The most commonly seen features are multifocal bilateral ‘ground-glass’ areas associated with con- solidation and a patchy peripheral distribution, with greater involvement of the lower lobes. A ‘reversed halo sign’ is also seen in some patients, which is identified as a focal area ofpatchy opacities surrounded by a peripheral ring with consolidation. Other findings include pleural effusion, cavitation, calcification, and lymphadenopathy (Parasher, 2020)

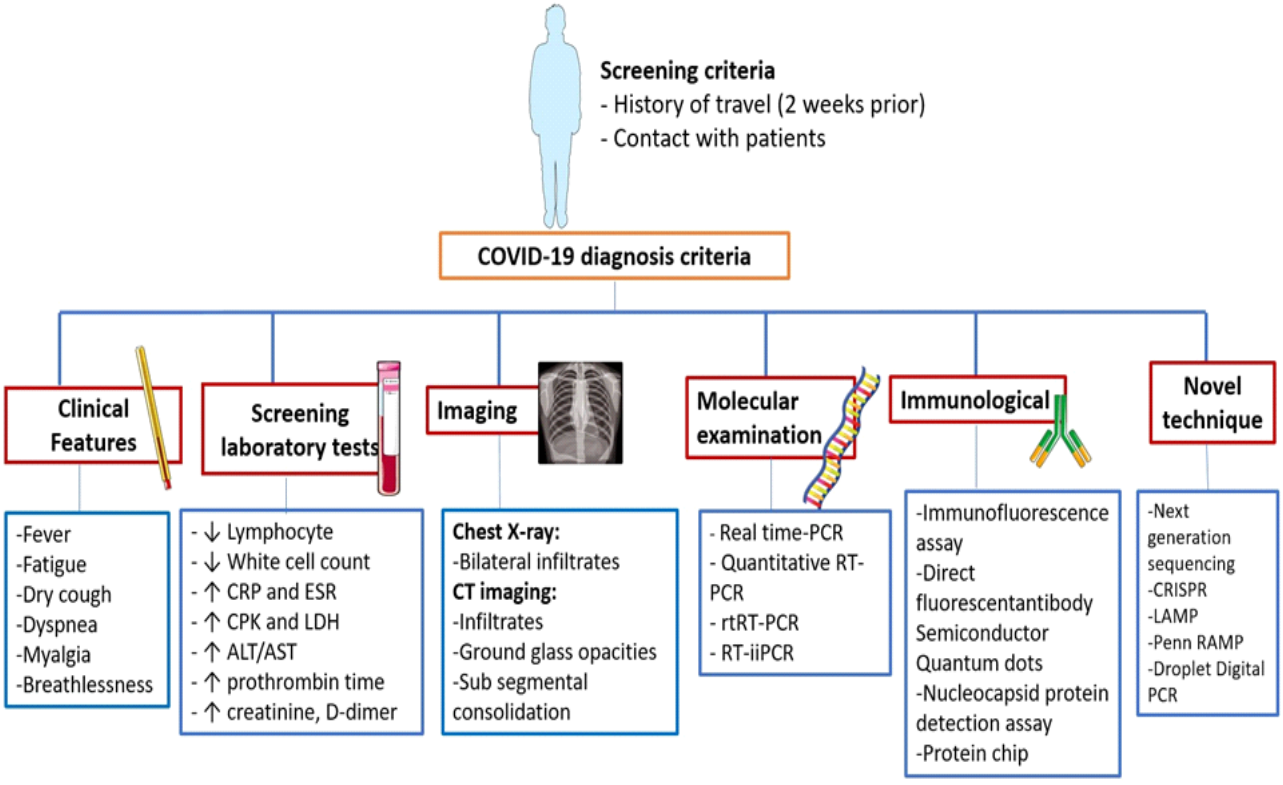

Figure08. Diagnostic protocol been recommended for COVID-19 (Mohamadian and *al*., 2021)

CHAPTER TWO

Next-Generation Sequencing (NGS) A Tool for SARS-CoV-2 Diagnosis, and reveals the progression of covid-19

“With the help of NGS, we can enhance our understanding of the virus and its underlying pathways impacting humans”

M a y 0 4, 2 0 2 0

Andreas Scherer, PhD

CEO of Golden Helix

II. NEXT-GENERATION SEQUENCING (NGS) A Tool for SARS-CoV-2 DIAGNOSIS, AND REVEALS THE PROGRESSION OF COVID-19

1. Introduction

While the pandemic has prompted an unprecedented global effort to find therapeutic targets and develop treatments and vaccines, to date, decisive remedies are lacking.

Recent experience with emerging infectious diseases, such as SARS, MERS, Zika and Ebola has demonstrated that NGS technologies represent powerful tools for tracing origins, spread and transmission chains of outbreaks, as well as for monitoring the evolution of the etiological agents. Accordingly, the COVID-19 pandemic has triggered unprecedented efforts for the development of effective real-time surveillance strategies based on sequencing of the genome of its causative agent with more than 100 000 complete or near complete SARS-CoV-2 having been deposited in dedicated repositories such as EpiCov and others. These data have already fostered several studies on the evolutionary dynamics of the virus, and the identification of variants of potential clinical relevance (Babb de Villiers and *al*., 2021).

2. A brief history "The birth of next-generation sequencing methods"

After the completion of the Human Genome Project (HGP), the ambition was to sequence a large number of genomes in order to study genetic variation and to carry out genome-wide association studies. Studies - GWAS), in which we try to identify links between genetic diseases and specific genetic profiles. For this purpose, the first NGS technology was the "454" method, launched in 2005 by the 454 Life Sciences company. Their 454 Genome Sequencer produced approximately 200,000 reads of 110 base pairs in length per run, that is, per cycle of operation (Gkazi, 2021).

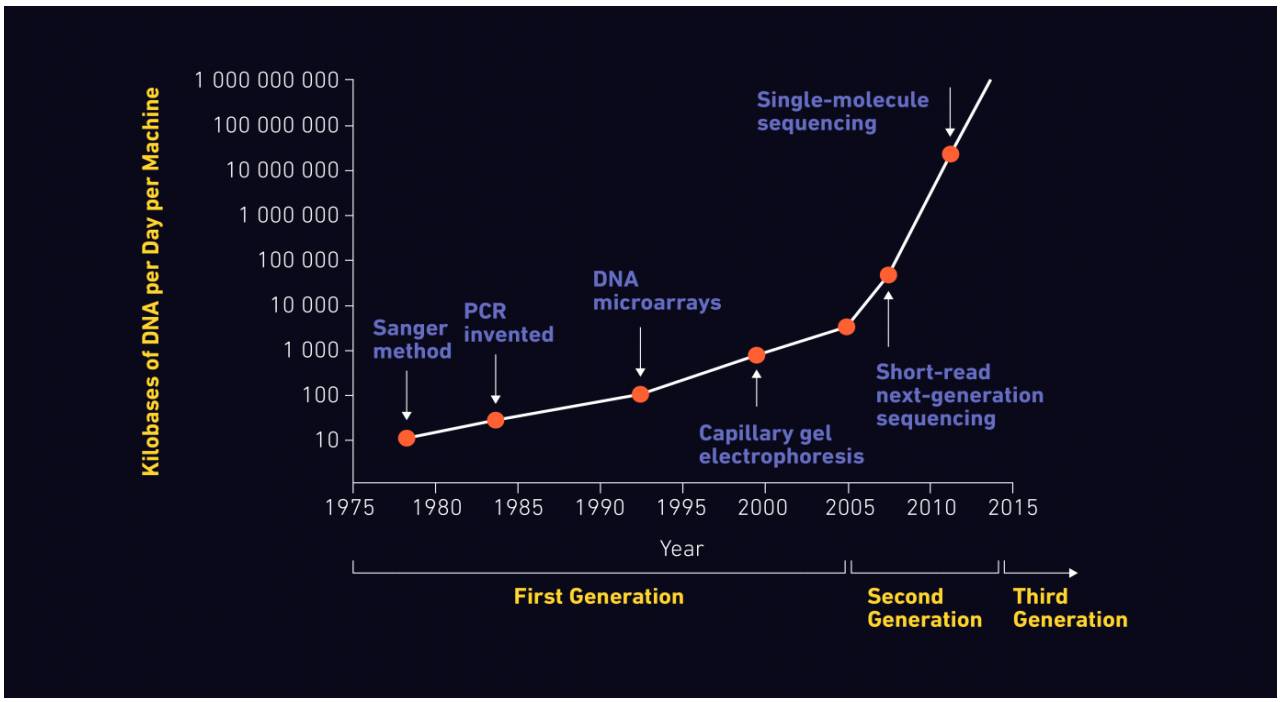

Figure 09. The evolution of sequencing methodologies (Gkazi, 2021).

Next-generation sequencing (NGS) methods have been developed. These methods share three major improvements over Sanger sequencing:

1- Instead of molecular cloning of DNA fragments followed by introduction into host cells and isolation of each clone individually, a library containing all of the fragments is made directly into a tube (in vitro). The Figure below schematically shows the procedure; DNA fragments, generated by random (enzymatic or mechanical) cuts of genomic DNA, are linked to small molecules of DNA of known sequences called "adapters". Random cuts generate fragments of a wide variety of sizes (between about 50 nt and 2000 nt).

A size selection is usually made for two reasons:

- Remove fragments shorter than the sequencing length.
- Remove excessively long fragments (larger than about 1000 nt).

This last step is important for NGS techniques which then require amplification by PCR (polymerase chain reaction), which is less efficient on long fragments. Note that while in the Figure below a single genomic DNA molecule is shown, in reality a library is made from a large number of copies of the genomic DNA molecules to be sequenced.

2- While the machines developed for the Human Genome Sequencing Project were only able to perform a few hundred Sanger sequencing reactions in parallel, NGS sequencers can perform millions or even billions of sequencing reactions in parallel.

3- NGS technologies do not separate the fragments by electrophoresis; the detection of the nucleotides incorporated by the polymerase is carried out directly after each cycle of incorporation (Van Dijk, 2021).

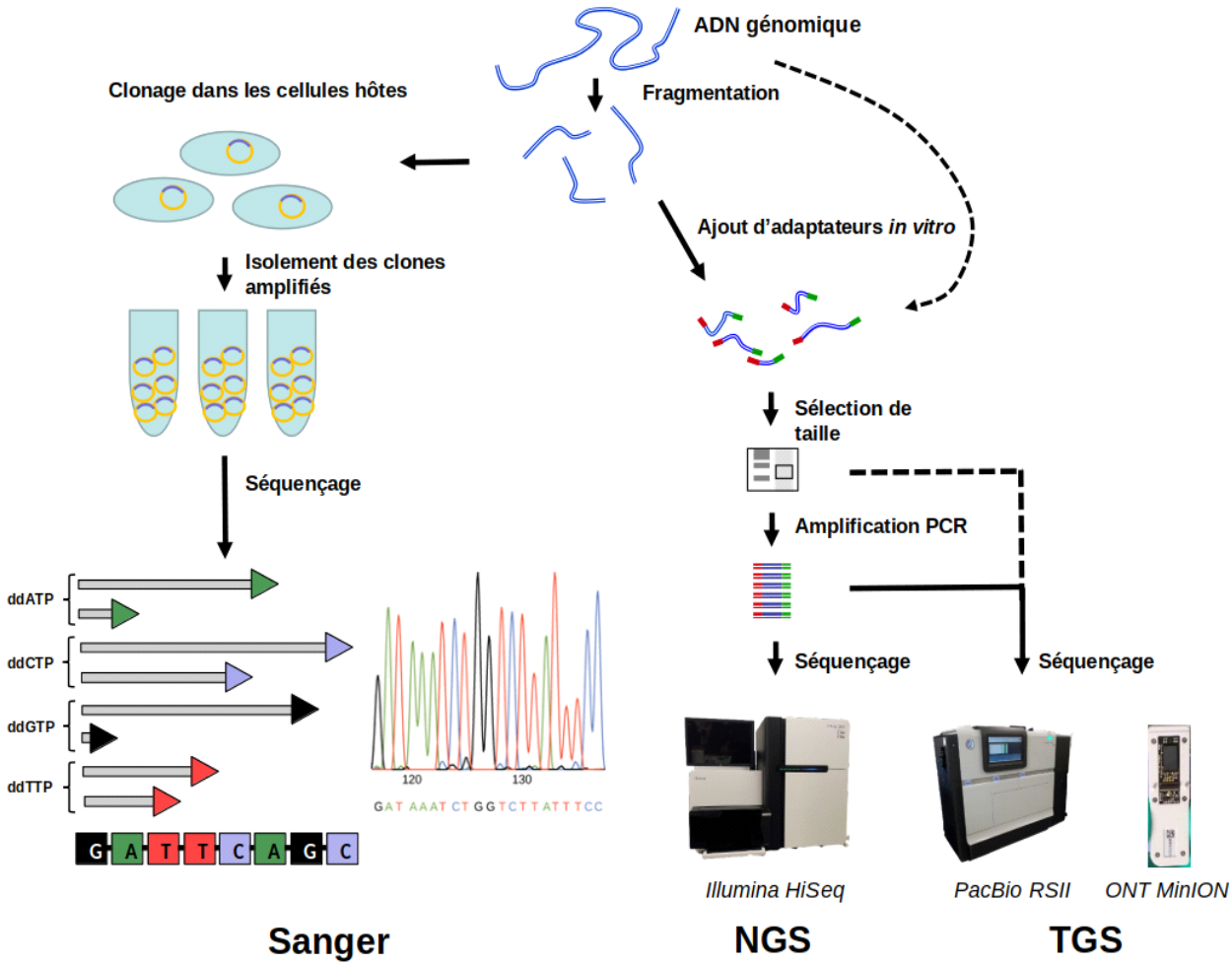

Figure 10. Comparison of Sanger, Next Generation (NGS, Next Generation Sequencing) and Third Generation (TGS, Third Generation Sequencing) methods
(Van Dijk, 2021).

3. Next generation sequencing definition

NGS, also called "high throughput sequencing" or else “second generation sequencing”, is a term for a variety of genetic sequencing techniques, which provide improvements to the initial process of Sanger sequencing. These techniques include Illumina sequencing (Solexa), Roche 454 sequencing, Ion Torrent: Proton / PGM sequencing, and SOLiD sequencing. These modes of

DNA and RNA sequencing use massively parallel processes for faster and more cost-effective operation than the Sanger method **(**https://www.hpe.com/fr/fr/what-is/next-gen-sequencing.html).

Those techniques were developed by different entities, but all have the same philosophy, that is to say:

- Amplification of DNA mainly by emulsion PCR (emPCR) or by solid-phase PCR (solid phase amplification).
- Succession of washing cycles and identification ("wash & scan"): incorporation of nucleotides in the reaction chamber, washing of the reaction chamber, capture of the image.

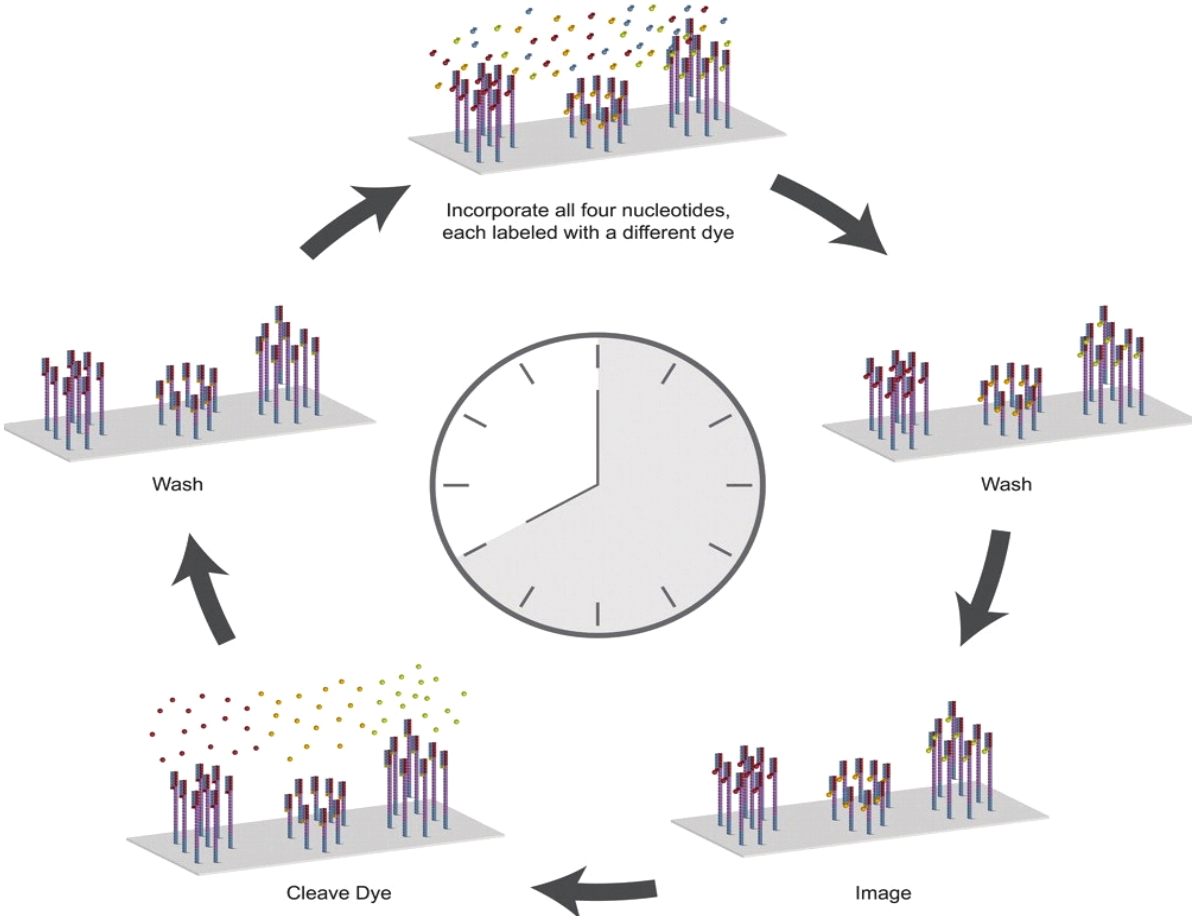

Figure 11. High throughput sequencing method (Yohan, 2021).

4. Main steps of 2G sequencing methods and next-generation sequencing library prep

Regardless of the NGS method chosen, there are several main steps that must be tailored to the target (RNA or DNA) and sequencing system selected.

4.1 Sample preparation (pre-processing)

Nucleic acids (DNA or RNA) are extracted from the selected samples (blood, sputum, bone

marrow etc.). Extracted samples are quality control (QC) checked, using standard methods (spectrophotometric, fluorometric or gel electrophoretic). If using RNA, this must be reverse transcribed into cDNA, however some library preparation kits may include this step.

4.2 Library preparation

Random fragmentation of the cDNA or DNA, typically by enzymatic treatment or sonication, is performed. The optimal fragment length depends on the platform being used. It may be necessary to run a small amount of fragmented sample on an electrophoresis gel when optimizing this process. These fragments are then end-repaired and ligated to smaller generic DNA fragments called adapters. Adapters have defined lengths with known oligomer sequences to be compatible with the applied sequencing platform and identifiable where multiplex sequencing is performed. Multiplex sequencing, using individual adapter sequences per sample, enables large numbers of libraries to be pooled and sequenced simultaneously in a single run. These pools of DNA fragments with adapters attached are known as a sequencing library (Babb de Villiers and *al.*, 2021).

Size selection may then be performed, by gel electrophoresis or using magnetic beads, to remove any fragments that are too short or too long for optimal performance on the sequencing platform and protocol selected. Library enrichment/amplification is then achieved using PCR. In techniques involving emulsion PCR, each fragment is bound to a single emulsion bead which will go on to form the basis of sequencing clusters. Amplification is often followed by a "clean-up" step (e.g., using magnetic beads) to remove undesired fragments and improve sequencing efficiency (Gkazi, 2021).

The final libraries can undergo QC checks using qPCR, to confirm DNA quality and quantity. This will also allow the correct concentration of sample to be prepared for sequencing.

4.3 Sequencing

Depending on the selected platform and chemistry, clonal amplification of library fragments may occur prior to sequencer loading (emulsion PCR) or on the sequencer itself (bridge PCR). Sequences are then detected and reported according to the platform selected.

4.4 Data analysis

The generated data files are analysed depending on the workflow used. Analysis methods are highly dependent on the aim of the study (Gkazi, 2021).

5. Next generation sequencing technologies

5.1 Reversible Terminator Technology "illumine solexa"

The technology of Illumina sequencing started with a concept by British scientists Shankar Balasubramanian and David Klenerman; it involved sequencing of single DNA molecules attached to microspheres. They founded Solexa in 1998, keeping single molecule sequencing in mind, but because of certain limitations, had to shift to sequencing clonally amplified DNA; the system was commercialized in 2006 as the Solexa Genome Analyzer (Voelkerding and *al*., 2009).

The Illumina Genome Analyzer uses flow cells consisting of optically transparent slides with eight individual lanes. Small oligonucleotide anchors are immobilized on the surfaces of these lanes. The template DNA to be sequenced is fragmented, phosphorylated at the 5′end, and adenylated to add a single A at the 3′end. Oligonucleotide adaptors are ligated to the DNA fragment, and the ligation is facilitated by the presence of a single T overhang on the adaptors. The adaptor-ligated oligonucleotides are complementary to the flow-cell anchors, and hence attach to the anchors. These DNA templates attached to the anchors are used to generate clusters of the same DNA fragment by amplification. A DNA fragment bends and hybridizes with its distal end to an adjacent anchor complementary to the distal end. On denaturation, both strands separate, and again bend and hybridize with their distal ends to adjacent anchors complementary to their distal ends. After multiple amplification cycles, a single DNA template makes a clonally amplified cluster with thousands of clonal molecules.

Millions of clusters of different template molecules can be generated per flow cell. For sequencing, the technology uses four fluorescently labeled nucleotides to sequence the tens of millions of clusters on the flow cell surface in parallel. In each growing chain, a single labeled deoxynucleoside triphosphate (dNTP) is added in each cycle. Due to the incorporation of the labeled nucleotide, DNA polymerization terminates, and the fluorescent dye is imaged to identify the incorporation. Then the label is enzymatically cleaved to allow incorporation of the next nucleotide.

5.1.1 Advantages

According to product information available from Illumina, in approximately 11–14 days, huge amounts of data are generated in the form of base pairs sequenced per run. From around 95 GB data coming out from nearly 150 bp long reads from both sides (2 x 150 bp) in the most widely cited platform Genome Analyzer IIx, the throughput has been significantly increased up

to 600 GB data with 2 × 100 bp reads in newer versions of the platform (e.g. HiSeq2500 or HiSeq 2000), resulting in low cost per base. Their bench top platform MiSeq produces approximately 5 GB of data for 2 × 150 bp sequencing, or 8 GB data with 2 × 250 bp sequencing. Large data and low cost per base renders the technology a good choice for many sequencing applications where large read length and de novo construction of a genome is not required (e.g. re-sequencing, ChIP sequencing, certain RNA sequencing projects, etc.).

5.1.2 Limitations

The major concern with the Illumina technology is that of de-phasing, which means different fragments in a cluster are sequenced with different phases; in other words, under- or over-incorporated nucleotides, because of block removal failure or other factors, result in fragments of varying lengths, which reduces precision in base calling at the 3′ ends of the fragments. De-phasing increases with increased read length. It is more common at sequences of invert repeats or GGC (Nakamura and *al*., 2011). Illumina technology produces reads of short length "micro-reads," hence assembly and downstream bioinformatics could be a challenge, especially for certain de novo sequencing. Longer run-time is also a limitation (Babb de Villiers and *al.*, 2021)

5.2 Sequencing by Ligation Technology "ABI Solide"

Sequencing by ligation technology is marketed by Applied Biosystems, USA. The name SOLiD stands for Small Oligonucleotide Ligation and Detection System. This technology was developed by George Church in 2005, and was further improved and distributed by Applied Biosystems in 2007 (Voelkerding and *al*., 2009). The principle of this sequencing relies upon the ability of DNA ligase to detect and incorporate bases in a very specific manner. In sequencing by ligation, DNA fragments attached to beads are clonally amplified by emulsion PCR. After PCR, specific primers hybridize to the adaptor sequence of the amplified templates on the beads, which provides a free 5′ phosphate group for ligation to the fluorescently labeled probes (called interrogation probes) instead of providing a 3′ hydroxyl group as in normal polymerase-mediated extension. The interrogation probe is 8 bp in length, where the first two bases are specific, and the rest of the 6 bases are degenerate. A set of four fluorescently labeled interrogation probes, consisting of one of 16 possible 2-base combinations at the end (e.g. TT, GT, TC, CG, etc.), compete for ligation to the sequencing primer. Upon ligation, fluorescence is captured, which is corresponding to the probe ligated. For the second cycle, the "fluor" of the attached probe is removed and a 5′ phosphate group is regenerated. Multiple cycles of ligation, detection, and cleavage are performed, with the number of cycles determining the eventual read length. Following a series of ligation cycles (usually seven), the extension product is removed and the

template is reset with a primer complementary to the n-1 position for a second round of ligation cycles. This process is repeated each time with a new primer with a successive offset (n-1, n-2, n-3, and so on). Thus the sequencing is divided into library preparation, emulsion PCR, bead deposition, sequencing, and primer reset. A 6–7-day long instrument run in a SOLiD 5500 system claims to generate sequence data at approximately 10–15 GB/day (total throughput 120–240 GB, 100 GB in the case of the SOLiD 4 system) with a read length of 75 bases (for mate-paired: 2 x 60 bp; for paired-end: 75 bp × 35 bp) and a sequence consensus accuracy of 99.99% (Voelkerding and *al*., 2009).

5.2.1 Advantages

The advantage of this technology is generation of sequencing data of comparatively higher accuracy than other sequencing methods. One of the reasons behind the high accuracy is sequencing with successive offset primer less by one bp so that each nucleotide of the template is sequenced twice; therefore, in order to miscall a SNP, two adjacent colors must be miscalled, which does not frequently happen.

5.2.2 Limitation

Among the limitations of this technology are that less data are output than with Illumina, and shorter read length, requiring close genome sequencing for mapping. Even the time taken for a whole run is about 6–7 day to complete, especially for bigger genomes.

5.3 Pyrosequencing Technology "Roche-454 GS FLX "

In 1993, Nyrén and his group published a novel sequencing method based on chemiluminescent detection of pyrophosphate released during polymerase-mediated deoxynucleoside triphosphate (dNTP) incorporation, which was later commercialized by 454 Life Sciences with technical refinement (Ronaghi and *al*., 1996) and the use of emulsion-based PCR. Later, in 2007, Roche acquired 454. In pyrosequencing, DNA to be sequenced is fragmented and subjected to its complementary strand synthesis by DNA polymerase. As the polymerase incorporates a nucleotide in the growing chain, a pyrophosphate molecule is released. This pyrophosphate, through a series of enzymatic reactions, is converted into ATP. Then ATP is used to enzymatically convert luciferin into oxyluciferin, which emits fluorescence that is recorded by the camera. By detecting this fluorescence, the incorporation of a nucleotide is confirmed. The identity of the incorporated nucleotide is known, as four dNTPs (dATP, dTTP, dCTP and dGTP) are introduced in the reaction separately in predefined cycles (Ronaghi and *al*., 1996).

454 Sequencing uses a massively parallel pyrosequencing system capable of sequencing up to 1,000 bp of DNA in a 23-hour run on its new Genome Sequencer FLX Plus instrument. The technology works by fragmenting the DNA into approx 800–1,000 bp in length (nebulization), ligating adaptors to DNA fragments, making a library, and attaching the library to small DNA-capture beads. The beads are compartmentalized into water-in-oil emulsion microvesicles, often called micro-reactors, where clonal multiplication of single DNA molecules bound to the beads occurs during emulsion PCR. After amplification, the emulsion is disrupted, and the beads containing clonally amplified template DNA are enriched. Each amplified DNA-bound bead is placed into a tiny well on a PicoTiterPlate consisting of around 3.4 million wells. A mix of enzymes such as DNA polymerase, ATP sulfurylase, and luciferase, are also packed into the well. The PicoTiterPlate is then placed into the sequencer machine for sequencing. The GS FLX platform can produce data of approximately 450 MB with a 400–600 bp read length. The new platform GS FLX Plus is, however, able to generate approximately 700 MB of data with a read length of 700–1,000 bp (Gupta and Gupta, 2014; Ronaghi and *al*., 1996).

5.3.1 Advantages

The advantage of 454 technology lies within its ability to sequence reads in the read length of 700–1000 bp. The longer read length is advantageous in terms of downstream bioinformatics, resulting in sequence assembly with longer contigs, higher N50 length, and less gaps, especially in de novo sequencing projects. Longer paired-end reads produced by the 454 platform also facilitate construction of better scaffolds (Gupta and Gupta, 2014).

5.3.2 Limitations

The 454 technology has certain limitations. The primary one is the difficulty in sequencing homopolymer repeats due to simultaneous incorporation of the same nucleotide, producing light that cannot be discriminated after a certain length (> 6 bp) with high accuracy. Another disadvantage of the technology is the generation of relatively low bases/run (around 700 Mb) as compared to other NGS technologies. This makes it relatively expensive technology and not of priority if re-sequencing is desired with a high X dept (Gupta and Gupta, 2014; Mardis, 2008).

5.4 Ion semiconductor sequencing

Ion semiconductor sequencing is a method of DNA sequencing based on the detection of hydrogen ions that are released during the polymerization of DNA. This is a method of

"sequencing by synthesis", during which a complementary strand is built based on the sequence of a template strand.

A microwell containing a template DNA strand to be sequenced is flooded with a single species of deoxyribonucleotide triphosphate (dNTP). If the introduced dNTP is complementary to the leading template nucleotide, it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers an ISFET ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence, multiple dNTP molecules will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.

This technology differs from other sequencing-by-synthesis technologies in that no modified nucleotides or optics are used. Ion semiconductor sequencing may also be referred to as Ion Torrent sequencing, pH-mediated sequencing, silicon sequencing, or semiconductor sequencing.

5.4.1 Advantages

The major benefits of ion semiconductor sequencing are rapid sequencing speed and low upfront and operating costs. This has been enabled by the avoidance of modified nucleotides and optical measurements.

5.4.2 Limitations

If homopolymer repeats of the same nucleotide (e.g. TTTTT) are present on the template strand (strand to be sequenced) then multiple introduced nucleotides are incorporated and more hydrogen ions are released in a single cycle. This results in a greater pH change and a proportionally greater electronic signal. This is a limitation of the system in that it is difficult to enumerate long repeats. This limitation is shared by other techniques that detect single nucleotide additions such as pyrosequencing. Signals generated from a high repeat number are difficult to differentiate from repeats of a similar but different number; e.g., homorepeats of length 7 are difficult to differentiate from those of length 8.

Another limitation of this system is the short read length compared to other sequencing methods such as Sanger sequencing or pyrosequencing. Longer read lengths are beneficial for de novo genome assembly. Ion Torrent semiconductor sequencers produce an average read length of approximately 400 nucleotides per read

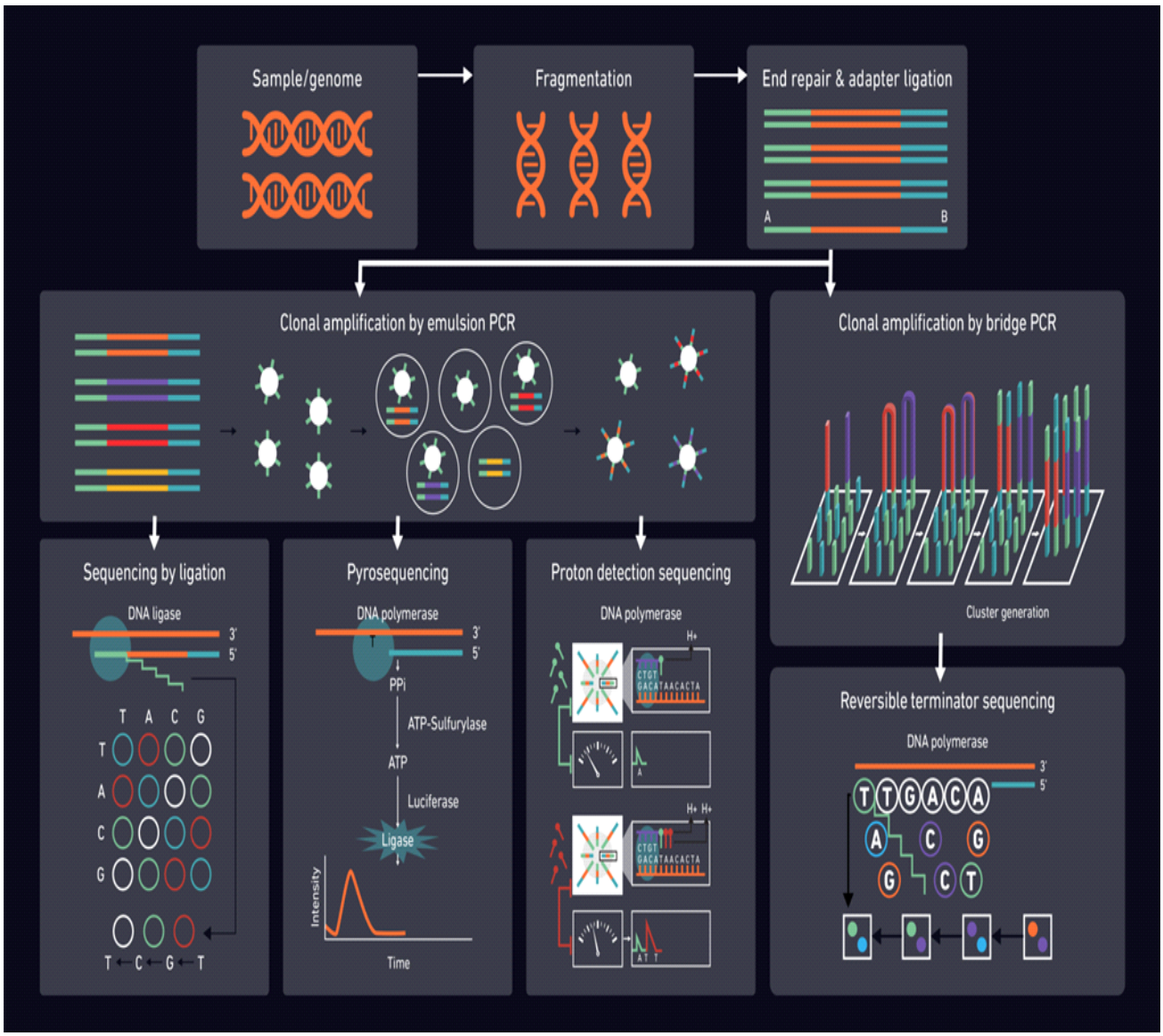

Figure 12. Diagram representing the principle 2G sequencing platforms and chemistries (Gkazi, 2021).

5.5 Applications of high throughput sequencing:

DNA sequencing by NGS can be applied within different applications, such as partial-exome (PES), whole-exome (WES) or whole-genome sequencing (WGS). The broad range of applications opens new and more affordable possibilities to study numerous cellular processes at the single-base resolution.

However, both WES and WGS produce massive amounts of data, which presents significant challenges for data storage, distribution, analysis and interpretation. In the nearer future, this will remain one of the main bottlenecks of all described approaches.

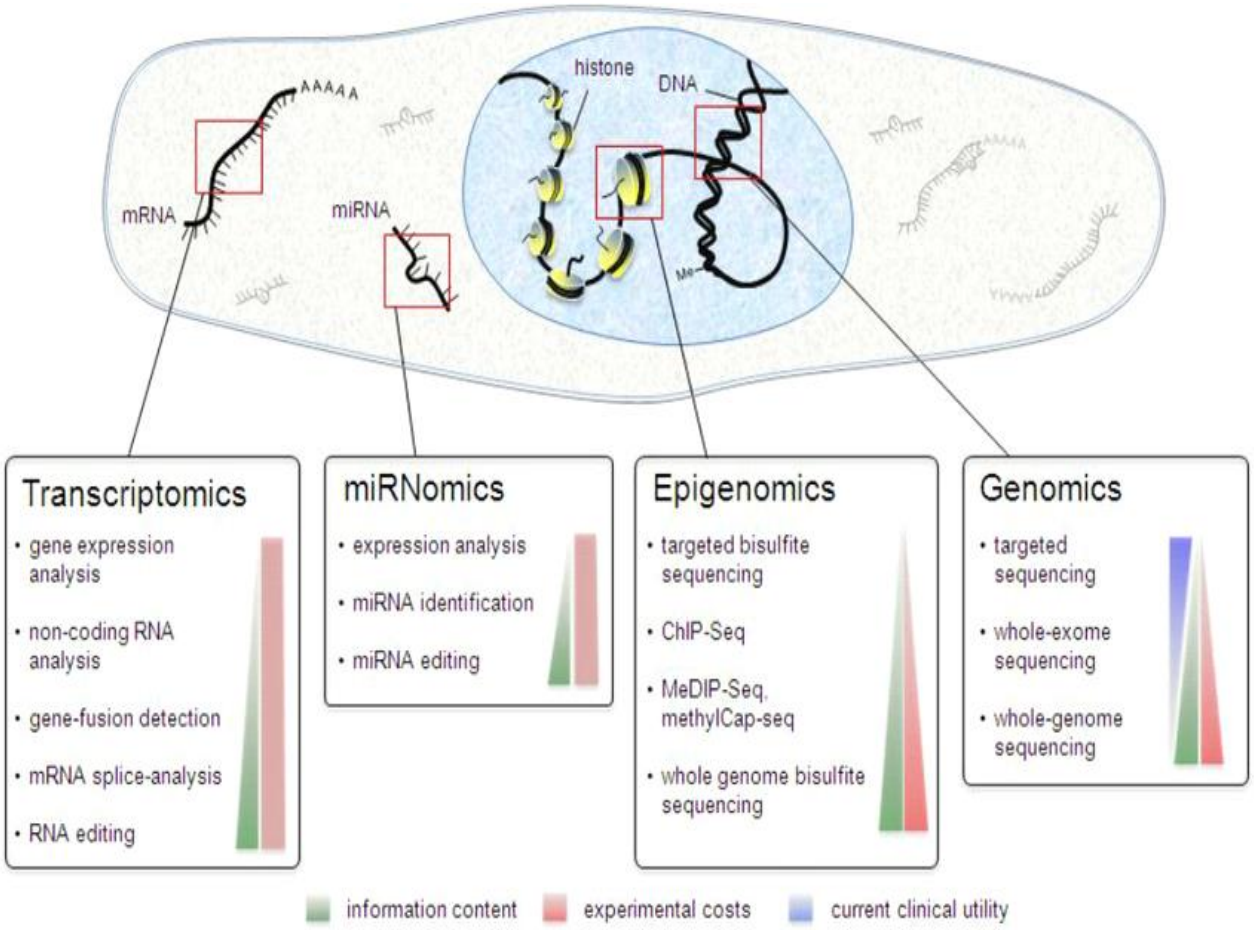

Figure 13. Next-generation sequencing applications. Schematogram depicting the different methods for transcriptomic, miRNomic, epigenomic and genomic studies. (Frese and al., 2013)

6. Third generation sequencing

The third generation of mass sequencing is symbolized by the sequencing of a single DNA molecule (SMS or “Single Molecule Sequencing”).Unlike the second generation, no DNA (or RNA) amplification is required to perform the measurements. Only one molecule is "read".

Several technologies and methods are at work here, each with advantages and disadvantages. It can roughly be classified into three categories:

1. Sequencing by synthesis, observation of DNA polymerase as it synthesizes the DNA strand.
2. Detection of the bases one after the other as the DNA sequence passes through a nanopore.
3. Direct observation of the DNA molecule using microscopy technique. https://www.hpe.com/fr/fr/what-is/next-gen-sequencing.html

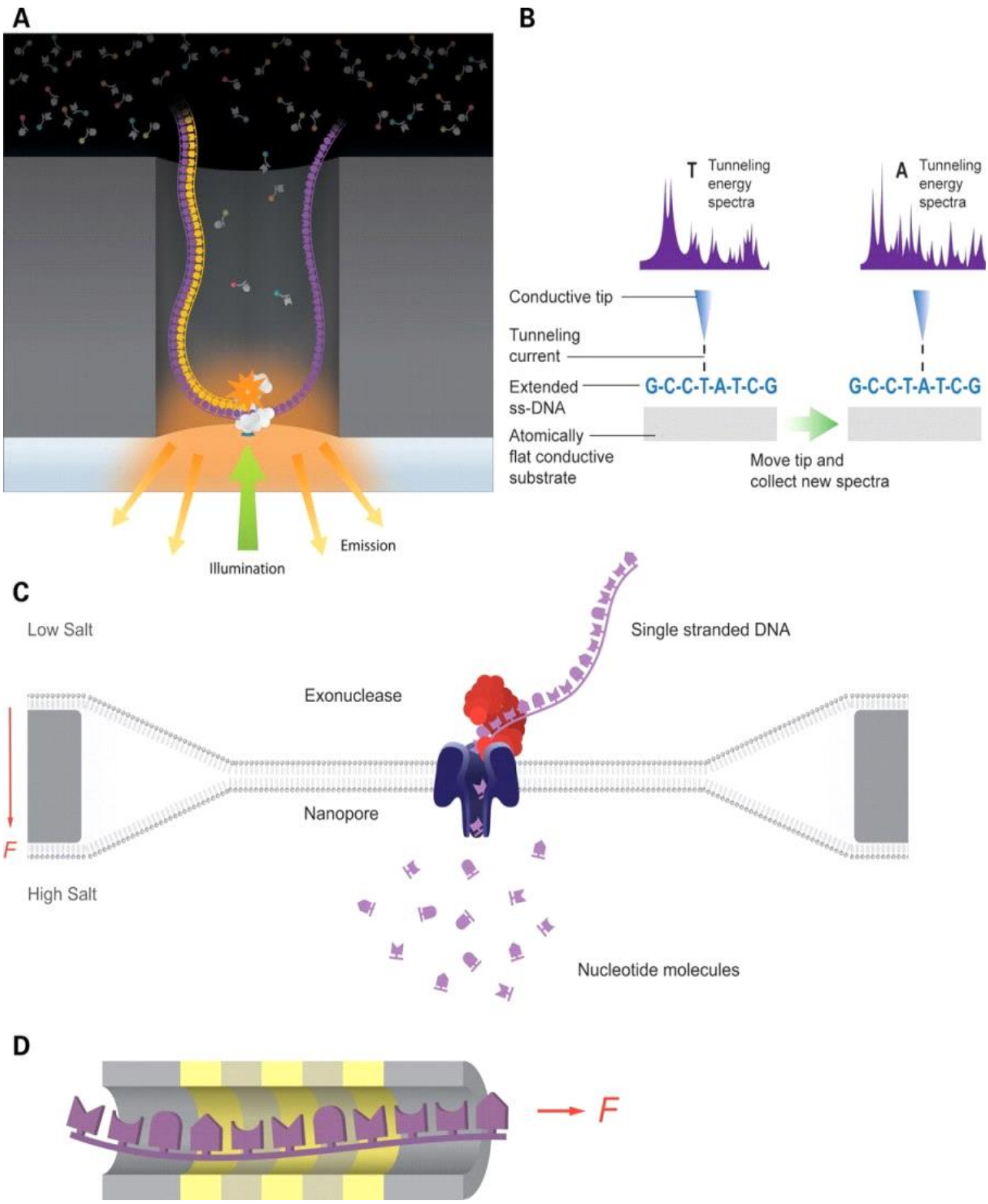

Figure 14. 3rd generation of sequencing tools.

(A) Pacific Biosciences - Direct observation of DNA synthesis in real time - **(B)** Reveo - Direct inspection of the DNA molecule by electron microscopy - **(C)** Oxford Nanopore - Measures the translocation of nucleotides by measuring ionic concentrations **(D)** IBM - reading of bases. https://www.hpe.com/fr/fr/what-is/next-gen-sequencing.html

6.1 Nanopore sequencing

Nanopore sequencing is a third generation approach used in the sequencing of biopolymers-specifically, polynucleotides in the form of DNA or RNA.

Using nanopore sequencing, a single molecule of DNA or RNA can be sequenced without the need for PCR amplification or chemical labeling of the sample. At least one of these aforementioned steps is necessary in the procedure of any previously developed sequencing approach. Nanopore sequencing has the potential to offer relatively low-cost genotyping, high mobility for testing, and rapid processing of samples with the ability to display results in real-time. Publications on the method outline its use in rapid identification of viral pathogens, monitoring ebola, environmental monitoring, food safety monitoring, human genome sequencing, plant genome sequencing,monitoring of antibiotic resistance, haplotyping and other applications.

7. NGS tools for Detection and sequencing SARS-CoV-2:

In late January 2020, Lu et *al*. reported SARS-CoV-2 genomic data from nine patients presenting with pneumonia of unknown origin at three hospitals in Wuhan, China. BAL and cultured isolates were used as samples. The patients' samples were negative for known respiratory pathogens, with five tested by the Chinese CDC and four by the BGI group in Beijing, China. NGS technology was used to sequence and identify the causative pathogen, with the BGI and CDC labs differing slightly in their sequencing techniques and bioinformatic processing pipelines. In both groups, gaps between contigs were connected using Sanger sequencing and terminal genome regions were identified via rapid amplification of cDNA ends (RACE).

At the BGI group, RNA extraction of BALF samples was carried out with a QIAamp Viral RNA Mini Kit, and a probe-captured technique was used to remove human nucleic acid material. Next, RNA was reverse transcribed to cDNA, second-strand synthesis was performed, and a DNA library was constructed. The DNA library was quantified with a Qubit method and transformed into a single strand circular library. Rolling circle amplification was used to construct DNA nanoballs, and they were subsequently qualified. The DNBSEQ-T7 high throughput sequencer from MGI was used with paired-end, 100 bp read lengths. High quality reads were filtered for human reads against the hg19 human reference genome with Burrow-Wheeler alignment software. The remaining data were aligned with published data on coronaviruses from the US National Center for Biotechnology Information. Mapped reads were assembled with SPAdes software to create a consensus genome sequence.

The Chinese CDC sequencing protocol similarly used the QIAamp Viral RNA Mini Kit to extract viral RNA from the clinical samples, followed by cDNA synthesis and second-strand synthesis. cDNA libraries were generated and then purified with Agencourt AMPure XP beads to remove contaminants. Following quantitation, the sequencing was carried out on MiSeq or iSeq platforms from Illumina. The terminal genome regions were identified by the use of Rapid amplification of cDNA ends (RACE) system from Invitrogen. Assembled genomes were confirmed with traditional Sanger sequencing. The raw sequencing reads were filtered via the same protocol used by the BGI group, and CLCBio software version 11.0.1 was used for de novo assembly, variant calling, and alignment. The bat-SL-CoVZC45 virus (containing 87.99% sequence similarity) was also used to perform a mapped assembly (Babb de Villiers and *al.*, 2021 ; John and *al*., 2021).

Sequencing yielded eight full genomes and two partial genomes (one patient's BALF sample was used to isolate the virus, which was also sequenced, yielding 10 total samples). The sequences were used to generate PCR-based assays, that were then used to confirm the presence of the SARS-CoV-2 virus, and cycle threshold (Ct) values ranged from 22.85 to 34.23 (John and *al*., 2021).

The results of sequencing the viral genome in this study yielded highly useful information during the early stages of the SARS-CoV-2 outbreak. Genomic analyses led to the revelation that, while the whole genome sequence of SARS-CoV-2 is highly similar to bat-SL-CoVZC45 (87.99% similarity) and bat-SL-CoVZXC21 (87.23%), the receptor binding domain (S1) sequence of the spike protein (S), was more similar to that of SARS-CoV, the virus responsible for the first SARS outbreak in the early 2000s. This evidence supports the suggestion that SARS-CoV-2 uses the ACE-2 receptor to gain entry into cells, the same route utilized by SARS-CoV. The utilization of ACE-2 receptors by SARS-CoV-2 has also been demonstrated in infectivity studies by Zhou et al. The phylogenetic analysis, made possible by the assembled sequences, allowed the classification of the virus, showing that the virus belongs to the subgenus Sarbecovirus, a member of the Betacoronavirus genus. The high sequence similarity (over 99.9%) among viral samples obtained from the nine patients in Wuhan provides evidence of very recent entry into the human population (John and *al*., 2021).

Other laboratories in China conducted parallel investigations at the onset of the outbreak, such as Zhu et *al*., who used a similar combination of Illumina and nanopore sequencing, RACE, and Sanger sequencing to identify and characterize the SARS-CoV-2 genomes extracted from three patient samples in Wuhan, China. Their bioinformatics pipeline included CLC Genomics

software, version 4.6.1; Muscle; and RAxML for phylogenetic analysis. Their sequencing protocol yielded more than 20,000 viral reads per sample, obtaining one full-length genome and two nearly full-length genomes. They similarly noted that contigs aligned with high similarity with bat-SL-CoVZC45. Published 24 January 2020, they reached similar conclusions to Lu and *al*. regarding the phylogenetic characterizations of the virus and used their de novo generated sequences to design primers for PCR-based diagnostic assays. Groups all over the world are now investigating possible diagnostic interventions made possible by NGS technology. Campos et al. reported the use of metatranscriptomic next-generation sequencing technology in the detection of SARS-CoV-2 in a nasopharyngeal swab specimen from a patient in Feira de Santana-Bahia, Brazil. They used the Ion S5 platform from ThermoFisher with an Ion 540™ chip and the Ion Total RNA-Seq kit v2. This platform uses an ion-semiconductor sequencing process, and they implemented the Low Input RiboMinus™ Eukaryote System v2 from ThermoFisher to remove rRNA from one sample. The rRNA-depleted library contained human transcripts as 77.29% of total reads, while the whole RNA library had 84.49% of total reads as human transcripts. Contigs from the rRNA-depleted library provided 29.9% genome coverage, while contigs from the non-depleted sample yielded only 5.4% genome coverage. Total genome coverage from all viral reads in the rRNA depleted sample was 59.9%. These results indicate that rRNA-depletion strategies may play a role in improving NGS diagnostic abilities.

Moore et al. have reported on the use of amplicon- and metagenomic-MiniION based sequencing in the identification of SARS-CoV-2 and co-infections, respectively. Amplicon-based NGS is a tool that is commonly used to provide highly specific data on the presence of organisms in a sample via primers targeting highly conserved areas of a genome. This is contrasted with metagenomic-NGS which takes a "shotgun" approach, identifying all genetic material in a sample, not just those that contain the highly conserved genetic region. Primers in this amplicon-based approach were designed with sufficient overlap that the sequence of SARS-CoV-2 could be reconstructed from the individual fragments. The study was limited as it only included two patients, both from the UK. Primers were designed for amplicon-based NGS sequencing of SARS-CoV-2 to generate approximately 1000 base pair fragments with roughly 200 base pair overlaps for sequence assembly, and the assay successfully sequenced the SARS-CoV-2 genome in both patients. For validation, they spiked samples with VP35 RNA from Ebola Virus as an internal control. The mapping software successfully identified the internal control RNA and also identified the presence of Fusobacterium periodonticum and human cytomegalovirus (human betaherpes virus 5) in addition to SARS-CoV-2 in the mNGS results.

The group used Oxford Nanopore Technology's (ONT's) cloud-based pipeline EPI2ME (WIMP rev. 3.2.2) workflow for bioinformatic analysis.

Patient 1 was sampled twice (two days apart) and patient 2 was sampled once, yielding total reads of 8,698,559 (78.6% of which were human reads), 9,890,327 (97.7%), and 5,849,966 (92.7%), respectively, during mNGS. The metagenomic approach did not provide uniform genome coverage among the three genomes, and the amplicon-based sequencing method provided a much higher read depth than the metagenomic approach. (John and *al*., 2021).

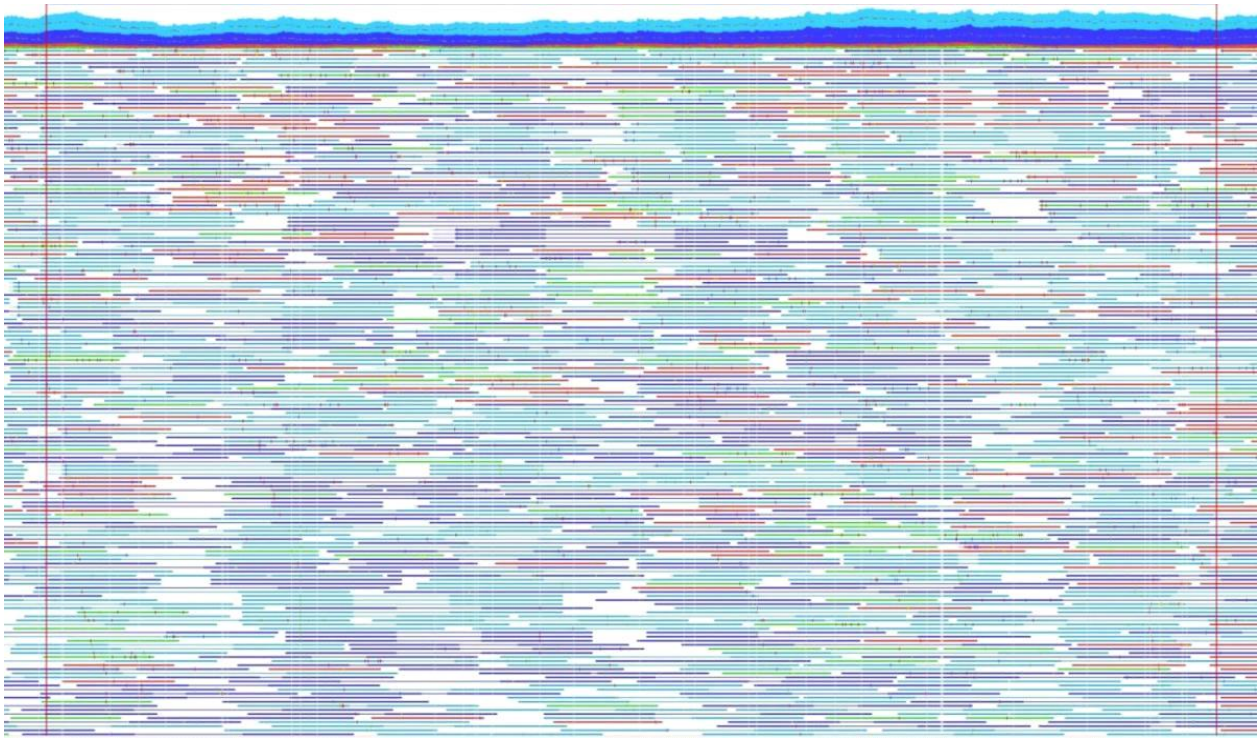

Figure 15. Whole genome sequence of the 2019-nCoV coronavirus, in one of the first French cases, made at the Pasteur Institute (Paris), using a unique Platform (P2M), open to all French National Reference Centers (Pasteur Institute, 2021)

8. Launch of the network of COVID-19 genome sequencing laboratories in Africa

In September 2020 the WHO and the Africa CDC launched a network of laboratories to reinforce genome sequencing of SARS-CoV-2 in Africa. Twelve specialised and regional reference laboratories in the network provide sequencing, data analysis and other technical support services to the countries where they are located as well as to neighbouring countries and countries in their sub-regions. The following were selected as the specialised continental reference sequencing research laboratories for emerging pathogens:

Redeemer's University African Centre of Excellence for Infectious Diseases (ACEGID), Nigeria; South African National Bioinformatics Institute (SANBI); and Kwazulu-Natal Research Innovation and Sequencing Platform (KRISP) in South Africa (Babb de Villiers and *al.*, 2021)

Genomes of SARS-CoV-2 from Africa have been sequenced in laboratories in the Democratic Republic of the Congo (DRC), Kenya, Egypt, Gambia, Ghana, Nigeria, Senegal, South Africa, Tunisia, and Uganda

Algeria sent its samples to a laboratory in Paris for sequencing (28 May 2020), and Hungary recently (said the responsible of bioinformatics platform, Hebbachi, K. at Pasteur institute, DELY IBRAHIM).

By May 2020, the Institut National de Recherche Biomédicale (INRB) in DRC contributed nearly 60% of the SARS-CoV-2 genome sequences from the African continent. It built up this capacity during the ongoing Ebola outbreak in the eastern part of the country.

Table 01. A network of laboratories to reinforcing genome sequencing of SARS-CoV-2 in Africa (Babb de Villiers and *al.*, 2021)

Submitting lab	Number of sequences
South Africa KRISP, KZN Research Innovation and Sequencing Platform	2090
Kenya KEMRI-Wellcome Trust Research Programme/KEMRICGMR-C Kilifi	512
South Africa National Institute for Communicable Diseases of the National Health Laboratory Service	459
Gambia MRCG at LSHTM Genomics lab	427
Democratic Republic of the Congo Pathogen Sequencing Lab, National Institute for Biomedical Research (INRB)	353
South Africa NHLS/UCT	316
Nigeria African Centre of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria	213
Senegal Institut Pasteur de Dakar	136
Uganda MRC/UVRI & LSHTM Uganda Research Unit	133
South Africa National Health Laboratory Service (NHLS), Tygerberg, Cape Town	119

Africa CDC and WHO, in collaboration with other partners, are providing Member States with sequencing equipment, reagents and technical support to accelerate the sequencing of SARS-CoV-2 in Africa.

This partnership between WHO and Africa CDC to establish a network of COVID-19 sequencing laboratories is very important in determining the response to a given strain of SARS-CoV-2 and in helping countries manage the localized or imported transmission (WHO, 2021).

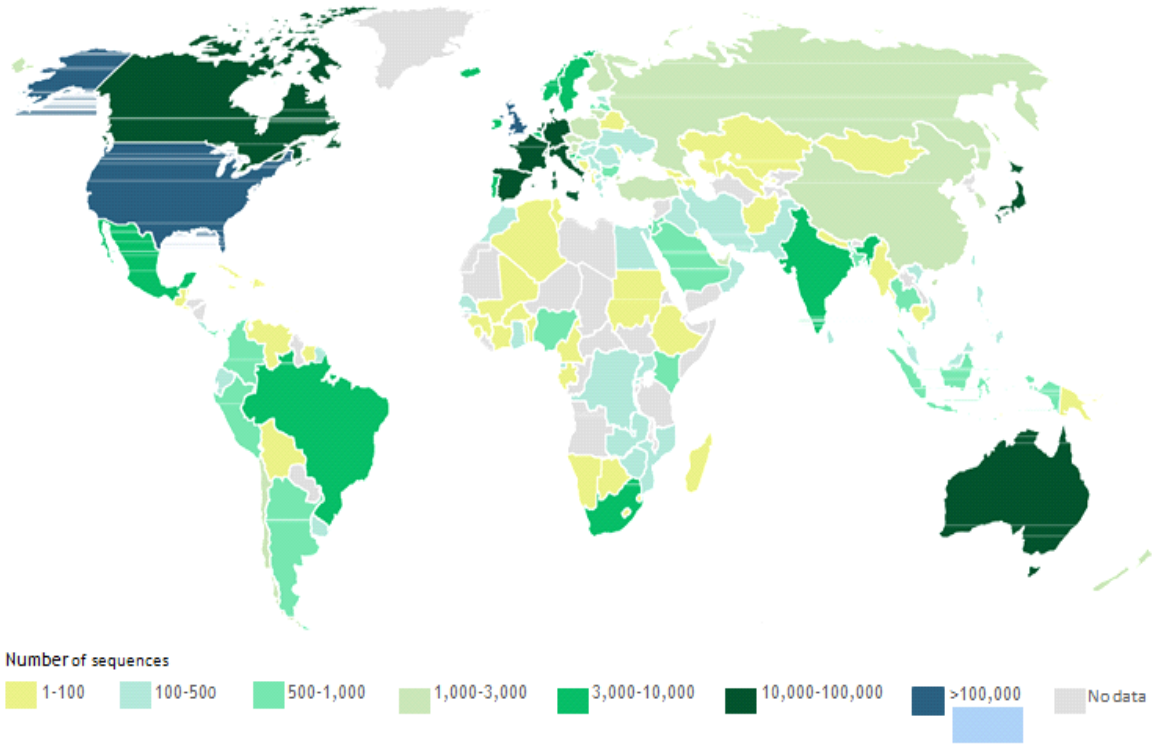

Figure 16. Map showing numbers of SARS-CoV-2 genome sequences uploaded to GISAID by 28 February 2021 (Babb de Villiers and *al.*, 2021)

CHAPTER THREE

SARS-COV-2 PHELOGENETIC ANALYSIS

"And thus, the forms of life throughout the universe become divided into groups subordinate to groups."

CHARLES DARWIN

- ON THE ORIGIN OF SPECIES -

1859

III. SARS-COV-2 PHELOGENETIC ANALYSIS

1. Introduction

Phylogenetic analysis using molecular data such as DNA sequence for genes and amino acid sequence for proteins is very common not only in the field of evolutionary biology but also in the wide fields of molecular biology. The reason is that DNA sequencing became very popular and a huge amount of sequence data of genes and proteins are available in the public online database.. The methods for phylogenetic analysis are improving along with the evolution of computer science. Thus, there are many methods to infer phylogenetic tree, and many programs for each method are available.

The origin of SARS-Cov-2 and its evolutionary relationship is still ambiguous. Several reports attempted to figure out this critical issue by genome-based phylogenetic analysis, yet limited progress was obtained, principally owing to the disability of these methods to reasonably integrate phylogenetic information from all genes of SARS-CoV-2.

2. Phylogenetic analysis

In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor.

The diagram is known as a phylogenetic tree. Phylogenetic analysis is important for gathering information on biological diversity, genetic classifications, as well as learning developmental events that occur during evolution.

With advancements in genetic sequencing techniques, phylogenetic analysis now involves the sequence of a gene to understand the evolutionary relationships among species. DNA being the hereditary material can now be sequenced easily, rapidly, and cost-effectively, and the data obtained from genetic sequencing is very informative and specific. Also, morphological estimates can be used to infer evolutionary developments, especially in cases where genetic material is not available (fossils) **(**Sanchari Sinha Dutta, 2021).

3. Phylogenetic tree

A phylogenetic tree, also known as a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor. Phylogenies are useful for organizing knowledge of biological diversity, for structuring classifications, and for providing insight into events that occurred during evolution. Furthermore,

because these trees show descent from a common ancestor, and because much of the strongest evidence for evolution comes in the form of common ancestry, one must understand phylogenies in order to fully appreciate the overwhelming evidence supporting the theory of evolution (Baum, 2008).

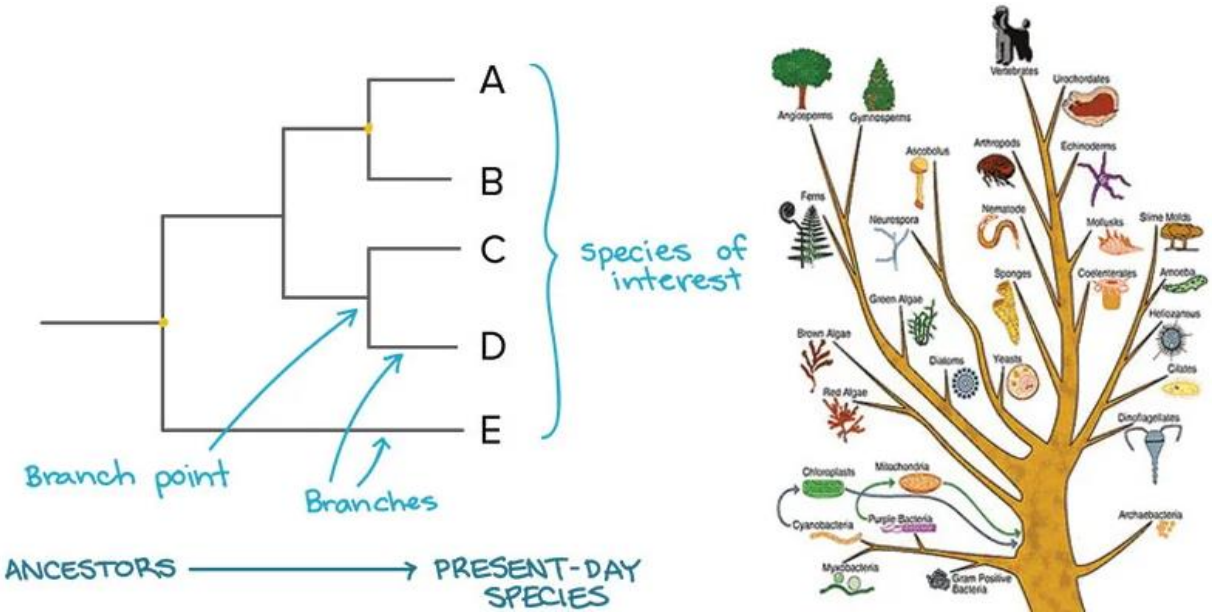

Figure 17. Representation of Phylogenetic tree (Aryal, 2019).

4. Tree-Building Methods

The most popular and frequently used methods of tree building can be classified into two major categories: phenetic methods based on distances and cladistic methods based on characters.

The former measures the pair-wise distance/dissimilarity between two genes, the actual size of which depends on different definitions, and constructs the tree totally from the resultant distance matrix. The latter evaluate all possible trees and seek for the one that optimizes the evolution.

4.1 Distance-Based Methods (phenetic)

The most popular distance-based methods are the unweighted pair group method with arithmetic mean (UPGMA), neighbor joining (NJ) and those that optimize the additivity of a distance tree (FM and ME) (Baxevanis and Ouellette, 2001).

4.1.1 UPGMA Method

This method follows a clustering procedure:

1- Assume that initially each species is a cluster on its own.

2- Join closest 2 clusters and recalculates distance of the joint pair by taking the average.

3- Repeat this process until all species are connected in a single cluster.

Strictly speaking, this algorithm is phenetic, which does not aim to reflect evolutionary descent. It assigns equal weight on the distance and assumes a randomized molecular clock. WPGMA is a similar algorithm but assigns different weight on the distances. UPGMS method is simple, fast and has been extensively used in literature. However, it behaves poorly at most cases where the above presumptions are not met.

4.1.2 Neighbor Joining Method (NJ)

This algorithm does not make the assumption of molecular clock and adjust for the rate variation among branches. It begins with an unresolved star-like tree. Each pair is evaluated for being joined and the sum of all branches length is calculated of the resultant tree. The pair that yields the smallest sum is considered the closest neighbors and is thus joined.

A new branch is inserted between them and the rest of the tree and the branch length is recalculated. This process is repeated until only one terminal is present. NJ method is comparatively rapid and generally gives better results than UPGMA method. But it produces only one tree and neglects other possible trees, which might be as good as NJ trees, if not significantly better. Moreover, since errors in distance estimates are exponentially larger for longer distances, under some condition, this method will yield a biased tree (Bruno and *al* .,2000; https://guava.physics.uiuc.edu/~nigel/courses/598BIO/498BIOonline-essays/hw2/files/hw2_li.pdf).

4.1.3 Weighted Neighbor-Joining (Weighbor)

This is a new method proposed recently. The Weighbor criterion consists of two terms; an additivity term (of external branches) and a positivity term (of internal branches), that quantifies the implications of joining the pair. Weighbor gives less weight to the longer distances in the distance matrix and the resulting trees are less sensitive to specific biases than NJ and relatively immune to the "long branches attraction/distraction" drawbacks observed with other methods (Bruno and *al* ., 2000).

4.1.4 Fitch-Margoliash (FM) and Minimum Evolution (ME) Methods

Fitch and Margoliash proposed in 1967 criteria (FM Method) for fitting trees to distance matrices (Baxevanis and Ouellette, 2001). This method seeks the least squared fit of all observed pair-wise distances to the expected distance of a tree. The ME method also seeks the tree with the minimum sum of branch lengths. But instead of using all the pair-wise distances as FM, it fixed the internal nodes by using the distance to external nodes and then optimizes the internal branch lengths FM and ME methods perform best in the group of distance-based methods, but

they work much more slowly than NJ, which generally yield a very close tree to these methods (Baxevanis and Ouellette, 2001).

4.2 Character-Based Methods (cladistic)

Distance-based methods are more rapid and less computationally intensive than character based methods, but the actual characters are discarded once the distance matrix is derived. On the other hand, character-based methods make use of all known evolutionary information, i.e. the individual substitutions among the sequences, to determine the most likely ancestral relationships.

4.2.1 Maximum parsimony (MP)

The criterion of MP method is that the simplest explanation of the data is preferred, because it requires the fewest conjectures. By this criterion, the MP tree is the one with fewest substitutions/evolutionary changes for all sequences to derive from a common ancestor. For each site in the alignment, all possible trees are evaluated and are given a score based on the number of evolutionary changes needed to produce the observed sequence changes. The best tree is thus the one that minimized the overall number of mutation at all site. MP works faster than ML and the weighted parsimony schemes can deal with most of the different models used by ML. However, this method yields little information about the branch lengths and suffers badly from long-branch attraction, which is the long branches have become artificially connected because of accumulation of inhomogous similarities, even if they are not at all phylogenetically related. MP yields more than one tree with the same score.

4.2.2 Maximum Likelihood (ML)

Like MP methods, ML method also uses each position in an alignment and evaluates all possible trees. It calculates the likelihood for each tree and seeks the one with the maximum likelihood.

For a given tree, at each site, the likelihood is determined by evaluating the probability that a certain evolutionary model has generated the observed data. The likelihood's for each site are then multiplied to provide likelihood for each tree.

ML method is the slowest and most computationally intensive method, though it seems to give the best result and the most informative tree (https://guava.physics.uiuc.edu/~nigel/courses/598BIO/498BIOonline-essays/hw2/files/hw2_li.pdf).

Table 02. Phylogenetic methods used with molecular sequence data (Christensen and Olsen, 2018).

Methods	Principle	Benefit	Drawback	Use and limitations	Options for extension
Neighbor Joining	Algorithmic	Simple and fast in relation to computational power	**Only one tree**	**Routine draft phylogeny**	**Bootstrap**
Maximum parsimony	Optimality criteria	Includes all informative positions of alignment and more trees are compared based on statistical criterion	**More trees can be "equally parsimonious"**	**Routine phylogeny of closely related sequences**	**Bootstrap**
Maximum likelihood	Optimality criteria	Includes all informative positions of alignment and more trees are compared based on statistical and probabilistic (likelihood criterion)	**With complex phylogenies the "maximum likelihood" tree will never be found**	**Routine phylogeny with fewer sequences**	**Bootstrap and likelihood ratio tests**
Bayesian (MrBayes)	Optimality criteria	Includes all informative positions of alignment and more trees are compared based on statistical and probabilistic	**Very complex phylogenies are not fully resolved**	**For complex phylogenies but not too complex**	

5. Sequence alignment

Sequence alignment is used to find out degrees of similarity between two (pairwise alignment) or more nucleic acid sequences of DNA or RNA and amino acid sequences of proteins.

It's the procedure of comparing two (pair-wise alignment) or more (multiple sequences) by searching for a series of individual characters or patterns that are in the same order in the sequences.

Pairwise alignment can be either global or local:

In global alignment, an attempt is made to align the entire sequence.

- If two sequences have approximately the same length and are quite similar, they are suitable for the global alignment.
- Suitable for aligning two closely related sequences Local alignment concentrates on finding stretches of sequences with high level of matches.

Local alignment concentrates on finding stretches of sequences with high level of matches

- Finds local regions with the highest level of similarity between the two sequences and aligns these regions without considering the alignment of rest of the sequence regions
- Suitable for aligning more divergent sequences
- Used for finding out conserved patterns in DNA or protein sequences.

5.1 Importance of sequence alignment in bioinformatics

By finding similarities between sequences;

1. Scientists can infer the function of newly sequenced genes by aligning the newly sequenced genes with sequences already in the database.

2. Predict new members of gene families.

3. Discovering evolutionary relationships or reconstruction of phylogeny to find whether two (or more) genes or proteins are evolutionarily related to each other in closely related species.

4. It can be used to predict the location and function of protein-coding and transcription-regulation regions in genomic DNA. Regulatory regions in genome are often conserved therefore presence of such conserved regions easily tells us the regulatory sites in the newly sequenced genes.

5. To find structurally or functionally similar regions within proteins

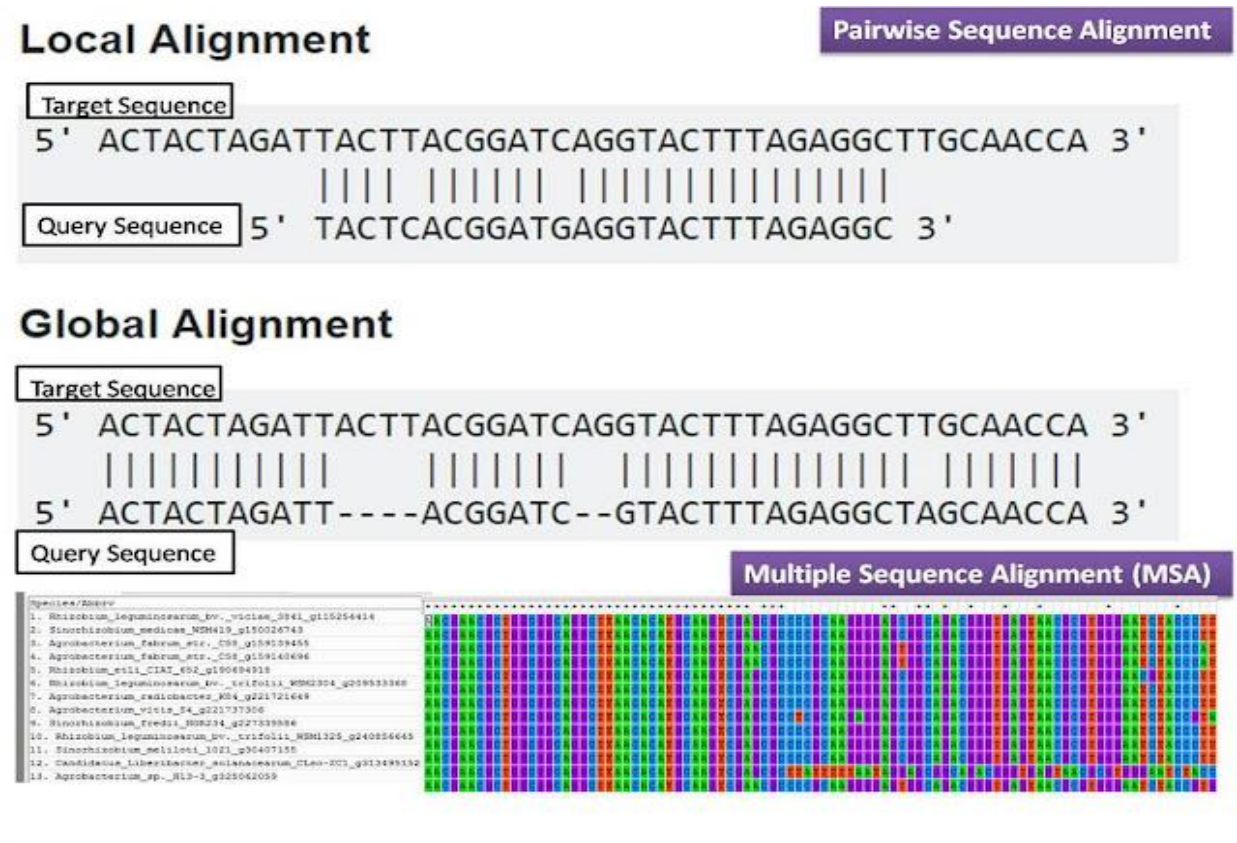

Figure 18. Difference between Pairewise alignment and Multiple Sequence Alignment
https://www.biologyexams4u.com/2016/05/importance-of-sequence-alignment-in.html

5.2 Pairwise Alignment VS Multiple Sequence Alignment

Table 03. Comparaison between Pairewise alignment and Multiple Sequence Alignment

Pairwise Alignment	Multiple Sequence Alignment (MSA)
An alignment procedure comparing two biological sequences of either protein, DNAor RNA.	An alignment procedure comparing three or more biological sequences of either protein, DNA or RNA.
Pairwise alignments can be generally categorized as global or local alignment methods.	MSA is generally a global multiple sequence alignment
Comparatively simple algorithm is used	Complex sophisticated algorithm is used
A general global alignment technique is the Needleman–Wunsch algorithm. A general local alignment method is Smith–Waterman algorithm.	A technique called progressive alignment method is employed. In this approach, a pairwise alignment algorithm is used iteratively, first to align the most closely related pair of sequences, then the next most similar one to that pair, and so on.

Applications: a) Primarily to find out conserved regions between the two sequences. b)Similarity searches in a database	**Applications:** a) To detect regions of variability or conservation in a family of proteins. b) Phylogenetic analysis (inferring a tree, estimating rates of substitution, etc.) c) Detection of homology between a newly sequenced gene and an existing gene family prediction of protein structure. d) Demonstration of homology in multigene families.
Examples of pairwise alignment tools: LALIGN BLAST EMBOSS Needle EMBOSS Water	**Examples of Multiple Sequence Alignment tools:** MUSCLE T-Coffee MAFFT CLUSTALW

6. Bioinformatics Tools for Phylogenetic Analysis

There are several bioinformatics tools and databases that can be used for phylogenetic analysis.

These include PANTHER, P-Pod, PFam, TreeFam, and the PhyloFacts structural phylogenomic encyclopedia. Each of these databases uses different algorithms and draws on different sources for sequence information, and therefore the trees estimated by PANTHER, for example, may differ significantly from those generated by P-Pod or PFam. As with all bioinformatics tools of this type, it is important to test different methods, compare the results, then determine which database works best (according to consensus results) for studies involving different types of datasets (https://www.biologyexams4u.com/2016/05/importance-of-sequence-alignment-in.html)

7. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2

The full-length genomic sequences and protein-coding sequences (CDSs) of 102 SARS-CoV-2, 5 SARS-CoV, 2 MERS-CoV, and 11 bat coronaviruses were downloaded from NCBI Severe acute respiratory syndrome coronavirus 2 data hub (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) and GenBank among genomic sequences of

SARS-CoV and bat coronaviruses, those showing high similarity with genomic sequences of SARS-CoV-2 were chosen. The integrity of sequences was checked, and the fragmented sequences were reconstructed. Finally, the datasets were constructed by labeling the sequences with the region of sampling and collection date.

7.1 Construction of phylogenetic tree with full-length genomic sequences

The full-length genomic sequences of 120 coronaviruses were aligned using the L-INS-i method of MAFFT v7.31023. Aligned sequences were converted into phylip file format by Clustal W24. Maximum likelihood (ML) trees based on full-length genomic sequences were constructed and estimated by PhyML program version 3.025 with 100 bootstraps resampling. The phylogenetic trees were visualized by FigTree v1.4.4.

7.2 Construction of phylogenetic supertrees

The matrix representation with parsimony (MRP)9,26 pseudo-sequence supertree22 was built.

Firstly, ten groups of CDSs for orthologous proteins in selected coronaviruses were organized using the OrthoMCL program27, with repeated sequences removed from the orthologous groups. The CDSs of 120 coronaviruses were assigned to their corresponding orthologous protein groups by custom-made scripts, and aligned by MAFFT23 with the L-INS-i method, followed with formation into phylip file by Clustal W24.

Secondly, ML phylogenies by using PhyML25 were employed to build source phylogenetic trees based on each CDSs, with 100 bootstrap replications. Thirdly, the members of each clade making up the selected bipartitions (above 55% bootstrap support) are assigned an A or T, and custom-made scripts were applied to retrieve the Baum-Ragan matrix pseudo-sequences. Fourthly, The pseudo-sequences of the coronaviruses were used to re-construct the phylogenetic supertree using P hyML25. The A/T substitutions were treated equally in the analysis, without systematic bias imported.

In addition, published supertree software Clann (version 4.2.4) was also used to construct traditional MRP supertree (with PAUP* version 4.0a16628) and MSSA (most similar supertree method) supertree, with default parameter settings29. L.U.St package version 2.030 was used to construct an approximated maximum likelihood supertree (Li and al., 2020).

In figure 19, the hosts and sampling locations of animal coronaviruses are enclosed in parentheses. The coding of SARS-CoV-2 viruses is the combination of the abbreviation of sampling location, sampling time, and Genbank accession. MERS-CoV clade, SARS-CoV clade,

and nine clades of SARS-CoV-2 are highlighted and labeled, respectively. The numbers along the branches mark the bootstrap values percentage out of 1000 bootstrap resamplings.

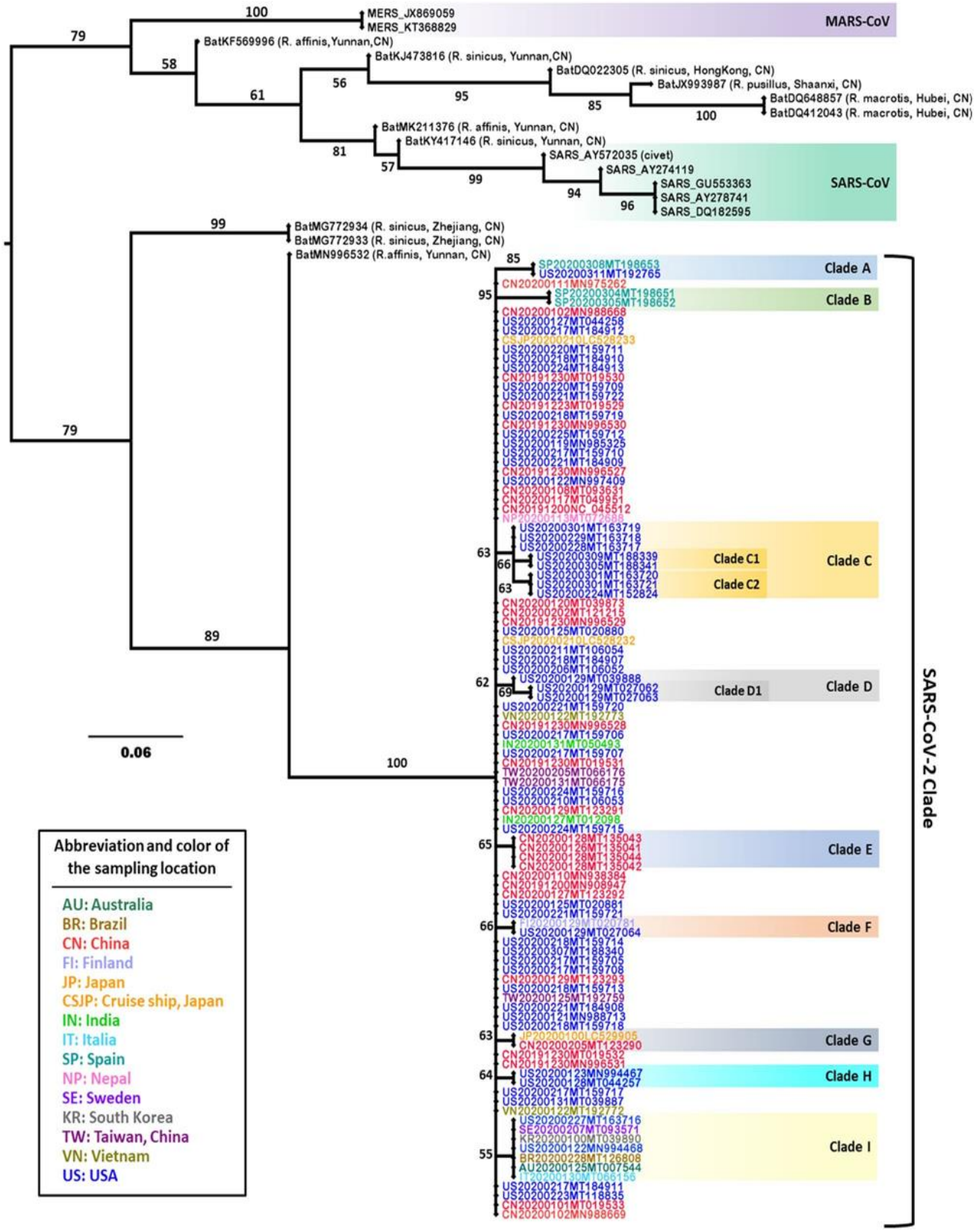

Figure 19. MRP pseudo-sequence supertree for SARS-CoV-2 constructed from protein source trees (Li and *al.*, 2020).

CONCLUSION

Due to the recent progress in virology, molecular biology, and pharmacology fields, we were quickly able to dissect and understand the COVID-19 causing virus structure, functions, lifecycle, and pathophysiological characteristics.

In this study, we discussed different types, obstacles and opportunities facing the application of next generation sequencing technologies for the diagnosis, surveillance, and study of SARSCoV-2 and other infectious diseases. At the early pandemic, Algeria has struggled on sequencing it sars-cov-2 samples, thing that required asking for support from some European countries (according to the Pasteur institute, sequencing is done in Algeria since the beginning of 2021).

We have seen how the supertree method is a powerful approach applied in the phylogenetic analysis of coronavirus. The distinct phylogenetic distance in the SARS-CoV-2 clade only can be detected by MRP pseudo-sequence supertree. Timely monitoring of the variation and evolution of SARS-CoV-2s would be favorable to treatment and control of COVID-19 and prevent its future outbreak.

REFERENCES

Altamimi, A., Ahmed, E.A. (2020). Climate factors and incidence of Middle East respiratory syndrome coronavirus. *Journal of Infection and Public Health*, **13**; 704-708. https://doi.org/10.1016/j.jiph.2019.11.011

Aryal, S. How to construct a Phylogenetic tree? . Microbe.notes [enligne]. February 4, 2019. Available from: https://microbenotes.com/how-to-construct-a-phylogenetic-tree/ accessed october 05,2021.

Babb de Villiers, C., Blackburn, L., Cook, S., Janus, J, Johnson, E and Kroese, M., (2021). Next generation sequencing for SARS-CoV-2. FIND 2021.

Baum, D. (2008). Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. *Nature Education*, 1(1):190.

Baxevanis, A. D and Ouellette, B.F. F. (2001). Bioinformatics: a practical guide to the analysis of genes and proteins- 2nd edition. *WILEY INTERSCIENCE* – 495P.

Bidar, B., Canco, P., Henry, C., Philipe, J M and *al*. (2020). Préparation au risque épidémique COVID – 19. Fédération Hospitalière de France (FHF). Available from: https://solidarites-sante.gouv.fr/IMG/pdf/guide-covid-19-phase-epidemique-v15-16032020.pdf

Bruno, W.J., Socci, N.D., Halpern, A.L. (2000). Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction, *Molecular Biology and Evolution*, 17; 1, 189–197- https://doi.org/10.1093/oxfordjournals.molbev.a026231

Cascella M, Rajnik M, Aleem A, et *al*. Features, Evaluation, and Treatment of Coronavirus (COVID-19) [Updated 2021 Jul 30]. In: *StatPearls* [Internet]. Publishing; 2021 Jan-. Available from:https://www.ncbi.nlm.nih.gov/books/NBK554776/?fbclid=IwAR2nAr0h_G5RgKy9tXgJbiNsDq_7k6G2EVdvBQvOW3CgcC9UHWVJuZPFPSI

Chilamakuri, R.; Agarwal, S. (2021). COVID-19: Characteristics and Therapeutics. *Cells*,10 ; 206. Available from : https://doi.org/10.3390/cells10020206

Christensen H., Olsen J.E. (2018) Short Introduction to Phylogenetic Analysis of Molecular Sequence Data. In: Christensen H. (eds). *Introduction to Bioinformatics in Microbiology. Learning Materials in Biosciences.* Springer, Cham. https://doi.org/10.1007/978-3-319-99280-8_6

Fisher, D., Heymann, D. (2020). Q&A: The novel coronavirus outbreak causing COVID-19. BMC Med **18,** 57. https://doi.org/10.1186/s12916-020-01533-w

Frese, S.K., Katus, H.A and Meder, B. (2013). Next-Generation Sequencing: From Understanding Biology to Personalized Medicine : A review. *Biology*. **2**, 378-398; doi:10.3390/biology2010378

John, G., Sahajpal,N.S., Ashis K. Mondal,A.K., Ananth, S., Williams, C.., Chaubey,A.,Rojiani,A.M and Kolhe, R. (2021). Next-Generation Sequencing (NGS) in COVID-19: A Tool for SARS-CoV-2 Diagnosis, Monitoring New Strains and Phylodynamic Modeling in Molecular Epidemiology. *Curr. Issues Mol. Biol*. 43, 845–867. https://doi.org/10.3390/cimb43020061

Hewlett Packard Enterprise. *WHAT IS NEXT GEN SEQUENCING ?* [Internet]. Available from : https://www.hpe.com/fr/fr/what-is/next-gen-sequencing.html. Accessed october 05,2021.

Institut Pasteur : *Covid-19 desease (NOVEL CORONAVIRUS).* [Updated 2021 Jan 13]. Internet]. Available from: https://www.pasteur.fr/fr/centre-medical/fiches-maladies/coronavirus-wuhan Accessed september 01,2021

Instiut pasteur. *INSTITUT PASTEUR SEQUENCES THE WHOLE GENOME OF THE CORONAVIRUS, 2019-NCOV*. Puplishing Jan 30,2021. [Internet]. Available from : https://www.pasteur.fr/en/press-area/press-documents/institut-pasteur-sequences-whole-genome-coronavirus-2019-ncov Accessed october 11,2021.

Importance of Sequence Alignment in Bioinformatics biology. (undated). exams4u.com [enligne]. Available from: https://www.biologyexams4u.com/2016/05/importance-of-sequence-alignment-in.html ; Last access october 13,2021.

Gupta, A.K and Gupta, U.D. (2014). Chapter 19 - Next Generation Sequencing and Its Applications, *Animal Biotechnology*, Elsevier Inc.

Gkazi, A.S. (2021). An Overview of Next-Generation Sequencing. In Genomics Research [Internet]. March 17, 2021. Available from: https://www.technologynetworks.com/genomics/articles/an-overview-of-next-generation-sequencing-346532 . Accessed october 02, 2021

Li, T., Liu, D., Yang, Y., Guo, J., Feng, Y., Zhang, X., Cheng, S and Feng, J. (2020). Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. *Sci Rep* **10,** 22366 https://doi.org/10.1038/s41598-020-79484-8

Mardis, E.R. (2008). Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*. 9; 1, 387-402

Nakamura, K., Oshima,T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa Y, Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, Md., Ogasawara, N., Kanaya, E. Sequence-specific error profile of Illumina sequencers. , *Nucleic Acids Research*, 39; 90. https://doi.org/10.1093/nar/gkr344

NCBI. Available from : https://www.ncbi.nlm.nih.gov/labs/virus/vssi/# last access October 13,2021

OMS. Lancement du réseau de laboratoires de séquençage du génome de COVID-19 en Afrique. (Sep 10,2020). Available from : https://www.afro.who.int/fr/news/lancement-du-reseau-de-laboratoires-de-sequencage-du-genome-de-covid-19-en-afrique

Parasher A. (2020). COVID-19: Current understanding of its pathophysiology, clinical presentation and treatment. *Postgrad Med*. 0 ; 1-9. Available from : doi:10.1136/postgradmedj-2020-138577

Ronaghi, M. (1998). A Sequencing Method Based on Real-Time Pyrophosphate. *Science* 281(5375):363, 365. doi:10.1126/science.281.5375.363

Shereen, A M., Khan, S., Kazmi, A., Bashir, N., Siddique, R. (2020). COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*, 24 ; 91-98.

Sanchari Sinha Dutta, Ph.D. What is Phylogenetic Analysis?. NWES MEDICAL LIFE SCIENCES. Updated: Mar 9, 2021. Available from: https://www.news-medical.net/health/What-is-Phylogenetic-Analysis.aspx. Last access: October 13,2021.

Tariq, A., Bhat, Ph.D., Maciej L., Goniewicz, Ph.D., Pharm.D., Yasmin M. Thanavala, Ph.D et *al.* (2020). SARS-CoV-2 Viral Load in Upper Respiratory Specimens of Infected Patients *N Engl J Med*; **382**:1177-1179. DOI: 10.1056/NEJMc2001737

Van Dijk, E., Thermes, C. (2021). La révolution de la génomique : les nouvelles méthodes de séquençage et leurs applications. Planete Vie [Internet]. Available from : https://planet-vie.ens.fr/thematiques/manipulations-en-laboratoire/la-revolution-de-la-genomique-les-nouvelles-methodes-de . Accessed : october 10,2021.

Van Doremale, N., Bushmaker, T., Morris, DH et al. (2020). Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. N Engl J Med, 382 (16) (published online March 17.) doi:10.1056/NEJMc2004973

Voelkerding, K.V., Dames, S.D., Durtschi, J.D. (2009) Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, 55, 641–658, https://doi.org/10.1373/clinchem.2008.112789

Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., & Tan, W. (2020). Detection of SARS-CoV-2 in Different Types of Clinical Specimens. JAMA, 323(18), 1843–1844. https://doi.org/10.1001/jama.2020.3786

Yohan J. (2021). Le séquençage, une histoire de générations. Bioinfo-fr. publishing : jun 05,2012. Available from : https://bioinfo-fr.net/le-sequencage

Yuki, K., Fujiogi, M and Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. *Clinical immunology (Orlando, Fla.)*, *215* ; 108427. https://doi.org/10.1016/j.clim.2020.108427