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

Meriem LAAMARTI[1]¶, Tarek ALOUANE[1]¶, Souad KARTTI[1], M.W. CHEMAO-ELFIHRI[1], Mohammed HAKMI[1], Abdelomunim ESSABBAR[1], Mohamed LAAMART[1], Haitam HLALI[1], Loubna ALLAM[1], Naima EL HAFIDI[1], Rachid EL JAOUDI[1], Imane ALLALI[2], Nabila MARCHOUDI[3], Jamal FEKKAK[3], Houda BENRAHMA[4], Chakib NEJJARI[5], Saaid AMZAZI[2], Lahcen BELYAMANI[6] and Azeddine IBRAHIMI[1]*

[1] Medical Biotechnology Laboratory (MedBiotech), Bioinova Research Center, Rabat Medical & Pharmacy School, Mohammed Vth University in Rabat, Morocco

[2] Laboratory of Human Pathologies Biology, Department of Biology, Faculty of Sciences, and Genomic Center of Human Pathologies, Faculty of Medicine and Pharmacy, Mohammed V University in Rabat, Morocco.

[3] Anoual Laboratory of Radio-Immuno Analysis, Casablanca, Morocco.

[4] Faculty of Medicine, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco.

[5] International School of Public Health, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco.

[6] Emergency Department, Military Hospital Mohammed V, Rabat Medical & Pharmacy School, Mohammed Vth University in Rabat, Morocco.

*** Corresponding author:** a.ibrahimi@um5s.net.ma

¶ These authors contributed equally to this work

Abstract

In late December 2019, an emerging viral infection COVID-19 was identified in Wuhan, China, and became a global pandemic. Characterization of the genetic variants of SARS-CoV-2 is crucial in following and evaluating their spread across countries. In this study, we collected and analyzed 3,067 SARS-CoV-2 genomes isolated from 59 countries during the first three months after the onset of this virus. Using comparative genomics analysis, we traced the profiles of the whole-genome mutations and compared the frequency of each mutation in the studied population. The accumulation of mutations during the epidemic period with their geographic locations was also monitored. The results showed 716 site mutations, of which 457 (64%) had a non-synonymous effect. Frequencies of mutated alleles revealed the presence of 39 recurrent non-synonymous mutations, including 10 hotspot mutations with a prevalence higher than 0.10 in this population and distributed in six genes of SARS-CoV-2. The distribution of these recurrent mutations on the world map revealed certain genotypes specific to the geographic location. We also found co-occurring mutations resulting in the presence of several haplotypes. Thus, evolution over time has shown a mechanism of co-accumulation and the phylogenetic analysis of this population indicated that this virus can be divided into 3 clades, including a subgroup-specific to the genomes of the United States. On the other hand, analysis of the selective pressure revealed the presence of several negatively selected residues that could be useful for considerations as therapeutic target design.

We have also created an inclusive unified database (http://moroccangenomes.ma/covid/) that lists all of the genetic variants of the SARS-CoV-2 genomes found in this study with phylogeographic analysis around the world.

Keywords: SARS-CoV-2, Hotspots mutations, Dissemination, Genomic analysis.

Introduction

The recent emergence of the novel, human pathogen Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in China with its rapid international spread poses a global health emergency. On March 11, 2020, World Health Organization (WHO) publicly announced the SARS-CoV-2 epidemic as a global pandemic. As of March 23, 2020, the COVID-19 pandemic had affected more than 190 countries and territories, with more than 464,142 confirmed cases and 21,100 deaths (1).

The new SARS-CoV-2 coronavirus is an enveloped positive-sense single-stranded RNA virus (2) and member of a large family named coronavirus which have been classified under three groups (2) two of them are responsible for infections in mammals (2), such us: bat SARS-CoV-like; Middle East respiratory syndrome coronavirus (MERS-CoV). Many recent studies have suggested that SARS-CoV-2 was diverged from bat SARS-CoV-like (3-4).

The size of the SARS-CoV2 genome is approximately 30 kb and its genomic structure has followed the characteristics of known genes of Coronavirus; the polyprotein ORF1ab also known as the polyprotein replicase covers more than 2 thirds of the total genome size, structural proteins, including spike protein, membrane protein, envelope protein and nucleocapsid protein. There are also six ORFs (ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10) predicted as hypothetical proteins with no associated function (5).

Characterization of viral mutations can provide valuable information for assessing the mechanisms linked to pathogenesis, immune evasion and viral drug resistance. In addition, viral mutation studies can be crucial for the design of new vaccines, antiviral drugs and diagnostic tests. A previous study (6) based on an analysis of 103 genomes of SARS-CoV-2 indicates that this virus has evolved into two main types. Type L being more widespread than type S, and type S representing the ancestral version. In addition, another study (7) conducted on 32 genomes of strains sampled from China, Thailand and the United States between December 24, 2019 and January 23, 2020 suggested increasing tree-like signals from 0 to 8.2%, 18.2% and 25, 4% over time, which may indicate an increase in the genetic diversity of SARS-CoV-2 in human hosts.

Therefore, analyzing viral mutations and monitoring of the evolution capacity over time of SARS-CoV-2 in a large population is necessary. In this study, we characterized the genetic

variants in 3067 complete genomes of SARS-CoV-2 to assess the genetic diversity of SARS-CoV-2 and to track mutations accumulation over time geographic location. On the other hand, we established selective pressure analysis to predict negatively selected residues which could be useful for the design of therapeutic targets. We have also created a database containing the comparative genomic analysis of the SARS-CoV-2 genomes, in order to facilitate comparison and research for the COVID-19 scientific community.

Materials and Methods

Data collection and Variant calling analysis

3067 sequences of SARS-CoV-2 were collected from the GISAID EpiCovTM (update: 02-04-2020) and NCBI (update: 20-03-2020) databases. Only complete genomes were used in this study. **Table S1** illustrating detailed information on the genomes downloaded. Genomes were mapped to the reference sequence Wuhan-Hu-1/2019 using Minimap v2.12-r847-dirty (8).The BAM files were sorted by SAMtools sort (9). The final sorted BAM files were used to call the genetic variants in variant call format (VCF) by SAMtools mpileup (9) and bcftools (9). The final call set of the 3067 genomes, was annotated and their impact was predicted using SnpEff v4.3t (10). First, the SnpEff databases were built locally using annotations of the reference genome NC_045512.2 obtained in GFF format from NCBI database. Then, the SnpEff database was used to annotate SNPs and InDels with putative functional effects according to the categories defined in the SnpEff manual (http://snpeff.sourceforge.net/SnpEff_manual.html).

Phylogentic analysis and geo-distribution

The downloaded full-length genome sequences of coronaviruses isolated from different hosts from public databases were subjected to multiple sequence alignments using Muscle v 3.8 (11). Maximum-likelihood phylogenetic trees with 1000 bootstrap replicates were constructed using RaxML v 8.2.12 (39), Tree was done using figtree v 1.4.4(12)

Selective pressure and modelling

We used Hyphy v2.5.8 (13) to estimate synonymous and non-synonymous ratio dN / dS (ω). Two datasets of 191 and 433 for orf1ab and genes respectively were retrieved from Genbank (http://www.ncbi.nlm.nih.gov/genbank/). After deletion of duplicated and cleaning the sequences, only 92 and 35 for orf1ab and spike proteins, respectively, were

used for the analysis. The selected nucleotide sequences of each dataset were aligned using Clustalw codon-by-codon and the phylogenetic tree was obtained using ML (maximum likelihood) available in MEGA X (14). For this analysis, four Hyphy's methods were used to study site-specific selection: SLAC (Single-Likelihood Ancestor Counting (15), FEL (Fixed Effects Likelihood) (15) , FUBAR (Fast, Unconstrained Bayesian AppRoximation) (16) and MEME (Mixed Effects Model of Evolution) (17). For all the methods, values supplied by default were used for statistical confirmation and the overall ω value was calculated according to ML trees under General Time Reversible Model (GTR model). The CI- TASSER generated models (https: //zhanglab.ccmb.med.umich.edu / COVID-19 /) of ORF1ab nonstructural proteins (NSP) 3,4,6,12,13,14 and 16 were used to highlight the sites under selective pressure on the protein, due to the lack of their experimental structures and the absence of suitable templates for homology modelling. On the other hand, the cryo-EM structure with PDB id 6VSB was used as a model for the gene S in its prefusion conformation. Structure visualization and image rendering were performed in PyMOL 2.3 (Schrodinger LLC).

Pangenome construction

115 proteomes of the genus *Betacorononavirus* were obtained from the NCBI database (update: 20-03-2020), of which 83 genomes belonged to the SARS-CoV-2 species and the rest distributed to other species of the same genus publicly available. These proteomes were used for the construction of pangenome at the inter-specific scale of *Betacoronavirus* and intra-genomic of SARS-CoV-2. The strategy of best reciprocal BLAST results (18) was implemented to identify all of the orthologous genes using Proteinortho v6.0b (19). Proteins with an identity above 60% and sequence coverage above 75% with an e-value threshold below 1e-5 were used to be considered as significant hits.

Results

SARS-CoV-2 genomes used in this study

In this study, we used 3,067 SARS-CoV2 complete genomes collected from GISAID EpiCovTM (update: 02-04-2020) and NCBI (update: 20-03-2020) databases. These strains were isolated from 59 countries (**Fig 1A**). The most represented origin was American strains with 783 (25.53%), followed by strains from England, Iceland, and China with 407

(13.27%), 343 (11.18%), 329 (10.73%), respectively. The date of isolation was during the first three months after the appearance of the SARS-CoV-2 virus, from December 24, 2019, to March 25, 2020 (**Fig 1B**). Likewise, about two-thirds of these strains collected in this work were isolated during March.

Allele frequencies revealed a diversity of genetic variants in six SARS-Cov-2 genes

To study and follow the appearance and accumulation of mutations, we traced the profiles of non-synonymous mutations, and we compared their frequency in the studied population. Remarkably, compared to the reference sequence Wuhan-Hu-1/2019, a total of 716 mutations were identified, including 457 (64%) non-synonymous and 188 (26%) synonymous mutations. The rest (10%) distributed to the intergenic regions. Focusing on non-synonymous mutations, the analysis of the frequencies of mutated alleles revealed the presence of 39 mutations with a prevalence greater than 0.006 (0.06% of the population), which corresponds to at least 20 / 3,067 SARS-CoV-2 genomes (**Fig 2**).

These recurring mutations were distributed in six genes with variable frequencies, of which the gene coding for replicase polyprotein (ORF1ab), spike protein, membrane glycoprotein, nucleocapsid phosphoprotein, ORF3a, and ORF8. Overall, ORF1ab harbored more non-synonymous mutations compared to the other five genes with 22 mutations, including three mutations located in NSP12-RNA- dependent RNA polymerase (M4555T, T4847I and T5020I), three in NSP13-helicase (V5661A, P5703L and M5865V), two in NSP5-main proteinase (G3278S and K3353R), two in NSP15-EndoRNAse (I6525T, Ter6668W), two in NSP3-multi domains (A876T and T1246I), one in NSP14-exonuclease (S5932F) and one in NSP4-transmembrane domain 2 (F3071Y). Likewise, spike protein harbored three frequent mutations, including V483A in receptor-binding domain (RBD). The rest of the mutations were found in nucleocapsid phosphoprotein (S193I, S194L, S197L, S202N, R203K and G204R), ORF3a (S193I, S194L, S197L, S202N, R203K and G204R), membrane glycoprotein (D3G and T175M) and ORF8 (V62L and L84S).

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

Interestingly, among all recurrent mutations, ten were found as hotspot mutations with a frequency greater than 0.10 in this study population (**Fig 2**). The most represented was D614G mutation at spike protein with 43.46% (n = 1.333) of the genomes, the second was

L84S at ORF8 found in 23.21% (n = 712). Thus, the gene coding for orf1ab had four mutations hotspots, including S5932F of NSP14-exonuclease, M5865V of NSP13-helicase L3606F of NSP6- transmembrane domain and T265I of NSP2 found with 17.02%, 16.56%, 14.38% and 10.66% of the total genomes, respectively. For the four other hotspot mutations were distributed in ORF3a (Q57H and G251V) and nucleocapsid phosphoprotein (R203K and G204R).

Geographical distribution and origin of mutations worldwide

Out of 3067 genomes, only 105 were found to be wild type and were particularly of Chinese origin. However, mutant strains were dispersed in different countries and represented up to 2962 strains with different genotype profiles. We performed a geo-referencing mutation analysis to identify region-specific loci. Remarkably, USA was the country with the highest number of mutations, with 316 (44% of the total number of mutations), including 173 (24%) singleton mutation (specific to USA genomes). While the Chinese genomes containing 22% of the total number of mutations, followed by France and New Zealand with 4% and 2%, respectively. It is interesting to note that among the 59 countries, 26 harbored singleton mutations. **Table S2** illustrates the detailed singleton mutations found in these countries. The majority of the genomes analyzed carried more than one mutation. However, among the recurrent non-synonymous, synonymous and intergenic mutations (**Fig 3**). We found G251V (in ORF3a), L84S (in ORF8) and S5932F (in ORF1ab) present on all continents except Africa and Austria (**Fig 4**). While F924F, L4715L (in orf1ab), D614G (in spike) and an intergenic variant at position 241 appeared in all strains except those from Asia. In Algeria, the genomes harbored mutations very similar to those in Europe, including two recurrent mutations T265I and Q57H of the ORF3a. Likewise, the European and Dutch genomes also shared ten recurrent mutations. On the other hand, continent-specific mutations have also been observed, for example in America, we found seven mutations shared in almost all genomes. In addition, two mutations at positions 28117 and 28144 were shared by the Asian genomes, while four different positions 1059, 14408, 23403, 25563 and 1397, 11083, 28674, 29742 were shared by African and Australian genomes (Supplementary material). The majority of these mutations are considered to be transition mutations with a high ratio of A substituted by G. The genome variability was more visible in Australia, New Zealand and America than in the rest of the

world.

Evolution of mutations over time

We selected the genomes of the SARS-CoV-2 virus during the first three months after the appearance of this virus (December 24 to March 25). We have noticed that the mutations have accumulated at a relatively constant rate (**Fig 5**). The strains selected at the end of March showed a slight increase in the accumulation of mutations with an average of 11.34 mutations per genome, compared to the gnomes of February, December and January with an average number of mutations of 9.26, 10.59 10.34 respectively. The linear curve in **Figure 5** suggests a continuous accumulation of SNPs in the SARS-CoV-2 genomes in the coming months. This pointed out that many countries had multiple entries for this virus that could be claimed. Thus in the deduced network demonstrated transmission routes in different countries.

The study of the accumulation of mutations over time showed a higher number of mutations in the middle of the outbreak (end of January). At the same time, an increase in the number of mutations in early April was also observed. The first mutations to appear were mainly located in the intergenic region linked to the nucleocapsid phosphoprotein and the orf8 protein. The T265I, D614G and L84S hotspot mutations located in orf1ab and Spike proteins respectively were introduced into the virus for the first time in late February.

Phylogeographical analysis of SARS-CoV-2 genomes

The phylogenetic tree based on the whole genome alignment showed the presence of 3 different clades (**Fig S1**) and demonstrate that SARS-CoV-2 is wildly disseminated across distinct geographical location. The results showed that several strains are closely related even though they belong to different countries. Which indicate likely transfer events and identify routes for geographical dissemination. For phylogenetic tree (http://moroccangenomes.ma/covid/) showed multiple introduction dates of the virus inside the USA with the first haplotype introduced related to the second epidemic wave in china.

Selective pressure analysis

Selective pressure on ORF1ab, gene harbored a high rate of mutations and on the Spike gene, indicated a single alignment-wide ω ratio of 0.571391 and 0.75951 for spike and

246 or1ab, respectively. Most sites for both genes had ω <1 values, indicating purifying
247 selection. We estimated eight sites under negative selection pressure (696, 1171, 2923,
248 3003, 3715, 5221, 5704 and 6267) and three sites under positive selection pressure (1473,
249 2244 and 3090) in ORF1ab. For spike, we found seven sites under negative selection
250 pressure (215, 474, 541, 809, 820, 921 and 1044), while only site 5 was found under
251 negative selection pressure (**Table 1**).
252 The modelling results of ORF1ab showed that the sites with positive selections were
253 distributed in NSP3 and NSP4. While the negatively selected codons were located in NSP3,
254 NSP4, NSP6, NSP12, NSP13, NSP14 and NSP16 (**Fig 6**). Thus, only one negatively
255 selected amino acid residue was observed on the receptor binding domain, suggesting that
256 there was no strong selective pressure in the region (**Fig 7**).

257 **Inter and intra-specific pan-genome analysis**
258 In order to highlight the structural proteins shared at the inter-specific scale between the
259 isolates of the genus *Betacoronavirus*, thus at the intra-genomic scale of SARS-CoV-2, we
260 have constructed a pan-genome by clustering the sets of proteins encoded in 115 genomes
261 available publicly in NCBI (update: 20-03-2020), including 83 genomes of SARS-Cov-2
262 and the rest distributed to other species of the same genus. Overall, a total of 1,190 proteins
263 were grouped into a pangenome of 94 orthologous cluster proteins, of which ten proteins
264 cluster were shared between SARS-CoV-2 and only three species of the genus
265 *Betacoronavirus* (BatCoV RaTG13, SARS-CoV and Bat Hp-betacoronavirus/
266 Zhejiang2013). Of these, BatCoV RaTG13 had more orthologous proteins shared with
267 SARS-CoV-2, followed by SARS-CoV with ten and nine orthologous proteins,
268 respectively (**Fig 8A**).

269 It is interesting to note that among all the strains used of *BetaCoronavirus*, the ORF8
270 protein was found in orthology only between SARS-RATG13 and SARS-CoV-2. In
271 addition, the ORF10 protein was found as a singleton for SARS-CoV-2. While, the analysis
272 of the pangenome at the intra-genomic scale of 83 isolates of SARS-CoV-2 (**Fig 8B**),
273 showed that ORF7b and ORF10 were two accessory proteins (proteins variable) in SARS-
274 CoV- 2 genomes, while the other proteins belonged to the core proteins of SARS-CoV-2
275 (conserved in all genomes).

Discussion

The rate of mutations results in viral evolution and variability in the genome, thus allowing viruses to escape host immunity, as well as drugs (20). Initial published data suggests that SARS-CoV-2 is genetically stable (21) which may increase the effectiveness of vaccines under development. The study on the genomic variation of SARS- CoV- 2 is very important for the investigation of pathogenesis, disease course, prevention, and treatment of SARS- CoV- 2 infection. In this study, we characterized the genetic variations in a large population of SARS-CoV-2 genomes (3,067). Our results showed a diversity of mutations detected with different frequencies. Overall, more than 450 non-synonymous mutations in SARS-CoV-2 genomes have been identified. The orf1ab gene contains 16 non-structural proteins (NSP1-NSP16) (22), and harbored more than 50 % of the frequent mutations, with a high mutation number 117, 61, and 61 in the NSP3, NSP12, and NSP2 Respectively. Both NSP2 and NSP3 were reported to be essential for correcting virus replication errors (23). On the other, hand recent studies have suggested that mutations falling in the endosome- associated- protein- like domain of the NSP2 protein, could explain why this virus is more contagious than SARS (24). The Replication enzymes NSP12 to NSP16 have been reported as antiviral targets for SARS-CoV (25). In the SARS-CoV-2 genomes, we found that NSP12 to NSP15 harbored nine recurrent non-synonymous mutations. Among them, eight identified as new mutations, including three (M4555T, T4847I, T5020I) in NSP12-RNA-dependent RNA polymerase, three (V5661A, P5703L and M5865V) in NSP13-Helicase, two (I6525T, Ter6668W) in NSP15-EndoRNAse. However, these new mutations must be taken into account when developing a vaccine using the ORF1ab protein sequences as a therapeutic target.

A high number of mutations were identified in the spike protein, an important determinant in pathogenicity that allows the virion attachment to the cell membrane by interacting with the host ACE2 receptor (26). Among all the frequent mutations in this protein, the V483A mutation has been identified in this receptor and found mainly in SARS-CoV-2 genomes isolated from USA. This result is consistent with the study of Junxian et al. (27). Eight stains from china, USA and France harbored V367F mutation previously described to enhance the affinity with ACE2 receptor (27).

Among the 716 mutations, ten were considered hypervariable genomic hotspots with high frequencies of mutated alleles such us the L3606F mutation detected in NSP6. The NSP6 protein works with NSP3 and NSP4 by forming double-membrane vesicles and convoluted membranes involved in viral replication (28). Besides, three previously reported mutations (M5865V, S5932F) and (R203K) were identified in ORF1ab and N respectively (20). Moreover, intraspecies pangenome analysis of SARS- CoV-2 showed that the six of the genes harboring hotspot mutations belong to the core genome.

Thus, under normal circumstances genomic variation increase the viruses spread and pathogenicity. This happens when the virus accumulated mutation enables its virulence potential (29). Genomic comparison of the studied population allowed us to gain insights into virus mutations occurrence over time and within different geographic areas. In the SARS-CoV virus, the SNP distribution is not random, and it is more dominant in critical genes for the virus. As it was already found in many studies (20,30). Our results confirmed what was previously described and elucidate the presence of numerous hotspot mutations. Besides, Co-occurrence mutations were also common in different countries all along with singleton mutations. In the case of the USA, the singleton mutations are driven by the single group that diverged differently due to the environment, the host, and the number of generations. These mutations are due to the low fidelity of reverse transcriptase (29, 31).

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

In this study, we have defined 3 major clades. The clustering of these genomes revealed the spread of clades to diverse geographical regions. We observed an increase of mutations over time following the first dissemination event from China. Specific haplotypes were also predominant to a geographical location, especially in the US. This study opens up new perspectives to determine whether one of these frequent mutations will lead to biological differences and their correlation with different mortality rates.

Among the 7 proteins of orf1ab that harbored sites under selective pressure, only NSP3 and NSP4 contains both residues under positive and negative selection. The modelling of NSP3 domains shows that only the negative selection site 1171 (Thr- 353), was located at

the conserved macro domain Mac1 (previously X or ADP-ribose 1" phosphatase) (32). This domain has been previously shown to be dispensable for RNA replication in the context of a SARS-CoV replicon (33). However, it could counteract the host's innate immune response (34). It was proposed that the 3Ecto luminal domain of NSP3 interacts with the large luminal domain of NSP4 (residues 112-164) to "zip" the ER membrane and induce discrete membrane formations as an important step in the generation of ER viral replication organelles (35, 36). As we have shown previously by the FEL, MEME and FUBAR methods, the orf1ab 2244 site coding for ILE-1426 is under positive selection pressure and since it is located on the luminal 3ecto domain of the NSP3 protein, this can be explained by a possible host influence on the virus in this domain.The results of selective pressure analysis revealed the presence of several negatively selected residues, one of which is located at the receptor-binding domain (GLN-474) and which is known by its interaction with the GLN24 residue of the human ACE2 (Angiotensin-converting enzyme 2) receptor (37). In general, it is well-known that negatively selected sites could indicate a functional constraint and could be useful for drug or vaccine target design, given their conserved nature and therefore less likely to change (38).

Conclusion

The SARS-CoV-2 pandemic has caused a very large impact on health and economy worldwide. Therefore, understanding genetic diversity and virus evolution become a priority in the fight against the disease. Our results show several molecular facets of the relevance of this virus. We identified ten non-synonymous hotspot mutations distributed in six of the virus genes with high frequencies of mutated alleles. We also were able to highlight an increase in mutations accumulation overtime and revealed the existence of three major clades in various geographic regions. Finally, the study allowed us to identify specific haplotypes by geographic location and provides a list of sites under selective pressure that could serve as an interesting avenue for future studies.

366 **Conflict of interest**

367 The authors declare that they have no competing interests.

368

369 **Acknowledgments**

370 We sincerely thank the authors and laboratories around the world who have sequenced and
371 shared the full genome data for SARS-CoV-2 in the GISAID database. All data authors
372 can be contacted directly via www.gisaid.org

373 This work was carried out under National Funding from the Moroccan Ministry of Higher
374 Education and Scientific Research (PPR program) to AI. This work was also supported, by
375 a grant to AI from Institute of Cancer Research of the foundation Lalla Salma.

376

377 **References**

378

379 1. World Health Organization. Infection prevention and control during health care
380 when COVID-19 is suspected: interim guidance, 19 March 2020. World Health
381 Organization. 2020. Available from:
382 https://apps.who.int/iris/handle/10665/331495
383 2. Enjuanes LD. Cavanagh K. Holmes MMC, Lai H. Laude P. Masters P et al.
384 (2000) Coronaviridae. In: Virus taxonomy. Classification and nomemclature
385 of viruses (M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B.
386 Carstens, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch,
387 C. R. Pringle, and R. B. Wickner eds.) Academic Press, San Diego. pp 835-
388 849.
389 3. Yeşilbağ K, Aytoğu G. Coronavirus host divergence and novel coronavirus
390 (Sars-CoV-2) outbreak. Clinical and Experimental Ocular Trauma and
391 Infection. 2020 Apr 23;2(1):1-9.
392 4. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal
393 origin of SARS-CoV-2. Nat Med. 2020;26: 450–452. DOI:10.1038/s41591-
394 020-0820-9

5. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579:265-269.
6. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 2020. DOI: 10.1093/nsr/nwaa036.
7. Li LQ, Huang T, Wang YQ, Wang ZP, Liang Y, Huang TB, et al. COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. J Med Virol. 2020 Mar 12. Doi: 10.1002/jmv.25757.
8. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100 .DOI: 10.1093/bioinformatics/bty191.
9. Li H, Hansaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–9. DOI:10.1093/bioinformatics/btp352
10. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6: 80-92 .DOI: 10.4161/fly.1969
11. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Re. 2004;32: 1792–1797. DOI: 10.1093/nar/gkh340
12. Rambaut, Andrew. "FigTree v1. 4." (2012).
13. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21: 676–679. DOI: 10.1093/bioinformatics/bti079
14. Kumar S, Stecher G, Li M, Knyaz C, Tamura. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35: 1547-1549. DOI: 10.1093/molbev/msy096.
15. Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino-acid sites under selection. Mol Biol Evol. 2005;22: 1208-1222. Doi:10.1093/molbev/msi105.
16. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SL, et al. Fubar: a fast, unconstrained Bayesian approximation for inferring selection. Mol Biol Evol. 2013;30: 1196-1205. DOI: 10.1093/molbev/mst030

17. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SL. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet. 2012;8: e1002764. Doi: 10.1371/journal.pgen.1002764
18. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12:962–968.
19. Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011;12:1–9. DOI:10.1186/1471-2105-12-124.
20. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 Mutation Hot Spots Include a Novel RNA-dependent-RNA Polymerase Variant. J Transl Med. 2020;18: 179. DOI: 10.1186/s12967-020-02344-6.
21. Su Y, Anderson D, Young B, Zhu F, Linster M, Kalimuddin S, et al. Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2. BioRxiv [Preprint]. 2020 [cited 2020 March 25]. Available from: https://www.biorxiv.org/content/10.1101/2020.03.11.987222v1
22. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. Gene Rep. 2020;19:100682. DOI:10.1016/j.genrep.2020.100682
23. Harcourt BH, Jukneliene D, Kanjanahaluethai A, Bechill J, Severson KM, Smith CM, et al. Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. J Virol. 2004;78:13600-13612. DOI:10.1128/JVI.78.24.13600-13612.2004
24. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. . COVID 2019: the role of the NSP2 and NSP3 in its pathogenesis. Journal of Medical Virology, 2020;92:584-588.
25. Subissi L, Imbert I, Ferron F, Collet A, Coutard B, Decroly E, et al. SARS-CoV ORF1b-encoded nonstructural proteins 12-16: replicative enzymes as

456 antiviral targets. Antiviral Res. 2014;101:122-30. DOI:
457 10.1016/j.antiviral.2013.11.006.
458 26. Hoffmann, Markus, et al. SARS-CoV-2 cell entry depends on ACE2 and
459 TMPRSS2 and is blocked by a clinically proven protease inhibitor.*Cell* (2020).
460 27. Ou J, Zhou Z, Dai R, Zhang J, Lan W, Zhao S. Emergence of RBD mutations
461 from circulating SARS-CoV-2 strains with enhanced structural stability and
462 higher human ACE2 receptor affinity of the spike protein. bioRxiv
463 2020.03.15.991844; doi: https://doi.org/10.1101/2020.03.15.991844
464 28. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ. Severe acute
465 respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce
466 double-membrane vesicles. mBio. 2013 Aug 13. pii: e00524-13. DOI:
467 10.1128/mBio.00524-13.
468 29. Stern A, Yeh MT, Zinger T, et al. The Evolutionary Pathway to Virulence of
469 an RNA Virus. Cell. 2017;169(1):35- 46.e19. doi:10.1016/j.cell.2017.03.013.
470 30. Wen F, Yu H, Guo J, Li Y, Luo K, Huang S. Identification of the hyper-variable
471 genomic hotspot for the novel coronavirus SARS-CoV-2. J Infect. 2020. pii:
472 S0163-4453(20)30108-0. DOI: 10.1016/j.jinf.2020.02.027.
473 31. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS.
474 Coronaviruses: an RNA proofreading machine regulates replication fidelity and
475 diversity. RNA Biol. 2011;8: 270–279. DOI: 10.4161/rna.8.2.15013.
476 32. Neuman BW. Bioinformatics and functional analyses of coronavirus
477 nonstructural proteins involved in the formation of replicative organelles.
478 Antiviral Res. 2016;135: 97-107. DOI: 10.1016/j.antiviral.2016.10.005.
479 33. Kusov Y, Tan J, Enrique A, Luis E, Hilgenfeld R. A G-quadruplex-binding
480 macrodomain within the “SARS-unique domain” is essential for the activity of
481 the SARS-coronavirus replication-transcription complex. Virology. 2015;484:
482 313–322. DOI: 10.1016/j.virol.2015.06.016.
483 34. Fehr AR, Channappanavar R, Jankevicius G, Fett C, Zhao J, Athmer J. The
484 Conserved Coronavirus Macrodomain Promotes Virulence and Suppresses the
485 Innate Immune Response during Severe Acute Respiratory Syndrome

Coronavirus Infection. mBio. 2016 Dec 13. pii: e01721-16. DOI: 10.1128/mBio.01721-16.

35. Hagemeijer MC, Monastyrska I, Griffith J, van der Sluijs P, Voortman J, en Henegouwen PM, et al. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. Virology. 2014;458: 125-135. DOI: 10.1016/j.virol.2014.04.027.
36. Qiu Y, Xu K. Functional studies of the coronavirus nonstructural proteins. STEMedicine. 2020;1:e39. DOI: 10.37175/stemedicine.v1i2.39.
37. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science. 2020;367:1444-1448. DOI: 10.1126/science.abb2762.
38. Mazumder R, Hu ZZ, Vinayaka CR, Sagripanti JL, Frost SD, Pond SL, et al. Computational analysis and identification of amino acid sites in dengue E proteins relevant to development of diagnostics and vaccines. Virus Genes. 2007;35:175-186. DOI: 10.1007/s11262-007-0103-2.
39. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. Doi: 10.1093/bioinformatics/btu033.

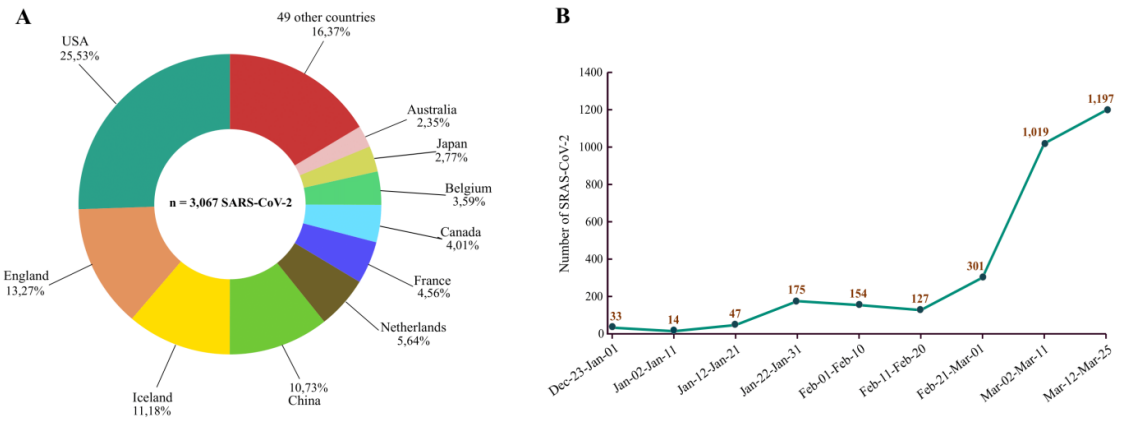

Figure 1: Distribution of the genomes of the 3,067 genomes used in this study by county and date of isolation. **A)** The pie chart represents the percentage of genomes used in this study according to their geographic origins. The colors indicate different countries. **B)** Number of genomes of complete pathogens, distributed over a period of 3 months from the end of December to the end of March.

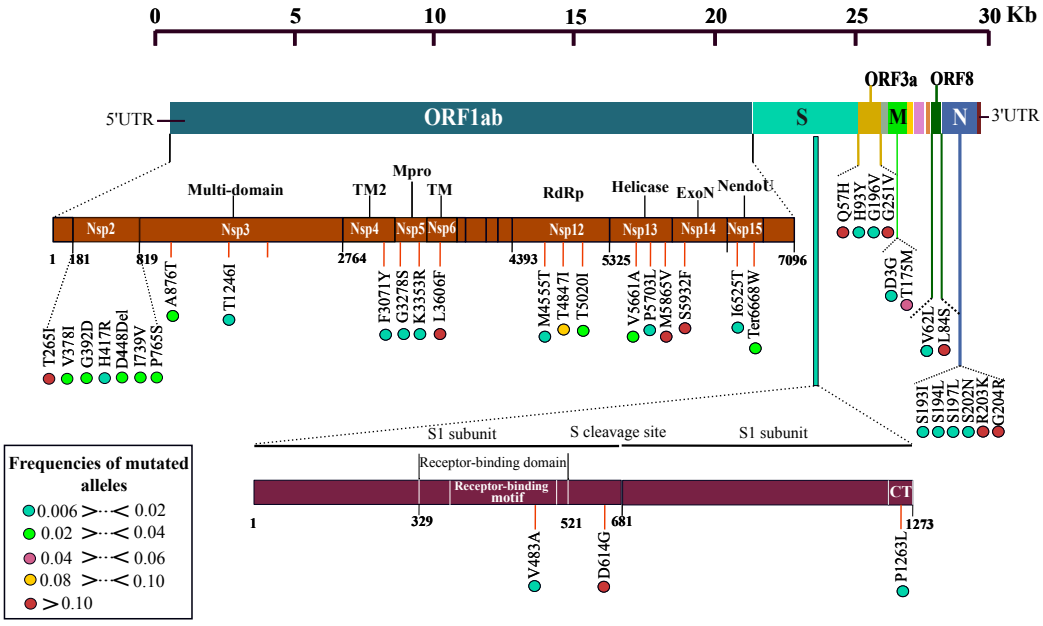

Figure 2: The linear diagrams represent genes distribution in the SARS-CoV-2 genome. Diagrams in garnet and brown represent the protein subunits of ORF1ab and S respectively. The presence of a mutation is represented by vertical lines. under each line, the most frequent variants are annotated as the amino-acid change at that specific site, and the frequency of mutations is presented by color-coded circles.

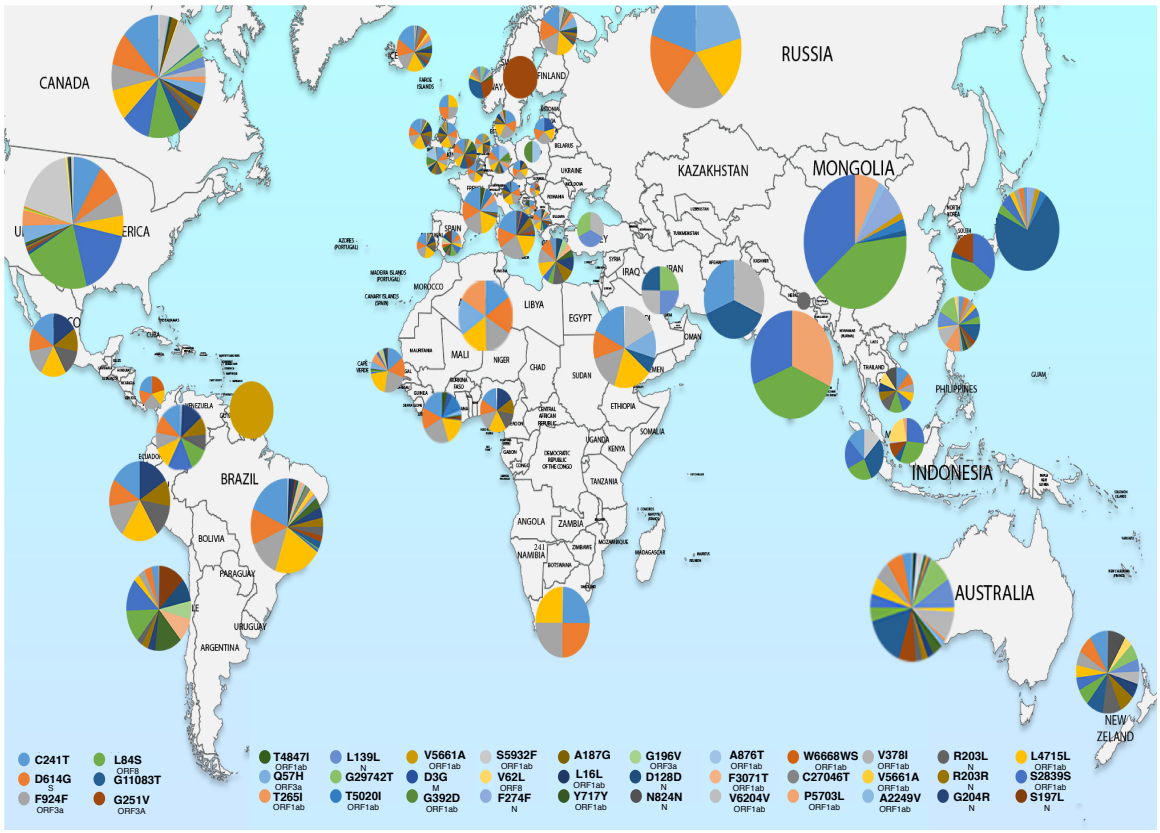

Figure 3: Map showing Geographical distribution of hotspot mutation in the studied population worldwide. The pie charts show the relative frequencies of haplotype for each population. The haplotypes are color coded as shown in the key.

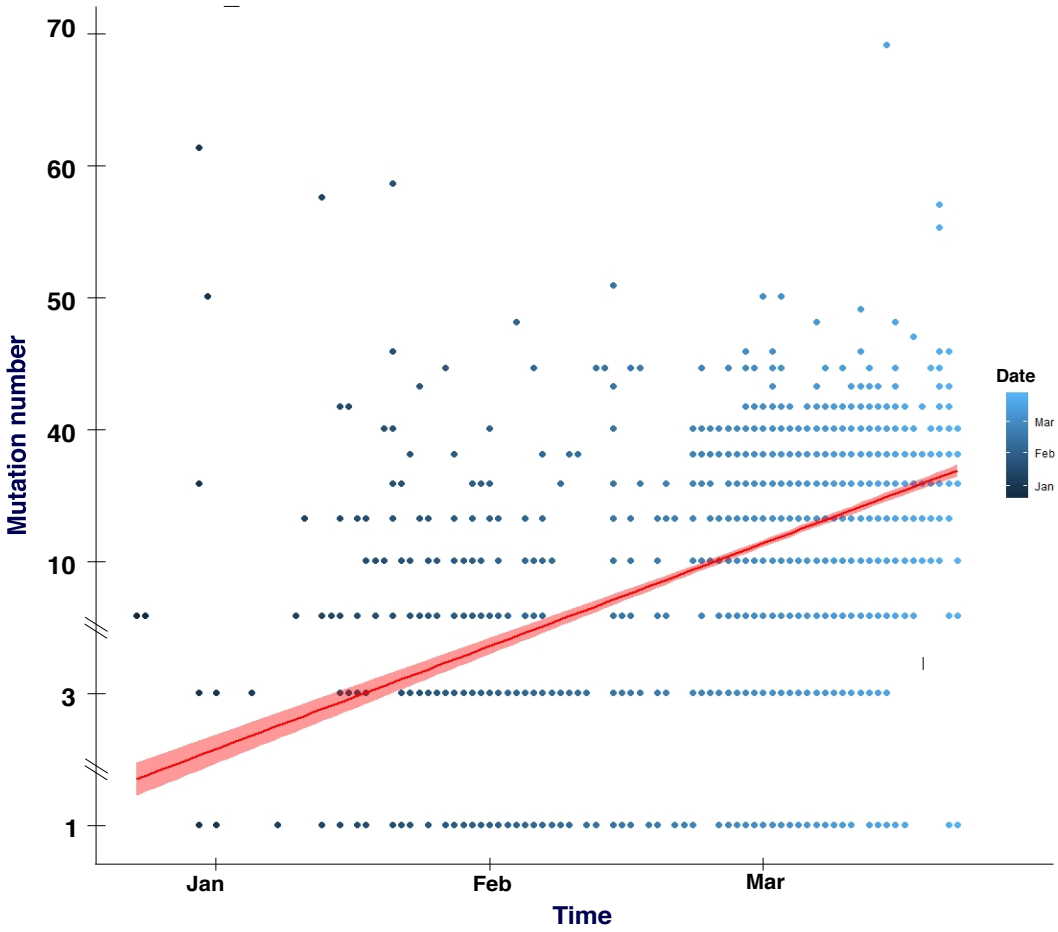

Figure 4 : The graph represents substitutions accumulation in a three months period. The accumulation of mutations increases linearly with time. The dots represent the number of mutations in a single genome. All substitutions were included non-synonymous, synonymous, intergenic.

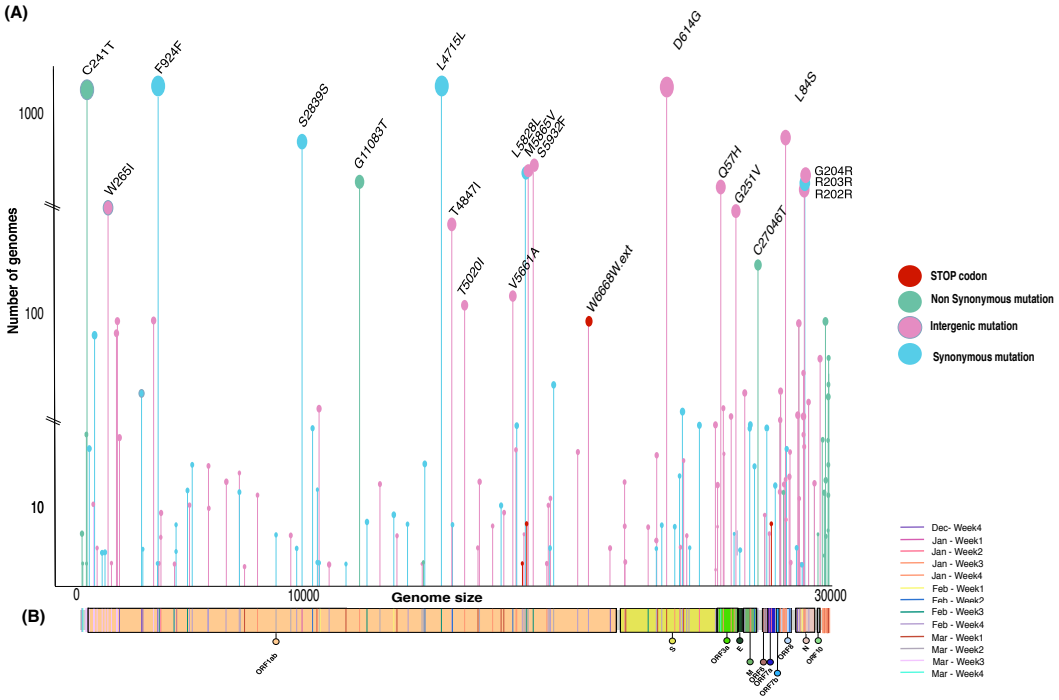

Figure 5 : Lollipop plots showing mutations distribution and frequency and time of the appearance in SARS-CoV-2 genome. A) The presence of a mutation is shown on the x-axis (lollipop), and the frequency of mutations is shown on the y-axis and correlates with the heights of the vertical lines representing each lollipop. Non-synonymous, synonymous, STOP Codon and intergenic mutations are presented as green, blue, red, and pink circles respectively. Predicted amino acid change has been represented for each hotspot mutation at the top of the circle. **B)** The diagram represents genes encoded in SARS-CoV-genome. The mutations are represented by a color-coded vertical line, each color represents a time period corresponding to the date of the first appearance of the mutation.

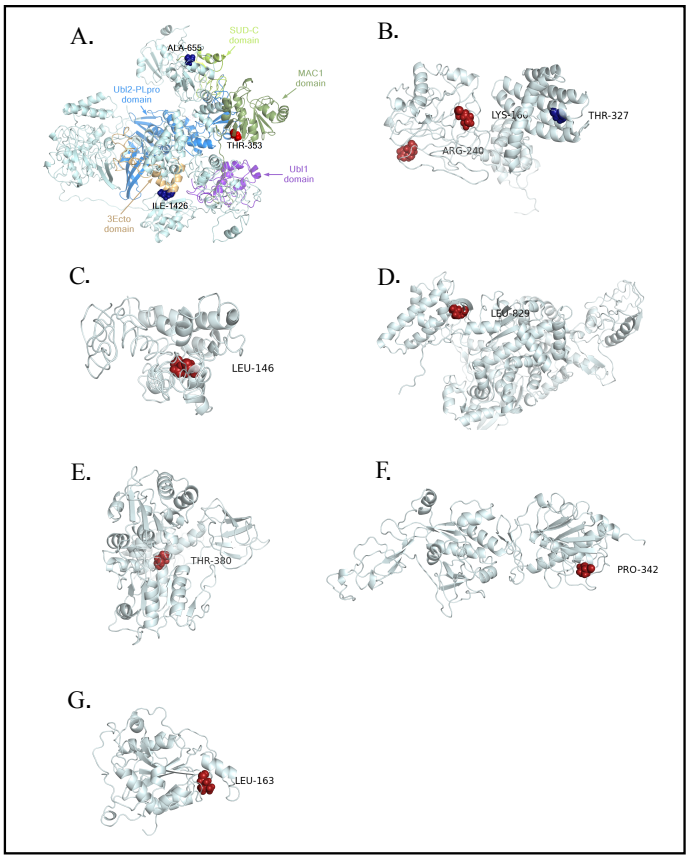

Figure6 : Structural view of selective pressure in orf1ab gene. The residue under the positive and negative selection is highlighted in blue and red respectively. The modeling of orf1ab non-structural proteins (NSP3, NSP4, NSP6, NSP12, NSP13, NSP14, and NSP16) harboring residues under pressure selection was produced using CI-TASSER. **A**. The NSP3 domains MAC1, Ubl1, Ubl2-PLpro, and SUD-C are color-coded in the 3D representation. The residues Ile-1426 and Ala-655 under negative selection are located respectively on 3Eco and SUD-C domains while Thr-353 residue under positive selection is shown on the MAC1 domain, **B**. 3D representation of the NSP4 protein, **C**. 3D representation of the NSP6 protein, **D**. 3D representation of the NSP12 protein, **E**. 3D representation of the NSP13 protein, **F**. 3D representation of the NSP14 protein, **G**. 3D representation of the NSP16 protein.

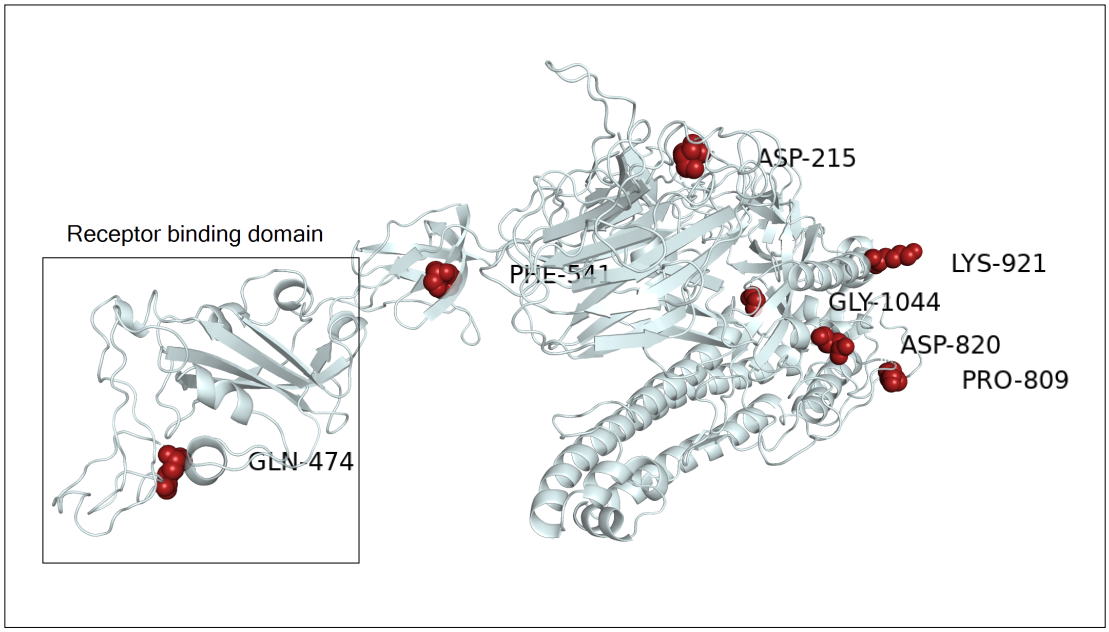

Figure 7: Structural view of selective pressure in spike gene. The negatively selected site in spike protein is highlighted in red. The only amino acid residue selected negatively on the receptor-binding domain corresponds to GLN-474. The cryo-EM structure with PDB id 6VSB was used as a model for the gene S in its prefusion conformation.

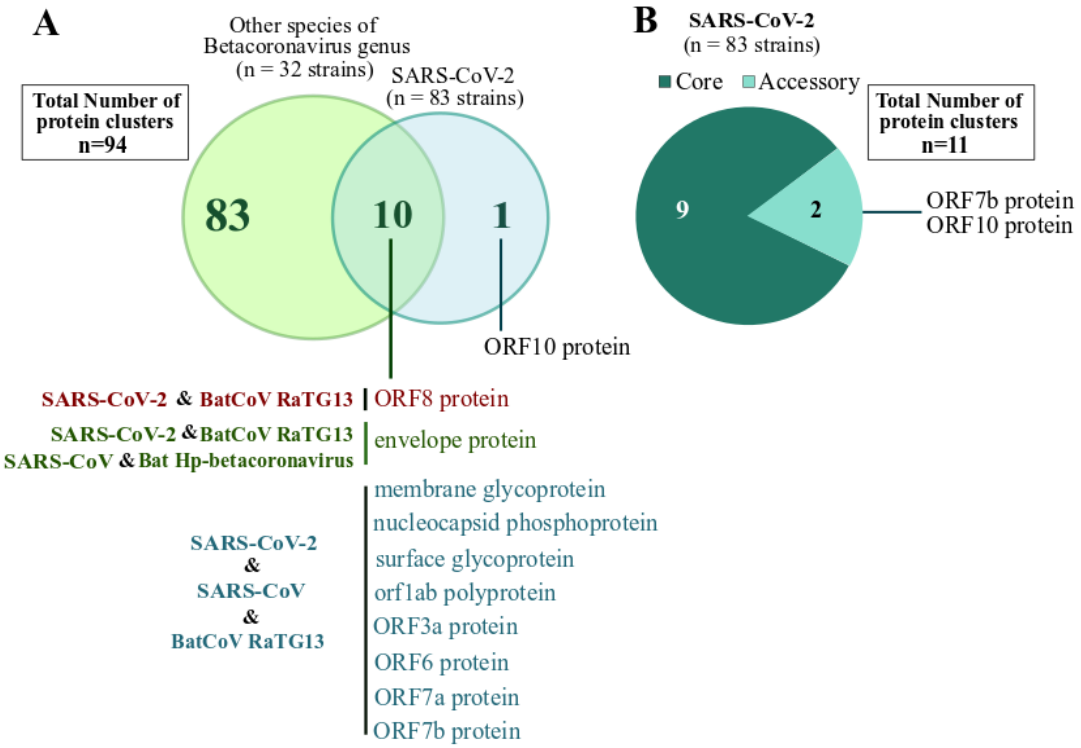

Figure 8 : Pangenome analysis of 32 from different *Betacoronavirus* species and 83 of SARS-CoV-2. (**A**) The Venn diagram represents the number of core, accessory, and unique proteins inside the *Betacoronavirus* genus. (**B**) The pie chart illustrates the core and accessory protein inside the SARS-CoV-2 specie.

Table 1 : Selective pressure analysis on the spike and orf1ab genes of SARS-CoV-2

Genes	ω	FEL method		MEME method	SLAC method		FUBAR method	
Spike	0.571391	PS	NS	PS	PS	NS	PS	NS
		-	Codons 215, 474, 809, 820, 921,	-	-	-	Codon 5	Codons 215, 474, 541,
orf1ab	0.75951	PS	NS	PS	PS	NS	PS	NS
		Codon 2244	Codons 1171, 2923, 3003, 3715, 5221, 5704, 6267, 6961	Codon2244	-	-	Codons 1473, 2244, 3090	-

652

653