SARS-CoV-2 virus genome structure and evolution

Chaimae SAMTAL¹, Nihal HABIB², Abdellah IDRISSI AZAMI², Sofia SEHLI², Bouchra CHAOUNI³, Mohamed BEKKALI⁴, Hicham BEKKARI⁵, Salsabil HAMDI⁶, Lahcen WAKRIM⁶, Najib Al IDRISSI⁶, Hassan GHAZAL⁷*

¹ Faculty of Sciences Dhar El Mehraz, USMBA, Fes, Morocco
² Faculty of Medicine and Pharmacy of Rabat, Rabat, Morocco
³ Faculty of Sciences of Rabat, Rabat, Morocco
⁴ Department of Genetics, Faculty of Sciences, University of Granada, Spain.
⁵ Environmental Health Laboratory Institut Pasteur du Maroc, Casablanca, Morocco
⁶ Virology Unit, Immunovirology Laboratory, Institut Pasteur du Maroc, Casablanca, Morocco.
⁷ School of Medicine, University Mohamed VI for Health Sciences, Casablanca, Morocco
⁸ Centre National de la Recherche Scientifique et Technique, Rabat, Morocco

*Corresponding Author E-mail: *hassan.ghazal@fullbrightmail.org

1. Introduction

Coronaviruses (CoVs), a large family of single-stranded RNA viruses, can infect animals and humans, causing respiratory, gastrointestinal, hepatic, and neurologic diseases (Weiss and Leibowitz, 2011). As the largest known RNA viruses, CoVs are further divided into four genera: α-Coronaviruses, β-Coronavirus, γ-Coronaviruses a
nd δ-Coronaviruses (Yang and Leibowitz, 2015). To date, six human coronaviruses (HCoVs) have been identified, including the alpha-CoVs HCoVs-NL63 and HCoVs-229E and the beta-CoVs HCoVs-OC43, HCoVs-HKU1, severe acute respiratory syndrome-CoV (SARS-CoV) (Drosten et al., 2003), and Middle East respiratory syndrome-CoV (MERS-CoV) (Zaki et al., 2012). New coronaviruses appear to emerge periodically in humans, mainly due to the high prevalence and wide distribution of coronaviruses, the large genetic diversity and frequent recombination of their genomes, and the increasing of the human-animal interface activities (Cui et al., 2019; Zhu et al., 2020).

However, they were not considered to be highly pathogenic to humans until the outbreak of severe acute respiratory syndrome (SARS) in 2002 and 2003 in Guangdong province, China (Drosten et al., 2003; Zhong et al., 2003), as the coronaviruses that circulated before that time in Humans mostly caused mild infections in immunocompetent people. Ten years after SARS, another highly pathogenic coronavirus, the Middle East respiratory syndrome coronavirus (MERS-CoV) emerged in Middle Eastern countries (Zaki et al., 2012). SARS coronavirus (SARS-CoV) uses angiotensin-converting enzyme 2 (ACE2) as a receptor and primarily infects ciliated bronchial epithelial cells and type II pneumocytes, whereas MERS-CoV uses dipeptidyl peptidase 4 (DPP4; also known as CD26) as a receptor and infects non ciliated bronchial epithelial cells and type II pneumocytes. SARS-CoV and MERS-CoV were transmitted directly to Humans from market civets and dromedary camels, respectively, and both viruses are thought to have originated in bats (Cui et al., 2019; Guan et al., 2003).

In late December 2019, several local health authorities reported clusters of patients with pneumonia of unknown cause, which were epidemiologically linked to a seafood market in Wuhan, Hubei Province, China (Zhu et al., 2020). The pathogen, a novel coronavirus (SARS-CoV-2) was identified by local hospitals using a surveillance mechanism for “pneumonia of unknown etiology” that was established in the wake of the 2003 SARS outbreak with the aim of allowing timely identification of novel pathogens(Q. Li et al., 2020; Zhu et al., 2020). On 30 January 2020, the World Health Organization (WHO) declared that CoVID-19 is a “public-health emergency of international concern” (X. Li et al., 2020). The pandemic is escalating rapidly.

### 2. Biology and Genomics

A Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV 2) appeared in the late December 2019 (Ye et al., 2020), this latter is a new member of the β-Coronavirus (βCoV) genus, Coronaviridae family which is a family of enveloped retroviruses (Park et al., 2020), the genome size of Coronaviridae species goes from 26 kb to 32 kb (Ye et al., 2020), as for the SARS-CoV 2, it has a genome size of 29.890 kb (Castillo et al., 2020), and it codes for 9744 amino acids (aa) (C. Wang et al., 2020), 75 aa make the envelope proteins, 222 aa make the membrane proteins, 419 aa for the nucleoproteins, 1273 aa for the spike as shown in Table 1 and the rest for 10 ORFs (C. Wang et al., 2020) (figure 1).

**Figure 1:** The composition of the SARS-CoV-2: The single-stranded RNA genomes of SARS-CoV encode two large genes, the ORF1a and ORF1b genes. The structural genes encode the structural proteins, spike (S), envelope (E), membrane (M), and nucleocapsid (N), which are common features to all coronaviruses. (Zhou et al., 2020)
Coronaviruses have non segmented genomes, 2/3 of their genomes contain large overlapping ORFs which are ORF-1a and ORF-1b as shown in figure 2, those regions code for replicase polyproteins (Ye et al., 2020).

Figure 2: Schematic representation of the genome organization of different Coronaviruses. The single-stranded RNA genomes of SARS-CoV and MERS-CoV encode two large genes, the ORF1a and ORF1b genes, which encode 16 non-structural proteins (nsp1–nsp16) that are highly conserved throughout coronaviruses. The structural genes encode the structural proteins, spike (S), envelope (E), membrane (M), and nucleocapsid (N), which are common features to all coronaviruses. The accessory genes (shades of green) are unique to different coronaviruses in terms of number, genomic organization, sequence, and function. The structure of each S protein is shown beneath the genome organization. The S protein mainly contains the S1 and S2 subunits. The residue numbers in each region represent their positions in the S protein of SARS and MERS, respectively. The S1/S2 cleavage sites are highlighted by dotted lines. SARS-CoV, severe acute respiratory syndrome coronavirus; MERS-CoV, Middle East respiratory syndrome coronavirus; CP, cytoplasm domain; FP, fusion peptide; HR, heptad repeat; RBD, receptor-binding domain; RBM, receptor-binding motif; SP, signal peptide; TM, transmembrane domain. (“La saga de COVID-19,” n.d.)

Coronaviruses are divided into 4 genera α-Coronaviruses, β-Coronaviruses, γ-Coronaviruses and δ-Coronaviruses (Guo et al., 2020), α-CoVs and β-CoVs are able to infect mammals (Guo et al., 2020). In fact, 5 of the human CoVs belong to the β-Coronaviruses (Ye et al.,

### Table 1. Structural proteins of coronaviruses and their function (Schoeman et al., 2019)

| Nucleoprotein or the N protein | 419 amino acid | -Binds to the CoV RNA genome, making up the nucleocapsid
| -Involved in the viral replication cycle and the cellular response to viral infections |
| Spike S1 | 1273 amino acid |
| -Facilitates the binding of the envelope viruses to the host cell by attracting the ACE2 receptor |
| S2 | -Mediates the virus fusion into the host cell |
| Matrix | 222 amino acid |
| -Plays a role in determining the shape of the virus envelope |
| -Binds to the other proteins in order to stabilize them |
| E protein | 75 amino acid |
| -Engages in the production and maturation of the virus |

Additionally, the ORF1 and ORF2 encodes for non-structural proteins of the coronavirus (Table 2).

### Table 2. Nonstructural proteins of coronaviruses and their function (Atsuti et al., 2020)

| nsp 1 and 3 | Inhibition of IFN signaling and blocking of host innate immune response by promotion of cellular degradation and blocks translation of host's RNA |
| nsp2 | Binding to prohibition protein |
| Nsp 3 and 5 | Promoting cytokine expression and cleavage of viral polyprotein |
| Nsp 4 and 6 | Contribute to structure of DMVs as transmembrane scaffold protein (DMVs formation) |
| Nsp 7/8 complex | Processivity clamp for RNA polymerase by arms hexadecameric complex |
| Nsp 9 | RNA binding protein phosphatase |
| Nsp 10, 16 and 14 | Stimulation of ExoN and 2-0-MT activity |
| Nsp 12 | Replication enzyme (RNA-dependent RNA polymerase) |
| Nsp 13 | RNA helicase, 5′ triphosphatase |
| Nsp 14 | Proofreading of viral genome |
| Nsp 15 | Viral endoribonuclease and chymotrypsin-like protease |
| Nsp 16 | Avoiding MDA5 recognition and inhibit innate immunity regulation |

Coronaviruses have non segmented genomes, 2/3 of their genomes contain large overlapping ORFs which are ORF-1a and ORF-1b as shown in figure 2, those regions code for replicase polyproteins (Ye et al., 2020).
ACE2 in the membrane-bound Angiotensin converting enzyme decreased in Wuhan after January 2020 (Tang et al., n.d.). Likely evolutionary (30%) (Tang et al., 2020) shows the existence of 2 major types of this virus, the L (it has a ‘CT’ haplotype at location 8782 and 24,144) which is more relevant (70%) and the S (it has a ‘TC’ haplotype at location 8782 and 24,144) which is a little bit rare (30%) (Tang et al., n.d.). In spite of that, the S type is evolutionary older than the L one, which makes it more likely the ancestor version, however, the L one was more found in the early stage of the outbreak, its frequency decreased in Wuhan after January 2020 (Tang et al., n.d.).

3. Integration of the viral genome into the host cell

Angiotensin converting enzyme 2 (ACE2) is a membrane-bound aminopeptidase that has a vital role in the cardiovascular and immune systems. ACE2 is involved in heart function and the development of hypertension and diabetes mellitus and other diseases (Figure 3). In addition, ACE2 has been identified as a functional receptor for coronaviruses (Turner et al., 2004). Enveloped viruses enter cells by binding their envelope glycoproteins to cell surface receptors followed by conformational changes leading to membrane fusion and delivery of the genome in the cytoplasm (Phogat and Dimitrov, 2003).

Results in the Cryo-EM experiments demonstrated that SARS-CoV-2 had a ten times higher affinity to ACE2 comparing to SARS-CoV, which was consistent with the higher efficiency of infection of SARS-CoV-2. These findings indicated that ACE2 might be crucial for the human infection of SARS-CoV-2 and for the progression and prognosis of COVID-19. Understanding the expression and activity of ACE2 under different physiological and pathological conditions may help to predict the susceptibility of SARS-CoV-2 in different cohorts of people and the clinical outcomes and prognosis of COVID-19 patients (Y. Li et al., 2020).

Unlike the ACE gene, which is located on human chromosome 17, the 40kb ACE2 gene is located on...
chromosome Xp22 and contains 18 exons, most of which resemble exons in the ACE gene. Whereas somatic ACE contains two active sites, ACE2 possesses only a single catalytic domain. Both ACE and ACE2 act as zinc metallopeptidases but of differing substrate specificities defining their distinct and counterbalancing roles in the RAS. Whereas ACE cleaves C-terminal dipeptide residues from susceptible substrates (a peptidyl dipeptidase), ACE2 acts as a simple carboxypeptidase able to hydrolyze Ang I, forming Ang 1–9 and Ang II to Ang 1–7 (Figure 4). ACE2 does not cleave bradykinin, further distinguishing its specificity from that of ACE while it is also insensitive to conventional ACE inhibitors. The C-terminal domain of ACE2 which has no similarity with ACE, is a homolog of a renal protein, collectrin, which regulates the trafficking of amino acid transporters to the cell surface, endowing ACE2 with multiple and distinctive physiological functions. It is the multiplicity of physiological roles that ACE2 plays that has allowed it to be hijacked by SARS-CoV-2 as a receptor, resulting in the COVID-19 pandemic. Structural studies have revealed the structures of both the SARS-CoV and much more recently, the SARS-CoV-2 in complex with ACE2 (Figure 4). In the case of SARS-CoV-2, the major spike glycoprotein (S1) binds to the N-terminal region of ACE2. The knowledge of the biology and physiology of ACE2 accumulated over the last 20 years since its discovery should provide a major stimulus to understanding some of the key steps in SARS-CoV-2 infection and its ultimate prevention (Gheblawi et al., 2020).

Figure 4: ACE2 has an extracellular facing N-terminal domain and a C-terminal transmembrane domain with a cytosolic tail. The N-terminal portion of the protein contains the claw-like protease domain (PD), while the C-terminal domain is referred to as the Collectrin-like domain (CLD). The receptor-binding domain (RBD) of SARS-CoV-2 binds with the PD of ACE2, forming the RBD-PD complex distinct from the ACE2 catalytic site (Gheblawi et al., 2020)

The binding affinity of SARS-CoV-2 with ACE2 appears to be stronger than SARS-CoV, due to alterations in several amino acid residues allowing for enhanced hydrophobic interactions and salt bridge formations, which may explain the considerably larger global influence of COVID-19 than the initial SARS (Shang et al., 2020). Moreover, SARS-CoV-2 has evolved to utilize a wide array of host proteases including cathepsin L, cathepsin B, trypsin, factor X, elastase, furin, and transmembrane protease serine 2 (TMPRSS2) for S-protein priming and facilitating cell entry following receptor binding. So far, TMPRSS2 and cathepsin L/B mediates S-protein priming of SARS-CoV-2, and camostat mesylate, a serine protease inhibitor combined with cathepsin L/B inhibitor, E-64d blocked SARS-CoV-2 entry.

The entry of both SARS-CoV and SARS-CoV-2 into cells is facilitated by the interaction between viral S-protein with extracellular domains of the transmembrane ACE2 proteins, followed by subsequent downregulation of surface ACE2 expression (Figure 5). In a cohort of 12
COVID-19 patients, circulating Ang II levels were markedly elevated compared to healthy controls (linearly correlated with viral load), providing a direct link between tissue ACE2 downregulation with systemic RAS imbalance, and facilitating the development of multi-organ damage from SARS-CoV-2 infections. Potential therapeutic strategies may include preventing the binding of human ACE2 and SARS-CoV-2 by blocking the receptor-binding domain (RBD) of the viral S-protein. In addition to this RBD blocking strategy, other possible treatment options may include localized use of ACE2-derived peptides, small molecule inhibitors, ACE2 antibody or single chain antibody fragment against ACE2 (Gheblawi et al., 2020).

After the endocytosis of the enzyme and the SARS-CoV-2 viral particles, the ACE2-mediated cardiovascular protection is lost Ang II levels grows with increased activity of angiotensin 1 receptors (AT1R) at the expense of ACE2/Ang 1–7 driven pathways leading to adverse fibrosis, hypertrophy, increased reactive oxygen species (ROS), vasoconstriction, and gut dysbiosis. ADAM17 mediated proteolytic cleavage of ACE2 is upregulated by endocytosed SARS-CoV-2 spike proteins. The activation of the AT1R by elevated Ang II levels which leads to furthermore ADAM17 activity. Similarly, the ADAM17 also cleaves its primary substrate releasing soluble TNF-α into the extracellular region where it has auto- and paracrine functionality. TNFα activation of its Tumor Necrosis Factor Receptor (TNFR) represents a third pathway rising the ADAM17 activity. Due to SARS-CoV-2 infection and together with comorbidities for instance diabetes and hypertension, TNF-α along with systemic cytokines are released which can lead to a cytokine storm (Figure 5) (Gheblawi et al., 2020)

Like SARS-CoV, SARS-CoV-2 use ACE2 that exist in lower respiratory tract as their receptor as shown in Figure 6 (Zhou et al., 2020), this receptor is integrated in the cross species transmission, and human to human transmission (Guo et al., 2020), the spike glycoprotein of the virus has 2 subunits, S1 and S2 (Guo et al., 2020), the S1 binds the virus to the host cell using the receptor-biding domain (RBD), while the S2 leads to the membrane fusion using heptad repeats (HR-1 and HR-2) (Xia et al., 2020), after that, the viral genome get released inside the host cytoplasm and it translates 2 polyproteins pp1a and pp1b that form the replication-transcription complex (RTC) in double-membrane vesicle (Guo et al., 2020), the formed RTC complex replicates continuously a set of sub-genomic RNAs that code for accessory and structural proteins (Guo et al., 2020), the new formed RNA, nucleocapsid and envelope proteins get assembled by the endoplasmic reticulum and Golgi in order to form the viral buds, the vesicles fuse with the cell membrane and release the new formed viruses (Guo et al., 2020).

**Figure 5**: Role of ACE2 in the pathogenesis of COVID-19 and the inflammatory response.
4. Phylogeny

The phylogenetic analysis of the isolated strains across the world show a similarity of 99.9% to 100% (C. Wang et al., 2020), the sequence alignment of many sequenced genomes showed a very rare mutation rate between different strains, the highest mutation rate is found in the ORF-1a and ORF-8 (Chan et al., 2020; Tang et al., n.d.; C. Wang et al., 2020). Yet the most common mutations were substitutions C > T in the position 8782 of ORF-1a and the position 29095 of the N gene and a substitution T > C in the position 28144 of the ORF-8, the last one causes amino acid to change in L84S in the ORF-8, and the first two were found outside of Wuhan (Chan et al., 2020).

A phylogenetic network was conducted during a study of the word SARS-CoV-2 strains using 160 complete genomes retrieved from GISAID database and using the closer Bat-CoV as an outgroup (Figures 6a). The network showed the existence of ancestral viral genomes alongside the newly mutated genomes (Forster et al., 2020).

The evolution study of the virus revealed that 87.6% to 95.6% of the non-synonymous mutations were removed by negative selection during the viral evolution, however 10 amino acids sites showed a high signals of positive selection evolution, three of them belong to the RBD of the spike (Tang et al., n.d.), this might explain the very high affinity of the RBD of SARS-CoV-2 to the human ACE2 regarding the SARS-CoV (Wrapp et al., 2020).

At a genomic level, the SARS-CoV-2 is closer to the Bat-CoV (96% of similarity) than the SARS-CoV (79.6% of similarity) (Figure7) (Zhou et al., 2020).
5. Conclusion

In summary, the origin of this virus that is responsible of COVID-19 pandemic, the deadliest infectious diseases to have emerged in recent history, is still unknown. Even with all the studies conducted around the world to understand the origins of the virus through decryption and studying its genome. Which shows that there is an urgent need for further studies that combine genomic, clinical and epidemiological data for better understanding of the origin, evolution and the impact of this virus.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References


Figure 7: Phylogenetic tree of different Coronaviruses (Zhou et al., 2020)