In-Silico Comparative Analysis of Egyptian SARS CoV-2 with Other Populations: a Phylogeny and Mutation Analysis

Lamis Sharawy*
Center for Informatics Sciences (CIS)
Nile University
Giza, Egypt.
lamis.yehia88@gmail.com

Maha Tantawy*
Center for Informatics Sciences (CIS)
Nile University
Egypt
Maha.tantawy@nu.edu.eg

Yara Ahmed
Center for Informatics Sciences (CIS)
Nile University
Egypt
Ya.Ahmed@nu.edu.eg

Ahmed Taha
Center for Informatics Sciences (CIS)
Nile University
Egypt
Ahm.Taha@nu.edu.eg

Omar Soliman
Faculty of Biotechnology
Misr University for Science and Technology
Egypt
https://orcid.org/0000-0002-6509-9895

Tamer M Ibrahim
Department of Pharmaceutical Chemistry, Faculty of Pharmacy
Kafrelsheikh University
Kafrelsheikh
https:/orcid.org/0000-0003-1016-8950

Mohamed El-Hadidi**
Center for Informatics Sciences (CIS)
Nile University
Giza, Egypt.
https://orcid.org/0000-0002-0360-3654

***Abstract*— In the current SARS-CoV2 pandemic, identification and differentiation between SARS-COV2 strains are vital to attain efficient therapeutic targeting, drug discovery and vaccination. In this study, we investigate how the viral genetic code mutated locally and what variations is the Egyptian population most susceptible to in comparison with different strains isolated from Asia, Europe and other countries in Africa. Our aim is to evaluate the significance of these variations and whether they constitute a change on the protein level and identify if any of these variations occurred in the conserved domain of the virus. The available Covid-19 complete genome nucleotide sequences on NCBI were gathered and filtered, and representative sequences were selected from each of the mentioned continents to make the population of our sample 1535 sequences. Multiple sequence alignment was conducted for all the 1535 sequences obtained from NCBI. For higher accuracy, we used the MAFFT iterative refinement method. Conserved domain extraction was carried out for all 1535 sequence for mutation evaluation. When the mutations were evaluated, Spike_D614G, NSP12_P323L, NS3_Q57H and N_R203K were found to be the most common amino acid substitutions among the viral isolates from Egypt. All retrieved mutations were processed and analyzed with principal component analysis (PCA). In general, no clear clusters were clustered based on the mutation pattern of different continents, including Africa, Asia, and Europe. However, PCA shows that the African mutation pattern is a partial subset of the complete European mutation pattern.**

Keywords—SARS-COV2, mutation, MAFFT, PCA, MSA

I. INTRODUCTION

The Coronaviridae family consists of many viruses that have genetic heterogeneity and that are differentiated into four genera: α-coronavirus, β-coronavirus, γ-coronavirus, and δ-coronavirus. The β-coronaviruses include SARS-CoV and MERS-CoV among others. SARS-CoV was defined in late 2002 and led to more than 8000 infected cases and 774 deaths. Bats act as a potential reservoir species. The 2012 MERS-CoV appeared in Saudi Arabia, where 35% of the infected cases were dead. Dromedary camels play a main role in viral transmission to humans. Recently, there is a new member of this genus that creates a lot of attention, SARS-CoV-2. It is a novel virus belonging to the β-coronavirus and is causing the current global pandemic [1][2].

Coronaviruses are large enveloped non - segmented positive - sense RNA viruses, generally cause enteric and respiratory diseases in animals and humans [3]. The first appearance of SARS-CoV-2 began in Wuhan-China, then it spread around the world. The novel virus was firstly named 2019-nCoV, thereafter, was called SARS-CoV-2 by World Health Organization (WHO)[2][4].

The genetic structure of SARS-CoV-2 consists of four main structural proteins: the spike (S) giving the crown appearance under electron microscopy, membrane (M), envelope (E), and the nucleocapsid (N) protein. The spike protein helps in fusion and viral entry by facilitating virus attachment to the host cell surface receptors. The viral shape is defined by a membrane protein. The envelope and nucleocapsid proteins assist in viral assembly and budding. With the spike protein's help, the virus attaches to the host cell receptor, thereafter, the viral replicase gene is translated, and the mature virus is formed by encapsidation [5][6].

The genome size of the SARS-CoV-2 is approximately 30 kb and its genomic structure has followed the characteristics of known genes of the coronavirus. The ORF1ab polyprotein is covering two-thirds of the viral genome and cleaved into many nonstructural proteins (nsp1 to nsp16). The third part of the SARS-CoV-2 genome codes for the main structural proteins (spike S, envelope E, nucleocapsid N and membrane M). Besides, six accessory proteins, encoded by, six ORFs, namely ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 which are predicted as hypothetical proteins [7].

(*) Equally contributed authors, (**) corresponding author, Email: melhadidi@nu.edu.eg.

According to WHO, the rate of death is 3.4% approximately, but the mortality number will be higher in patients with cancer (5.6%), Hypertension (6%), Chronic Respiratory disease (6.3%), Diabetes (7.3%) and Cardiovascular Disease (10.5%)[1]. Currently, there is no proven therapeutic medication for COVID-19, and conventional protocols, including cardiorespiratory ventilation support, are the main approach. Genomic analysis and comparative multiple sequences SARS-CoV2 could help developing effective therapeutic agents.

II. METHOD

A. Workflow

The pipeline for our project is described here, and detailed explanations will be presented in the following subsections. First, data was collected from NCBI (https://www.ncbi.nlm.nih.gov/). Second, data was prepared (renaming and segmentation). Third, Multiple Sequence Alignment (MSA) was used to align the sequences together. Fourth, the conserved domains were extracted from the aligned sequences to be studded. Fifth, the phylogenetic tree was constructed. Finally, all retrieved sequences were submitted to CoVsurver and PCA analysis was performed the output.

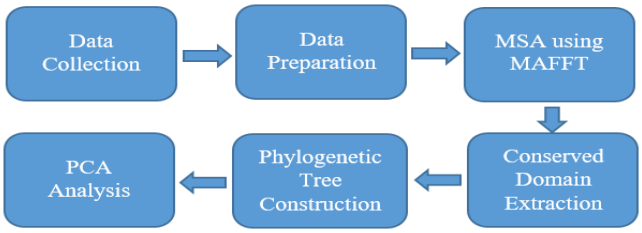


Figure 1 Block diagram for the work flow of our project.

Data Collection

Complete Nucleotides sequences of SARS-CoV2 were downloaded from Egypt and the surrounded countries in Africa, Europe and Asia (as the pandemic started). Table1 shows the number of sequences downloaded for each country in the previously mentioned continents. The process of accessing and retrieving data from NCBI is totally automated using the NCBI utility Entrez Direct (EDirect) Under Ubuntu 18 operating system.

TABLE I. NUMERIC DISTRIBUTION OF SARS-COV2 IN OUR DATABASE

Continent	Country
Africa	Zambia (1), Tunisia (8), South Africa (1), Nigeria (1), Morocco (10), Egypt (75)
Asia	Bahrain (9), Bangladesh (170), China (120), Georgia (8), Hong Kong (27), India (394) Iran (5), Israel (2), Kazakhstan (4), Malaysia (4) Nepal (1), Pakistan (3), Saudi Arabia (29) South Korea (6), Sri Lanka (4), Taiwan (32) Thailand (25), Timor-Leste (186), Turkey (34) Viet Nam (2)
Europe	Belgium (1), Czech Republic (23) Finland (1), France (84), Germany (59), Greece (98), Italy (14), Netherlands (16), Poland (31) Russia (2), Serbia (8), Spain (31), Sweden (1)
	Total no. of sequences = 1535

B. Data Preparation

After retrieving sequences from NCBI database, sequences name were renamed to the corresponding geographic location for better visualization and interpretation. The last step of data preparation involves pooling all sequence FASTA files into one file.

C. Multiple Sequence Alignment (MSA)

MSA plays an important role in evolutionary and functional analyses of biological sequencing, reconstructing phylogenetic trees, predicting 3D structure, and identifying the conserved regions.

MAFFT was chosen to be used here as the MSA tool (MAFFT-7.470-Linux64 version) due to its accuracy and speed compared to other alignment tool MAFFT (alignment using fast Fourier transform) offers various multiple alignment strategies. These strategies are classified into three types, (a) the progressive method, (b) the iterative refinement method with the weighted sum of pairs (WSP) score, and (c) the iterative refinement method using both the WSP and consistency scores. In our study, the iterative refinement method (FFT-NS-i) which is accurate, but slower than the other techniques, was used for the 1535 sequences. The max iteration was 1000, but actually, MAFFT reached its best results after just 16 iterations. Penalties for both gap opening and extension were set to the default values; 1.53 and 0.123 respectively. The phylogenetic tree was constructed using MAFFT online tool using Neigbor Jioning (NJ) method (https://mafft.cbrc.jp/alignment/server/phylogeny.html).

D. Conserved Domain Extraction

We used the Emboss cons tool (version 6.6.0) for conserved domain extraction. Cons calculates a consensus sequence from a multiple sequence alignment. For conserved region extraction, the used identity of the Emboss cons is 1533 for the best conserved region extraction.

E. Principle component (PCA) analysis

All retrieved sequences were submitted to CoVsurver (https://corona.bii.a-star.edu.sg/) to detect present mutation compared with the NCBI Refseq sequence (NC_045512.2) of complete SARS-COV2 genome using default parameters of CoVsurver. The retrieved mutations were processed using an in-house developed python script to generate a well-formatted table contain mutations from Africa, Asia and Europe. Moreover, this script searches for conserved mutations that appear in Africa, Asia and Europe and outputs some statistics about mutation frequency in each continent. The table exported from this script generates a table that was loaded into RStudio and submitted to prcomp function from stats V3.6.2 library to perform the PCA analysis.

III. Result & Discussion

In this study, 75 Egyptian SARS-COV-2 viral sequences were collected and uploaded to the NCBI database. The SARS-CoV-2 reference genome of Wuhan was used to detect mutations in the Egyptian genomes. We aimed to detect the most common variations of SARS-CoV-2 sequences isolated from EGYPT compared to the reference sequence from China (NC_045512.1). Among all mutations, Spike_D614G, NSP12_P323L, NS3_Q57H, N_R203K were found to be the most common amino acids substitutions among the vital isolates in Egyptian.

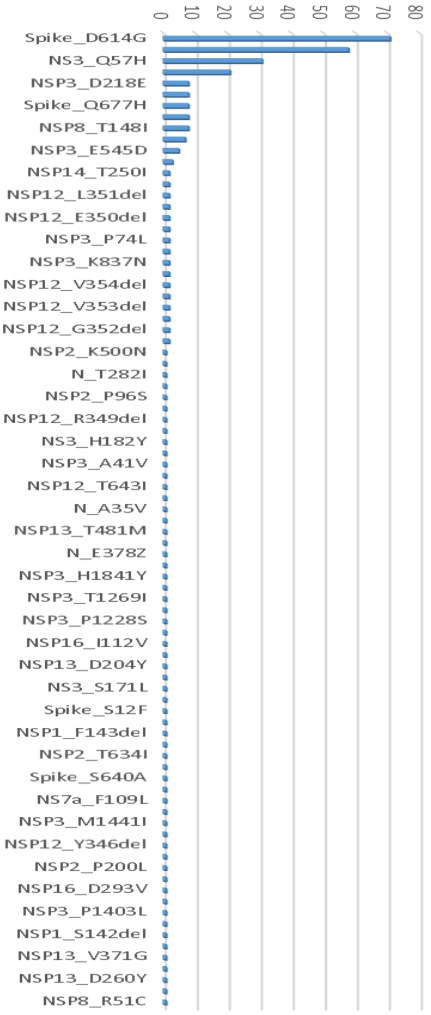


Figure 1. Mutation numbers for the Egyptian sequences

A. *D614G mutation:*

It's a variant founds on the surface of the spike protein promotor of SARS-CoV-2 codon 614 change Aspartic acid (D) with Glycine (G) by alteration of A-to-G nucleotide at position 23,403 in the Wuhan reference strain [8], beginning in Europe (Belgium) then moved together around the world as a consistent set of variations, in our study, we found 94% (n=71) D614G between Egyptian strains. This mutation is associated with potentially higher viral loads in COVID-19 patients by stabilizes the interaction between the S1 and S2 domains, limiting S1 shedding, but not with disease severity, moreover, D614G appears as part of a set of linked mutations in viral nsp3 and RdRp proteins [9].

B. *P323L mutation:*

Viral RNA synthesis is produced by nonstructural proteins (NSPs), an encoding region of the ORF1a and ORF1b viral polyproteins. RNA-dependent RNA polymerases (RdRps) (**also named NSP12**) are multi-domain proteins act as catalyze RNA-template dependent formation of phosphodiester bonds between ribonucleotides in the presence of divalent metal ion. NSP12 is high homology in SARS-CoV-2 then SARS-CoV, suggesting that its function and mechanism of action might be well conserved, this protein is vital for replication, pathogenesis and is considered as a crucial target for antiviral candidates in CoVs[10]. C14408T mutation on NSP12 is a missense mutation that leads to an amino acid change from proline to leucine at position 323 (P323L) in RNA polymerase protein, 77% (n=58) in our study. This mutation found commonly all over the world, mainly in Africa, then S. American and Europe in the third. The previous studies found that most of the isolates carry C14408T and A23403G (D614G in Spike glycoprotein) variations simultaneously in all continents[11]. RdRps act as primary targets for antiviral drug production. Many RdRp inhibitors have been considered to SARS-CoV-2 like Favipiravir, Galidesivir, Remdesivir and Ribavirin. So any mutation in the critical residues for drugs may lead to drug resistance and these molecules disable to bind to RdRp. Therefore, P323L mutations lead to significant changes in the protein secondary structure and are located in the interface domain (residues A250-R365) of the RdRp protein. This domain helps in the coordination of N and C terminal domains of RdRp [12]. The previous studies suggested that Simeprevir, Filibuvir and Tegobuvir bind to RdRp at a putative docking site (a hydrophobic cleft) that includes phenylalanine at 326th position. The mutation identified in our patients is very close to the docking site and the substitution of (P) amino acids to (L) at 323 might interfere with the interaction of these drugs with RdRp and also causing stabilization of the protein structure[13].

C. *NS3-Q57H mutation:*

Q57H (synonymous amino acid substitutions) is detected in ORF3a among 75 SARS-CoV-2 Egyptian strain in our research. Q57H is identified in 43% (n=31) of these strains. G25563T variation (NS3-Q57H) also causes the amino acid exchange of glutamine to histidine in residue 57 (Q57H), possibly converting it to a proteolytic cleavage site but perhaps leaving the corresponding viroporin function undisturbed. G25563T Orf3a is required for efficient in vitro and in vivo replication in SARS-CoV2, has been implicated in inflammasome activation, apoptosis, necrotic cell death and has been observed in Golgi membranes. The Q57H mutation in ORF3a protein might change important functional domains linked to virulence, infectivity, ion channel formation, and virus release leading to amino acid substitutions in the viral protein sequence [14]. The mutations are substitution mutations in the coding regions, resulting in amino acid sequence changes (missense mutation; non-synonymous mutations) and it is deleterious variant.

D. *N protein (R203k mutation):*

N protein is a 422 amino acid phosphoprotein that links the envelope to the +RNA genome and it consists of an N-terminal (NTD) and a C-terminal (CTD) domain. The CTD

allows the genome incorporation into the new virion and oligomerization of N proteins, the NTD interacts with the M protein to form virion particles. It is briefly involved in phosphorylation, oligomerization, and N to M protein interaction[15]. Due to the N protein's various roles, it may have a potential value in vaccine development[16][17].

R203k mutation is done along with N_G204R mutation (28881-28883: GGG>AAC), R203k mutation appeared from the alteration of Guanine to adenosine, and alteration of R (arginine) into K (lysine)[18]. This mutation is located in the SR-rich region, which is known to be intrinsically disordered. It first appeared in a SARS-CoV-2 sequence from northern Europe (Netherlands) and then occurred 13828 times in 73 countries. We found 28% (n=21) R substituted to K at 203. In the previous studies using PROVEAN, as a variant function prediction tool for both R203k and G204R mutations and the PROVEAN suggested a deleterious effect as a result of the two amino acid substitutions.

E. MSA

The phylogenetic tree was constructed using MAFFT (v7.471) online tool via Neighbor Joining method as shown in Figure 2.

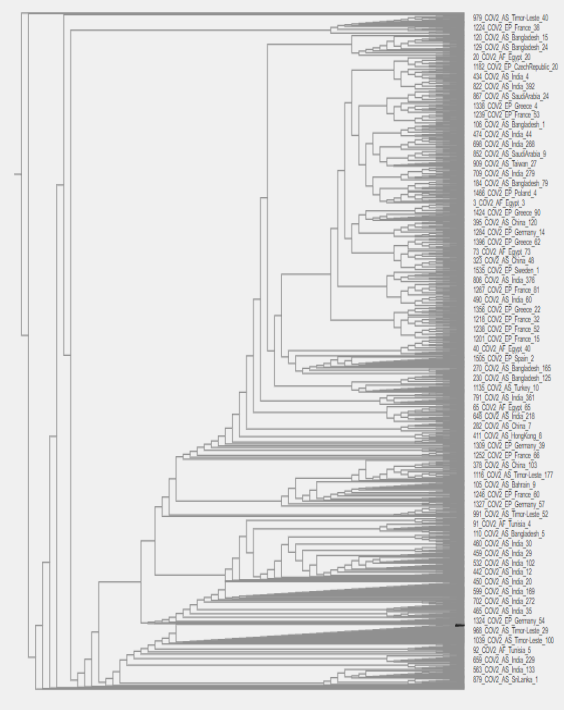


FIGURE 2. THE PHYLOGENETIC TREE OF 1535 SEQUENCES

F. Conserved Region

The percentage of the similar base pairs in the conserved region within these 1535 sequences is 80.1 %, which means that these sequences are highly similar in structure to each other. This may indicate that a single vaccine, if found, to work effectively on these three continents' populations is high.

G. Principle component (PCA) analysis

Our analysis detects 6805 mutations affecting 15 coding regions of SARS-COV2 from different continents. As shown in figures 3 and 5, the highest number of mutations occurs in spike protein; however, we found only two mutation were conserved among all analyzed sequences located in spike protein. Among 6805 mutations, there were 12 mutations appears in all continents, as shown in table 2. As shown in figure 6, PCA analysis shows that Africa, Asia and Europe share all the African mutation pattern. In contrast, the Asian mutation pattern shows a more diverse pattern, and this correlated to the number of available sequences retrieved from Asia.

TABLE II. SUMMARY OF EACH MUTATION AND IT'S PROTEIN

Mutation	Protein	Mutation	Protein	Mutation	Protein
D614G	Spike	Q57H	NS3	V198I	NSP2
P323L	NSP12	T85I	NSP2	L37F	NSP6
T428I	NSP3	L5F	Spike	S913L	NSP12
G15S	NSP5	L84S	NS8	Q1884H	NSP3

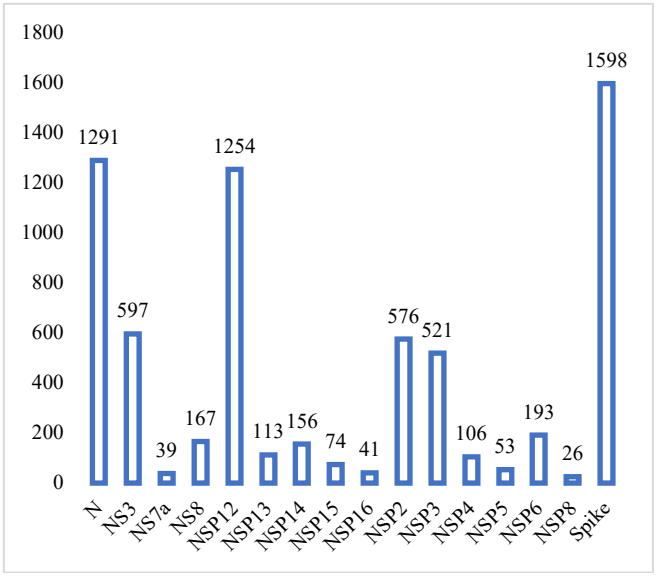


FIGURE 3: DISTRIBUTION OF THE NUMBER OF MUTATION PER EACH CODING REGION.

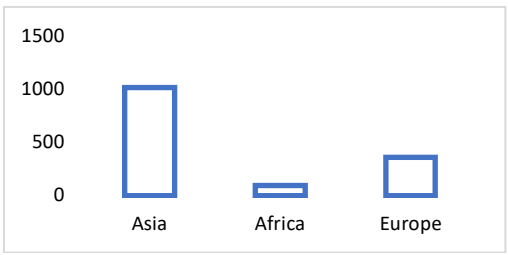
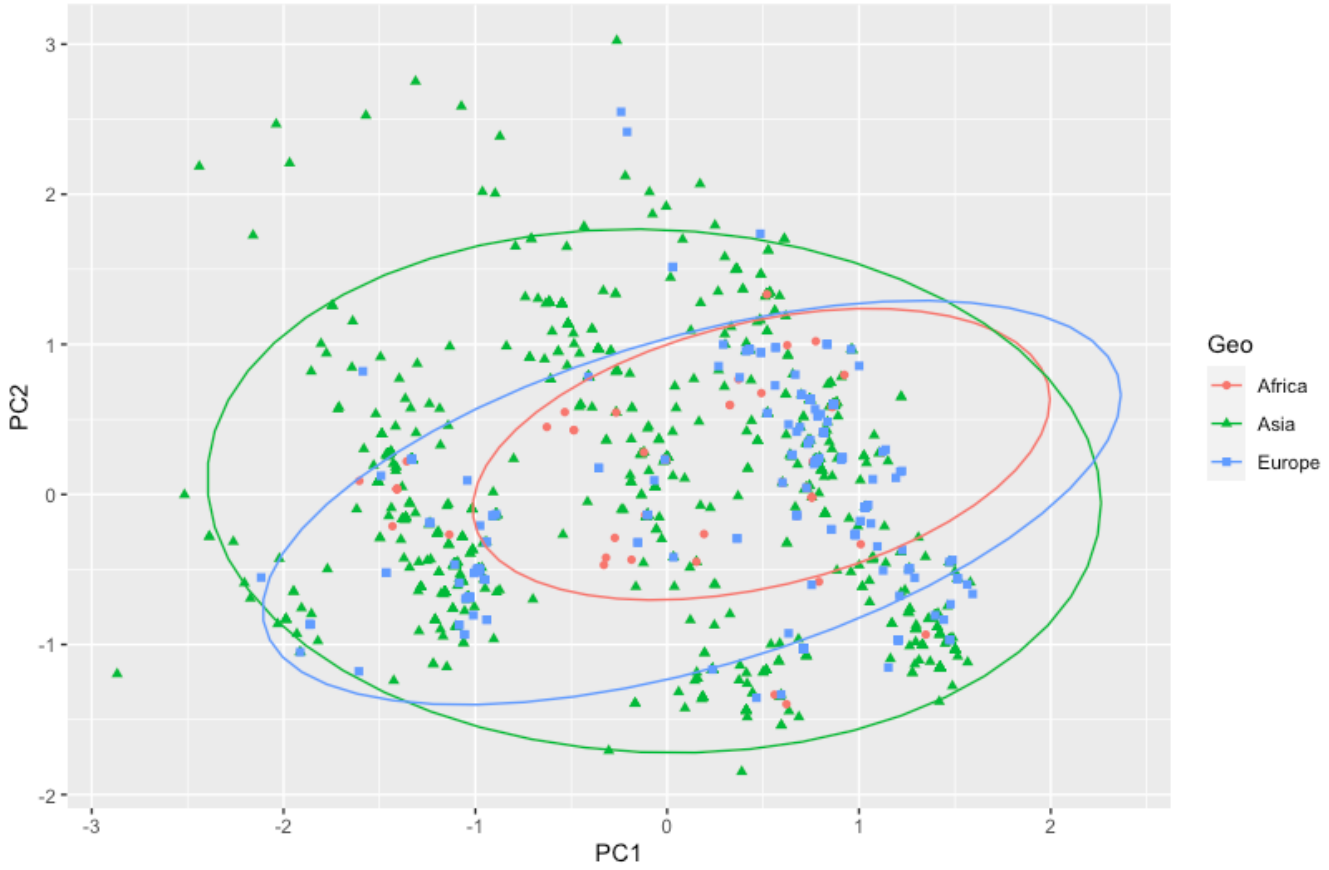


FIGURE 4. GRAPHICAL SUMMARY OF NUMBER OF MUTATION IN EACH CONTINENTS



REFERENCES

[1] F. safari Hamideh Amirfakhryan, "Outbreak of SARS-CoV2: Pathogenesis of infection and cardiovascular involvement," no. January, 2020.

[2] F. Perrotta, M. Gabriella, M. Cazzola, and A. Bianco, "Severe respiratory SARS-CoV2 infection : Does ACE2 receptor matter ?," no. January, 2020.

[3] Y. C. Li, "The neuroinvasive potential of SARS - CoV2 may play a role in the respiratory failure of COVID - 19 patients," no. February, pp. 24–27, 2020.

[4] A. B. Gussow, N. Auslander, G. Faure, Y. I. Wolf, F. Zhang, and E. V Koonin, "Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses," 2020.

[5] T. Chang, D. Yang, M. Wang, K. Liang, and P. Tsai, "Genomic analysis and comparative multiple sequences of SARS-CoV2," pp. 537–543, 2019.

[6] B. Robson, "COVID-19 Coronavirus spike protein analysis for synthetic vaccines , a peptidomimetic antagonist , and therapeutic drugs , and analysis of a proposed achilles ' heel conserved region to minimize probability of escape mutations and drug resistance," no. January, 2020.

[7] T. A. Meriem LAAMARTI1, "Large scale genomic analysis of 3067 SARS- CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations," vol. 3, 2020.

[8] M. Eaaswarkhanth, A. Al Madhoun, and F. Al-mulla, "Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality?," no. January, 2020.

[9] L. Zhang, C. B. Jackson, H. Mou, A. Ojha, and E. S. Rangarajan, "The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity," 2020.

[10] A. B. Demİr, D. Benvenuto, H. Abacioğlu, S. Angeletti, and M. Ciccozzi, "Identification of the nucleotide substitutions in 62 SARS-CoV-2 sequences from Turkey," no. April, pp. 178–184, 2020.

[11] D. T.-B. Osman Mutluhan UGUREL, Oguz ATA3, "An updated analysis of variations in SARS-CoV-2 genome," pp. 157–167, 2020.

[12] M. Pachetti et al., "Emerging SARS - CoV - 2 mutation hot spots include a novel RNA - dependent - RNA polymerase variant," J. Transl. Med., pp. 1–9, 2020.

[13] G. B. Chand and A. Banerjee, "Identi fi cation of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure," pp. 1–11, 2020.

[14] O. N. Mutations, "SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis," vol. 5, no. 3, pp. 1–7, 2020.

[15] N. J. S. Agnes. P. Chan, Yongwook Choi, "CONSERVED GENOMIC TERMINALS OF SARS-COV-2 AS CO-EVOLVING FUNCTIONAL ELEMENTS AND POTENTIAL THERAPEUTIC TARGETS," 2020.

[16] E. B. Dhurvas Chandrasekaran Dinesh, Dominika Chalupska, Jan Silhan, Vaclav Veverka, "Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein," 2020.

[17] J. Mu, J. Xu, L. Zhang, T. Shu, D. Wu, and M. Huang, "SARS-CoV-2-encoded nucleocapsid protein acts as a viral suppressor of RNA interference in cells," pp. 1–4, 2020.

[18] A. E. Castillo et al., "Phylogenetic analysis of the first four SARS - CoV - 2 cases in Chile."