Genomic diversity and hotspot mutations in 30,983 SARS-CoV-2 genomes: moving toward a universal vaccine for the "confined virus"?

Tarek ALOUANE[1¶], Meriem LAAMARTI[1¶], Abdelomunim ESSABBAR[1], Mohammed HAKMI[1], EL Mehdi BOURICHA[1], M.W. CHEMAO-ELFIHRI[1], Souad KARTTI[1], Nasma BOUMAJDI[1], Houda BENDANI[1], Rokaia LAAMRTI[2], Fatima GHRIFI[1], Loubna ALLAM[1], Tarik AANNIZ[1], Mouna OUADGHIRI[1] , Naima EL HAFIDI[1], Rachid EL JAOUDI[1], Houda BENRAHMA[3], Jalil ELATTAR[4], Rachid MENTAG[5], Laila SBABOU[6], Chakib NEJJARI[7], Saaid AMZAZI[8], Lahcen BELYAMANI[9] and Azeddine IBRAHIMI[1]*

[1] Medical Biotechnology Laboratory (MedBiotech), Bioinova Research Center, Rabat Medical & Pharmacy School, Mohammed Vth University in Rabat, Morocco.
[2] MASCIR, Rabat Design, Rabat, Morocco.
[3]Faculty of Medicine, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco.
[4] Laboratoire Riad, hay Riad, Rabat, Morocco.
[5] Biotechnology Unit, Regional Center of Agricultural Research of Rabat, National Institute of Agricultural Research, Rabat, Morocco.
[6] Microbiology and Molecular Biology Team, Center of Plant and Microbial Biotechnology, Biodiversity and Environment, Faculty of Sciences, Mohammed V University of Rabat.
[7] International School of Public Health, Mohammed VI University of Health Sciences (UM6SS), Casablanca, Morocco.
[8] Laboratory of Human Pathologies Biology, Faculty of Sciences, Mohammed V University in Rabat, Morocco.
[9] Emergency Department, Military Hospital Mohammed V, Rabat Medical & Pharmacy School, Mohammed Vth University in Rabat, Morocco.

Short title: Large-scale SARS-CoV-2 genome analysis

*** Corresponding author:**
e-mail: a.ibrahimi@um5s.net.ma

¶ These authors contributed equally to this work.

Abstract

The Coronavirus disease 19 (COVID-19) pandemic has been ongoing since its onset in late November 2019 in Wuhan, China. To date, the SARS-CoV-2 virus has infected more than 8 million people worldwide and killed over 5% of them. Efforts are being made all over the world to control the spread of the disease and most importantly to develop a vaccine. Understanding the genetic evolution of the virus, its geographic characteristics and stability is particularly important for developing a universal vaccine covering all circulating strains of SARS-CoV-2 and for predicting its efficacy. In this perspective, we analyzed the sequences of 30,983 complete genomes from 80 countries located in six geographical zones (Africa, Asia, Europe, North & South America, and Oceania) isolated from December 24, 2019 to May 13, 2020, and compared them to the reference genome. Our in-depth analysis revealed the presence of 3,206 variant sites compared to the reference Wuhan-Hu-1 genome, with a distribution that is largely uniform over all continents. Remarkably, a low frequency of recurrent mutations was observed; only 182 mutations (5.67%) had a prevalence greater than 1%. Nevertheless, fourteen hotspot mutations (> 10%) were identified at different locations, seven at the ORF1ab gene (in regions coding for nsp2, nsp3, nsp6, nsp12, nsp13, nsp14 and nsp15), three in the nucleocapsid protein, one in the spike protein, one in orf3a, and one in orf8. Moreover, 35 non-synonymous mutations were identified in the receptor-binding domain (RBD) of the spike protein with a low prevalence (<1%) across all genomes, of which only four could potentially enhance the binding of the SARS-CoV-2 spike protein to the human receptor ACE2.

These results along with the phylogenetic analysis demonstrate that the virus does not have a significant divergence at the protein level compared to the reference both among and within different geographical areas. Unlike the influenza virus or HIV viruses, the slow rate of mutation of SARS-CoV-2 makes the potential of developing an effective global vaccine very likely.

Keywords: SARS-CoV-2, genetic evolution, divergence, hotspot mutations, spike protein.

Introduction

The year 2019 ended with the appearance of groups of patients with pneumonia of unknown cause. Initial evidence suggested that the outbreak was associated with a seafood market in Wuhan, China, as reported by local health authorities (1). The results of the investigations led to the identification of a new coronavirus in affected patients (2). Following its identification on the 7th of January 2020 by the Chinese Center for Disease Control and prevention (CCDC), the new virus and the disease were officially named SARS-CoV-2 (for Severe Acute Respiratory Syndrome CoronaVirus-2) and COVID-19 respectively by the World Health Organization (WHO) (3). On March 11, 2020, WHO publicly announced the SARS-CoV-2 epidemic as a global pandemic.

This virus is likely to remain and continue to spread unless an effective vaccine is developed or a high percentage of the population is infected in order to achieve collective immunity. The development of a vaccine is a long process and is not guaranteed for all infectious diseases. Indeed, some viruses such as influenza and HIV have a high rate of genetic mutations which makes them prone to antigenic leakage (4,5). It is therefore important to assess the genetic evolution of the virus and more specifically the regions responsible for its interaction and replication within the host cell, in order to predict whether a vaccine could be effective. The identification of conserved and variable regions in the virus could help guide the design and development of anti-SARS-CoV-2 vaccines.

The SARS-CoV-2 is a single-stranded positive-sense RNA virus belonging to the genus *Betacoronavirus.* 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 (6). The ORF1ab polyprotein is covering two thirds of the viral genome and cleaved into many nonstructural proteins (nsp1 to nsp16). The third part of the SARS-CoV-2 genome codes for the main structural proteins (spike S, envelope E, nucleocapsid N and membrane M). In addition, six ORFs, namely ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10, are predicted as hypothetical proteins with no known function (7).

Protein S is the basis of most candidate vaccines, it binds to membrane receptors in host cells via its RBD and ensures a viral fusion with the host cells (30). Its main receptor is the angiotensin-converting enzyme 2 (ACE2), although another route via CD147 has also been described (8,9). The glycans attached to protein S assist the initial attachment of the

virus to the host cells and act as a coat that helps the virus to evade the host's immune system. In fact, A previous study has showed that glycans cover about 40% of the surface of the spike protein, however, the ACE2-RBD was found to be the largest and most accessible epitope (10). Thus, it may be possible to develop a vaccine that targets the spike RBD provided it remains accessible and stable over the time. Hence the importance of monitoring the introduction of any mutation that could compromise the potential effectiveness of a candidate vaccine.

In this study, we assessed the genetic diversity and determined the stability of SARS-CoV-2 in 30,983 complete genomes. A particular focus on the geographic distribution of genetic variants and its implications for vaccine design and development have been also reported.

Materials and Methods

Data collection and Variant calling analysis

Full-length viral nucleotide sequences of 30,983 SARS-CoV-2 genomes were collected from the GISAID EpiCovTM (update: 26-05-2020) (11) (**Additional file 1: Table S1)**. The genomes were obtained from samples collected from December 24, 2019 to May 13, 2020. Genome sequences were mapped to the reference sequence Wuhan-Hu-1/2019 (Genbank ID: NC_045512.2) using Minimap v2.12-r847 (12). The BAM files were sorted by SAMtools sort (13). The final sorted BAM files were used to call the genetic variants in variant call format (VCF) by SAMtools mpileup and BCFtools (13). The final call set of the 30,983 genomes, was annotated and their impact was predicted using SnpEff v 4.3t (14). For that, the SnpEff databases were first built locally using annotations of the reference sequence Wuhan-Hu-1/2019 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).

Divergence analysis

To calculate the divergence of the SARS-Cov-2 genomes from the reference Wuhan-Hu-1 genome, the 30,983 genomes were sorted by the continent and country of origin. The divergence was first calculated by estimating the similarities of the genomes with the reference by grouping all the genomes within each country using CD-Hit (15). The percentage of similarity was then recovered to 100%.

To calculate the percentage of divergence by the country we used the following formula:

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

$A = Similarity\ percentage$
$B = Number\ of\ genomes\ with\ Similarity\ value\ equal\ A$
$C = Tatal\ genomes\ by\ country \wedge countinent$
$D = Percentage\ of\ Divergence$

Phylogenetic and spatio-dynamic analysis

We generated a phylogenetic and, divergence trees, as well as a genomic epidemiology map of 1,025 SARS-CoV-2 isolates (**Additional file 1: Table S2)** using NextStrain tools

(https://nextstrain.org) (16), Maximum-likelihood trees were inferred with IQ-TREE85 v1.5.5 (17) under the GTR model. The evolutionary rate and time to the Most Recent Common Ancestor (TMRCA) was estimated using ML dating in the Tree Time package (18). The choice of the isolates was based on geographic location and date, from each continent, approximately 200 genome sequences were retrieved from GISAID database. The strains were chosen based on the date of isolation (one genome per week). A total 1200 genome was retrived at first and further filtered by length to get the final number of 1,025 genome.

D614G mutagenesis analysis

To investigate the possible impact of the most frequent D614G mutation, we conducted an *in-silico* mutagenesis analysis based on the CryoEM structure of the spike protein in its pre-fusion conformation (PDB id 6VSB). Modelling of the D614G mutation was done using UCSF Chimera (19). Then, the mutant structure was relaxed by 1000 steps of steepest descent (SD) and 1000 steps of conjugate gradient (CG) energy minimizations keeping all atoms with more than 5A from G614 fixed. Comparative analysis of D614 (wild type) and G614 (mutant) interactions with their surrounding residues was done in PyMOL 2.3 (Schrodinger L.L.C).

RBD mutations and Spike/ACE2 binding affinity

Modelling of RBD mutations was performed by UCSF chimera (19) using the 6M0J structure of the SARS-CoV-2 wild-type spike in complex with human ACE2 as a template. Mutant models were relaxed by 1000 steps of SD followed by 1000 steps of CG minimizations keeping all atoms far by more than 5A from the mutated residues fixed. Changes in the binding affinity of the spike/ACE2 complex for each spike mutant were estimated by the MM-GBSA method using the HawkDock server (20).

Results

Diversity of genetic variants of SARS-CoV-2 in different geographic areas

In order to characterize and monitor the appearance and the accumulation of SARS-CoV-2 mutations, we analysed 30,983 genomes belonging to the six geographical zones (according to GISAID database) and distributed in 80 countries as follow : 214 from Africa, 368 from South America, 1,590 from Oceania, 2,111 from Asia, 6,825 from North America and 19,875 from Europe. The genomes were obtained from samples collected from December 24, 2019 to May 13, 2020.

Among all the genomes analyzed, only 778 (2.5%) were invariants and were particularly native to Asia, representing 10.8% of the total Asian genomes, followed by North America and Europe with 2.38% and 1.87%, respectively. For the other genomes, 97.5% shows at least one mutation. A total of 3,206 variant sites were detected compared to the reference genome Wuhan-Hu-1/2019 (NC_045512.2).

We have analyzed the type of each mutation, highlighting the prevalence of these mutations both in all genomes (worldwide) and in each of the geographical areas studied (**Fig 1**). Our finding revealed that 67.96% of mutations were non-synonymous (64.16% have missense effects, 3.77% produce a gain or loss of stop codon and 0.33% produce a loss of start codon), 28.60% were synonymous while 3.43% of mutations were located in the intergenic regions, mainly in the untranslated regions (UTR). The comparison of prevalence of types of mutations in the six geographical areas shows a similar trend and that the distribution of these events was largely homogeneous.

Remarkably, in the set of the identified mutations, only 182 (5.67%) were recurrent, with a frequency greater than 1% of the sequences, in the six geographical areas (**Fig 2**). The distribution of these recurrent mutations along the viral genome revealed that 10 genes at least show one of these mutations. ORF1ab harbored more than half of them (58.79%), followed by proteins S and N, with 10.44% in each. Thus, the orf3a protein displayed 6% of mutations, while the orf8, M, orf7a and orf6 proteins contained 3.30%, 2.75%, 1.65% and 1.10% of mutations, respectively, whereas, proteins E and orf7b harbored only 0.55% of mutations each (**Fig 2**). Regarding the geographic distribution of recurrent mutations, we identified 72 mutations in Oceania, 68 mutations in Africa, 54 mutations in Asia, 51 mutations in South America, 46 mutations in North America and 31 mutations in Europe. Among the 182 recurrent mutations, 116 (63.73%) had a non-synonymous effect. Among them, 73 were observed in the ORF1ab gene and distributed in thirteen non-structural proteins (**Fig 3**), including nsp3-Multi domain, nsp2 and nsp12-RNA

dependent RNA polymerase (RdRp) having more mutations, with 18, 12 and 10, respectively. It is also interesting to note that none of the nine recurrent non-synonymous mutations identified in the S protein were located in RBD.

Geographical distribution of the SARS-CoV-2 hotspot mutations

Analysis of all sequences revealed fourteen non-synonymous mutations with significant spread (> 10%) in each of six geographical areas that were considered as hotspot mutations (**Fig 4**). Eight mutations of them are present in the ORF1ab gene and are distributed in seven regions coding for nsp2 (T265I), nsp3 (T2016K), nsp6 (L3606F), nsp12 (T5020I and T4847I), nsp13 (M5865V), nsp14 (D5932T) and nsp15 (Ter6668W). Moreover, three mutations in protein N (R203K, G204R and P13L) and one mutation in each of the three proteins; S (D614G), orf3a (Q57H) and orf8 (L84S).

Remarkably, we found a different pattern of mutations in the six geographical regions. Only two hotspot mutations were common in the six geographical regions: the high-frequency D614G (in S) mutation found in 40% of the Asian genomes at 80% and the Q57H mutation (in orf3a) in the African genomes. In addition, seven hotspot mutations were uniques to a single geographical region. Among them, two mutations T2016K (in nsp3) and P13L (in N) were predominant in Asia, two mutations M5865V (in nsp13) and D5932T (in nsp14) in North America, one T4847I (in nsp12) in Europe, one Ter6668W (in nsp 15) in South America and one T5020I (in nsp12) in Africa. However, the other five hotspot mutations were variable between the six geographical regions, including two R203K and G204R (in N) that were particularly predominant in the African, European, South American and Oceanian regions. Whereas, L3606F (in nsp6) was common in African, European, Asian and Oceanian regions. Thus, L84S (in orf8) was common in the Asian region, North America and the Oceania. In addition, T265I (in nsp2) was common in Asia, North America, South America and Oceania.

The distribution of hotspot mutation patterns of SARS-CoV-2 over time

In order to determine the appearance of each mutation, we analyzed each genome in each geographical area over time, classifying them per date of sample collection (according to GISAID database). We have noticed that the distribution of the fourteen hotspot mutation patterns is dynamic during the propagation of SARS-CoV-2 in different geographic regions. Of the fourteen hotspot mutations, four appeared for the first time during the

month of January 2020, seven during the month of February, while the remaining three mutations were observed from the first two weeks of March (**Fig 5**).

The high frequency D614G mutation (in S) appeared in January 24, 2020 in the Asian region (in China), after a week it was also observed in Europe (Germany), then gained dominance over time due to its spread in several European countries. In February this mutation was also reported in Africa and South America to spread to North America and Oceania in the middle of February.

The L84S mutation (in orf8) appeared in early stages in January 05 in Asia (China), then the emergence of this mutation was observed in the second half of February in the North American and Oceanic regions, particularly in the USA and Australia, respectively. In addition, L3606F (in nsp6) was emerged from the third week of January in Asia (in China), then was propagated in late February in Europe (in Italy and France), in the middle of February in Oceania and spread for the first time in March 20 in African regions (DRC and Senegal). While the mutation D5932T (in nsp14) appeared in January 19 particularly in North America (USA).

Among the mutations that emerged for the first time in February, triple mutations in protein N (R203K, G204R and P13L) have been observed to gain prominence during the pandemic. Interestingly, the emergence time of the R203K and G204R mutations was similar in Africa, Europe, South America and Oceania. Indeed, both mutations appeared in 24th of February in Europe (Netherlands and Austria) and three days later (27th of February) they raised in Africa (in Nigeria). Moreover, the P13L mutation emerged in the Asian region only from the end of February. We also noticed an increase in the frequency of the Q57H mutation (in orf3a) in late February in Africa (Senegal), Europe (in France and Belgium) and North America (in USA and Canada), then was observed in the three other regions at the beginning of March; Asia (Taiwan), Oceania (Australia) and South America (Brazil). Thus, we found that the T265I (in nsp2) mutation first appeared in late February in both Asia (Taiwan) and North America (USA), while it was spread at the beginning of March in South America (Brazil) and Oceania (Australia). In addition, the T4847I mutation (in nsp12) gained predominance from the second week of February in Europe (United Kingdom), then after one month spread to the oceanic region (in Australia). While the frequency of the M5865V mutation (in nsp13) has increased particularly in North America (USA).

For the three hotspot mutations that emerged in the first half of March, which was specific to a geographic region, including T2016K (in nsp3) their frequency increased from

04-March only in Asia (Taiwan). Likewise, the Ter6668W mutation (in nsp15) gained predominance only in South America from 09 March in Asia (Chile), while T5020I (in nsp12) appeared in mid-March in Africa (Senegal).

Mutagenesis of D614G and impact of RBD mutations on the binding ability of spike to ACE2

As shown the Figure 6, the non-synonymous D614G mutation did not have an impact on the two- or three-dimensional structure of the spike glycoprotein. However, D614 residue in the wild type spike is involved in three hydrogen bonds; one with A647 in the same subunit (S1) and two bonds with THR-859 and LYS-854 located at S2 subunit of the adjacent protomer (**Fig 6A**). The substitution of D614 by G in the mutant spike resulted in the loss of the two hydrogen bonds with THR-859 and LYS-854 in the S2 subunit of the adjacent protomer (**Fig 6B**). Such modification could result in a weak interaction between S1 and S2 subunits and thus increases the rate of S1/S2 cleavage which will improve the virus entry to host cells.

To evaluate the effect of RBD mutations on the binding affinity of the spike protein to ACE2, the MM-GBSA method was employed to calculate the binding affinity of 35 spike mutants to ACE2. Four mutations potentially enhance the binding affinity of spike/ACE2 complex by more than -1.0 kcal/mol, while nine were shown to potentially reduce its affinity by more than 1.0 kcal/mol (**Table 1**). However, the remaining 22 did not significantly affect the binding affinity of spike to ACE2.

Clustering and divergence of SARS-CoV-2 genomes in six different regions

In order to associate countries sharing the same trends in the frequencies of recurrent mutations; therefore probably having the same rate of accumulation of mutations, first of all, we individually normalized these frequencies for each country. Then we calculated a distance matrix using the Jaccard method. We observed a formation of two main clusters (cluster 1 and cluster 2) (**Fig 7**), each subdivided into sub-cluster (SC) and harbored from countries belonging to the six geographical areas. Our finding revealed that, the countries of cluster 2 were closer to each other compared to cluster 1. We also observed fifteen SCs with a significant distance of less than 0.5 (**Fig 7, Table 2**). Remarkably, most of the countries grouped in SC are close geographically. For example, Kazakhstan and Georgia, two geographically close countries were grouped in SC-2 and are having a significant distance of 0.22.

As shown in (**Figure 8A)** the genome of the circulating strain seems to be more divergent in Africa followed by South America and Asia while it is the closest to the reference strain in North America, followed by Europe and Oceania.

The highest percentage of divergence in Europe, North America, Asia, South America, and Africa was observed in England (1.93%) Canada (0.2%), Philippine (6.97%), Ecuador (3,5%) and Ghana (6.88%), respectively. While the lowest percentage was shown in Lithuania (0,01), USA (0,06%), Nepal (0.03%), Chile (0.06%), Guam (0.36%) and Egypte (0,08%). Interestingly, in Africa, the North African SARS-CoV-2 genome showed high similarity with that of south African Countries. However, this could be difficult to evaluate as some countries could have a high rate of infections with a virus that already has a high rate of divergence (and vice versa) in addition to the limited number of genomes sequences deposit in the database.

The phylogenetic divergence tree **(Fig 8B)** shows that the highest and lowest divergence rate was among genomes originally from Asia, followed by these from Europe and Africa, which is concordant with the similarity study. Using the clade nomenclature from the Nextstrain, we can identify two main clads; first one “A2” from Europe which called A2 which is the most divergent and the Asian Clad “ B” less divergent. Nonetheless, genomes from America and Africa were scattered through the Tree with no specific clad.

Divergence in the subclade I varied from 0.04300 to 0.0432, with the lowest rate in South Africa and DRC the latter also genomes with the highest divergence rate. The strains in this clad were mainly close to genomes from North America and Europe, which harbored a mutation C15324T.

The second subclad, still contains the highest divergence in African genomes. Just like the first clad, all the genomes cluster with strains from Europe or North America. The highest divergence is observed in Senegal and Ghana. North African genomes on the other hand showed divers clustering pattern, Algeria clustered close to North America and Europe, while genomes from Egypt clustered with Asian genomes from Arab countries such as Saudi Arabia and Jordan.

The distribution of genomes over times shows that the majority of sequences belongs to a period from March to May and suggests that the African genome is are more diverses since they went through many generation cycles before getting to Africa.

Phylogentics and spatio dynamics of SARS-CoV-2

The topology of the maximum likelihood phylogenetic tree **(Fig 9)** shows a clear clustering: First Cluster containing mainly strains from Asia while the second contains strains from Europe sharing the clade specific mutations at the position C18060T and C14408T respectively.

For each cluster we identified different clades: cluster I containing two main clades A1a and B1 harboring mainly strains from Asia, North America, and Asia, Europe, respectively. However, cluster II harbored three clades: B2, A2, A2a without a specific pattern.

The distribution of African genomes across the phylogenetic tree showed a close relationship with different continents. In the first clade, African genomes, mainly from south West Africa, clustered with Asia and showed the lowest divergence rate. Meanwhile, genomes clustering in the European clade were distributed into three sub clades all sharing the three-pattern mutations mostly common in Europe : G28881A, G28882A, and G28883C.

The map (**Additional file 1: Fig S1**) showing the spatio-dynamics of SARS-Cov-2 provides an insight into the viral strains potential geographical origin based on the sample used. The map shows a complex and interconnected network of strains. From these samples, strains from Asia appear to have diverged and resulted in other strains in all the investigated regions. European strains seem to have given rise to strains in North and South America and Africa, with multiple divergent strains within Europe itself. Similarly, with less frequency, strains from South and North America appear to be related to some divergent strains in Europe and Asia. African strains could be the source of a few, mainly in Europe.

Discussion

Due to the rapid spread and mortality rate of the new SARS-CoV-2 pandemic, the development of an effective vaccine against this virus is of a high priority (21). The availability of the first viral sequence derived during the COVID-19 epidemic, Wuhan-Hu-1, was published on January 5, 2020. From this date, numerous vaccination programs were launched (22,23). Furthermore drugs and vaccines should target relatively invariant and highly constrained regions of the SARS-CoV-2 genomes, in order to avoid drug resistance and vaccine escape (24). For this, monitoring genomic changes in the virus is essential and plays a pivotal role in all of the above efforts, due to the appearance of genetic variants which could have an effect on the efficacy of vaccines. In this study, we investigated the genetic diversity of the 30,983 complete SARS-CoV-2 genomes isolated from six geographic areas. Our results showed three different situations of the mutations identified; (i) mutations which have developed and are gaining a predominance in the six geographical areas; (ii) mutations which have been predominant only in certain geographical regions and (iii) mutations that are apparently expanding, but of low frequency in all the isolates studied.

A total of 3,206 variant sites were identified, of which only 5.67% recurrent with a frequency greater than 1% of all assessed sequences. While 94.33% of the mutations found at low frequency, including 49.68% unique (specific mutation to one genome). These results indicate that, so far, SARS-CoV-2 has evolved through a non-deterministic process and that random genetic drift has played the dominant role in the spread of unique mutations around the world, suggesting that the SARS-CoV-2 is not yet adapted to its host, and epidemiological factors are particularly responsible for the global evolution of this virus. Although some hotspot mutations have increased their frequencies, they have been systematically gained their predominance outside Asia. For example, the hotspot mutation Q57H (in orf3a), in February 21, 2020, has not yet been observed among isolates from China, while it has emerged in Europe and also spread in isolates from North America. Likewise, we were able to identify seven other hotspot mutations (> 10%) which gained their predominance over time in the SARS-CoV-2 genomes outside Asia; the double mutations R203K-G204R of the protein N (in Europe, South America and Oceania), M5865V of nsp13 and D5932T of nsp14 (in North America), T5020I of nsp12 (in Africa), T4847I of nsp12 (in l 'Europe and Oceania), Ter6668W from nsp15 (in South America).

It is interesting to note that the orf10 protein was the only invariant protein in recurrent mutations. In our recent study (25), we reported the absence of this protein in certain SARS-CoV-2 genomes, which could explain a rate of low-frequency mutations of this protein.

The analysis of the genomes divergence showed the highest divergence in Africa and South America and the lowest in North America and Europe. Interestingly, while china is the source of the original variant, it had a higher divergence rate than most countries in Europe and North America including the most affected ones like Italy, Spain, France and USA. This might suggest an early and fast transmission of the infections to Europe and North America by a variant that is genetically close to the original strain. A fast transmission would mean a single source leading to multiple infections thus giving the virus fewer life cycles to change. When the spread of the disease is slow, the virus would go through multiple propagation cycles and different hosts and thus is likely to be more divergent than the original variant. Indeed, the disease affected Africa and South America later on, likely with variants that are more divergent. This is in line with a previous study describing a continuous tendency of the virus to diverge over time (26). Moreover, it has been shown that mutational accumulations seems to assist the virus adaptation to the human host by optimization of codon usage through synonymous mutations (27).

However, genomes clustering did not have a clear pattern regarding their geographical distributions. The regions studied seem to have strains that were introduced from different areas, reflecting the migration and globalization effect as reported before (28, 29).

The phylogenetic tree confirms that the disease transmission to Africa has probably three sources. The least divergent variants were clustered with Asia while the most divergent ones clustered with Europe and North America. This distribution points to different sources of infection; countries in Western and Northern Africa likely had their carriers coming from European countries.

As the virus spreads more widely around the world, it is important to follow appearance of new mutations development that could compromise the effectiveness of a candidate vaccine. The SARS-CoV-2 S protein is the primary target of most vaccine candidates that are under development, due to their key role in mediating entry of the virus and is known to be immunogenic (30,31). Analysis of S protein in the SARS-CoV-2 genomes that we collected in this study revealed a hotspot mutation at position 614 that involves the

change from a large amino acid residue (aspartic acid) to a small hydrophobic one (glycine). This mutation was most frequent in the genomes of six geographical areas, with a frequency of 40% of the Asian genomes to 88% of the African genomes. Several studies had already indicated its emergence (25, 32-35). The D614G mutation is proximal to the S1 cleavage domain of the spike glycoprotein (36). Our finding showed that this mutation induces a loss of two hydrogen bonds between the S1 and S2 subunits of neighboring protomers and may therefore increase the cleavage rate of these subunits in the pre-fusion state of the spike protein to allow its conformational transition to the post-fusion state that is associated with membrane fusion during virus entry (37). However, our structural modeling of this mutation has shown no substantial impact on the secondary or tertiary structure of the spike protein which makes the likelihood of this variant to significantly affect a potential epitope in this region very low. From a clinical perspective, there is in fact no evidence that the SARS-CoV-2 gene mutation D614G affects its transmissibility or virulence since there was no association between this D614G substitution and the hospitalization status of the patients (38). With this in mind, we believe that this will not present a serious issue for vaccine development and that a universal candidate vaccine for all circulating strains of SARS-CoV-2 may be possible.

On the other hand, the RBD of S protein, allows the virus to bind to the ACE2 receptor on host cells (39,40). Mutations in this receptor are a likely route to escape recognition of antibodies, as described in other viruses (41,42). In more than 30,000 genomes analyzed, only 35 non-synonymous mutations in RBD have been identified. These mutations observed at low frequency> 1% in all isolates, the calculated binding free energy of mutant RBDs of spike protein in complex with ACE -2 demonstrate that in 89% (n = 31) of the cases these mutations decrease or does not significantly affect the ability of the virus to bind to the host cells receptor. While only 11% (n = 4) potentially enhance the binding of the viral spike protein to ACE-2 and therefore could influence the pathogenicity of SARS-CoV-2.

SARS-CoV-2 has only been identified recently in the human population; a short time compared to the adaptive processes that could take years to occur. Although we cannot predict whether adaptive selection will be observed in SARS-CoV-2 in the future, the main conclusion is that the currently circulating SARS-CoV-2 viruses constitute a homogeneous viral population.

This study is a comprehensive report on the amino acid mutations that spread during the SARS-CoV-2 pandemic, from December 24 to May 13. We can therefore be cautiously optimistic that, up to now, genetic diversity should not be an obstacle to the development of a broad protection SARS-CoV-2 vaccine. Therefore, ongoing monitoring of the genetic diversity of SARS-CoV-2 will be essential to better understand the host-pathogen interactions that can help inform drug and vaccine design.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

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

This work was carried out under National Funding from the Moroccan Ministry of Higher Education and Scientific Research (Covid-19 Program) to AI. This work was also supported by a grant to AI from Institute of Cancer Research and the PPR-1 program to AI.

References

1. Mackenzie JS, Smith DW. COVID-19: a novel zoonotic disease caused by a coronavirus from China: what we know and what we don't. Microbiology Australia. 2020;41(1):45-50.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine. 2020.

3. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery. 2020.

4. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely high mutation rate of HIV-1 in vivo. PLoS biology. 2015;13(9).

5. Rouse BT, Sehrawat S. Immunity and immunopathology to viruses: what decides the outcome? Nature Reviews Immunology. 2010;10(7):514-26.

6. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell host & microbe. 2020.

7. Malik YA. Properties of Coronavirus and SARS-CoV-2. The Malaysian Journal of Pathology. 2020;42(1):3-11.

8. Jaramillo CB, Cevallos D, Sanches-SanMiguel H, Unigarro L, Zalakeviciute R, Gadian10 N, et al. Clinical, molecular and epidemiological characterization of the SARS-CoV2 virus and the Coronavirus disease 2019 (COVID-19), a comprehensive literature review.

9. Wang K, Chen W, Zhou Y-S, Lian J-Q, Zhang Z, Du P, et al. SARS-CoV-2 invades host cells via a novel route: CD147-spike protein. BioRxiv. 2020.

10. Grant OC, Montgomery D, Ito K, Woods RJ. Analysis of the SARS-CoV-2 spike protein glycan shield: implications for immune recognition. bioRxiv. 2020.

11. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017;22(13).

12. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.

13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

14. 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. 2012;6(2):80-92.

15. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9.

16. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121-3.

17. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-74.

18. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus evolution. 2018;4(1):vex042.

19. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry. 2004;25(13):1605-12.

20. Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020.

21. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science. 2020;368(6493):860-8.

22. Amanat F, Krammer F. SARS-CoV-2 vaccines: status report. Immunity. 2020.

23. Zhang J, Zeng H, Gu J, Li H, Zheng L, Zou Q. Progress and prospects on vaccine development against SARS-CoV-2. Vaccines. 2020;8(2):153.

24. Tu Y-F, Chien C-S, Yarmishyn AA, Lin Y-Y, Luo Y-H, Lin Y-T, et al. A review of SARS-CoV-2 and the ongoing clinical trials. International journal of molecular sciences. 2020;21(7):2657.

25. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri M, Hakmi M, Essabbar A, et al. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geodistribution and a rich genetic variations of hotspots mutations. bioRxiv. 2020.

26. Sheikh JA, Singh J, Singh H, Jamal S, Khubaib M, Kohli S, et al. Emerging genetic diversity among clinical isolates of SARS-CoV-2: Lessons for today. Infection, Genetics and Evolution. 2020:104330.

27. Dilucca M, Forcelloni S, Giansanti A, Georgakilas A, Pavlopoulou A. Temporal evolution and adaptation of SARS-COV 2 codon usage. bioRxiv. 2020.

28. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences. 2020;117(17):9241-3.

29. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 2020:104351.

30. Du L, He Y, Zhou Y, Liu S, Zheng B-J, Jiang S. The spike protein of SARS-CoV—a target for vaccine and therapeutic development. Nature Reviews Microbiology. 2009;7(3):226-36.

31. Le TT, Andreadakis Z, Kumar A, Roman RG, Tollefsen S, Saville M, et al. The COVID-19 vaccine development landscape. Nature reviews Drug discovery. 2020;19(5):305-6.

32. Kiyotani K, Toyoshima Y, Nemoto K, Nakamura Y. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. Journal of Human Genetics. 2020;65(7):569-75.

33. Eaaswarkhanth M, Al Madhoun A, Al-Mulla F. Could the D614 G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? International Journal of Infectious Diseases. 2020.

34. Lorusso A, Calistri P, Mercante MT, Monaco F, Portanti O, Marcacci M, et al. A “One-Health” approach for diagnosis and molecular characterization of SARS-CoV-2 in Italy. One Health. 2020:100135.

35. Stefanelli P, Faggioni G, Presti AL, Fiore S, Marchi A, Benedetti E, et al. Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. Eurosurveillance. 2020;25(13):2000305.

36. Veljkovic V, Perovic V, Paessler S. Prediction of the effectiveness of COVID-19 vaccine candidates. F1000Research. 2020;9(365):365.

37. Tang L, Schulkins A, Chen C-N, Deshayes K, Kenney JS. The SARS-CoV-2 Spike Protein D614G Mutation Shows Increasing Dominance and May Confer a Structural Advantage to the Furin Cleavage Domain. Preprints. 2020.

38. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell. 2020.

39. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, et al. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for

development of RBD protein as a viral attachment inhibitor and vaccine. Cellular & molecular immunology. 2020;17(6):613-20.

40. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581(7807):221-4.

41. Wong AH, Tomlinson AC, Zhou D, Satkunarajah M, Chen K, Sharon C, et al. Receptor-binding loops in alphacoronavirus adaptation and evolution. Nature communications. 2017;8(1):1-10.

42. Rockx B, Donaldson E, Frieman M, Sheahan T, Corti D, Lanzavecchia A, et al. Escape from human monoclonal antibody neutralization affects in vitro and in vivo fitness of severe acute respiratory syndrome coronavirus. The Journal of infectious diseases. 2010;201(6):946-55.

Table 1: Table represent the effect of mutations on the binding affinity of spike RBD with ACE2 evaluated by MMGBSA binding free energy calculation (ΔG_{Bind}). Only mutations resulting in ΔG_{Bind} modification by more than 1.0 kcal/mol are shown.

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

ΔΔG: binding free energy difference between mutated and wild type complexes.

Table 2: Correlation matrix heatmap showing the Jaccard distance between countries based on the SARS-CoV-2 mutation frequencies. Only the distance between countries <0.5 are shown.

Sub-Cluster	**Cluster**	**Countries**	**Distance**	**Geographic areas**
SC-1	Cluster 1	Brunei - Guam	0.22	Asia, Oceania
SC-2	Cluster 2	Kazakhstan - Georgia	0.22	Asia
SC-3	Cluster 2	Nigeria - Serbia - Croatia – Ireland - Peru	0.26	Africa, Europe, South America
SC-4	Cluster 2	Vietnam - Jordan	0.27	Asia
SC-5	Cluster 2	Sri Lanka - Kuwait	0.3	Asia
SC-6	Cluster 2	Greece - Portugal	0.32	Europe
SC-7	Cluster 2	Singapore - Thailand	0.35	Asia
SC-8	Cluster 2	Finland - Poland	0.35	Europe

SC-9	Cluster 2	Slovenia - Jamaica	0.35	Europe, North America
SC-10	Cluster 2	Denmark - Iceland	0.36	Europe
SC-11	Cluster 2	Germany - Russia	0.36	Europe
SC-12	Cluster 1	Hungary - Latvia	0.41	Europe
SC-13	Cluster 1	Chile - Brazil	0.43	South America
SC-14	Cluster 1	Iran - Pakistan	0.43	Asia
SC-15	Cluster 2	Netherland - Belgium - Austria	0.49	Europe

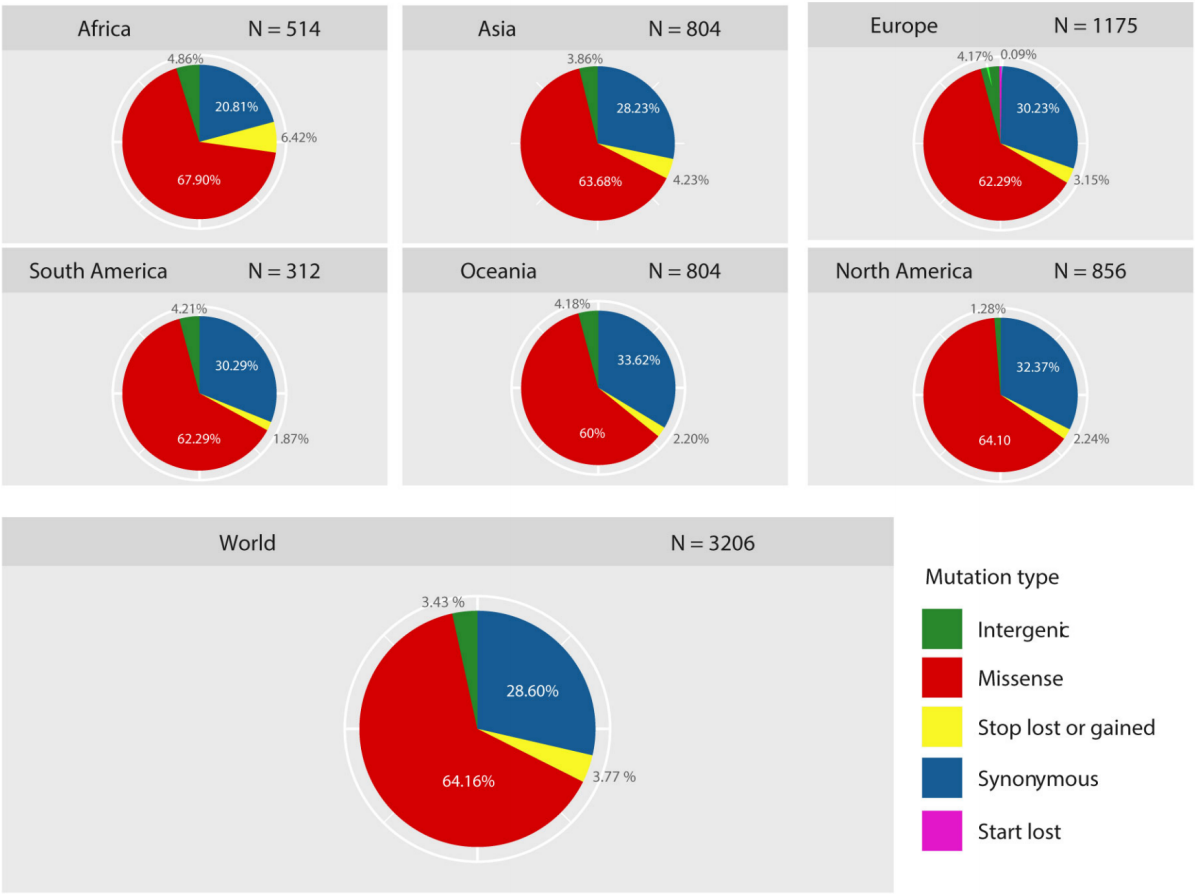

Figure 1: The prevalence of mutations effects of SARS-CoV-2 in different geographic regions. Pie charts showing the global and stratified continent-wide distribution of the SARS-CoV-2 mutation types. A total of 30,983 SARS-CoV-2 genomes were compared to the reference Wuhan-Hu-1 genome. The prevalence distribution of each type of mutation is largely uniform in the six geographic areas.

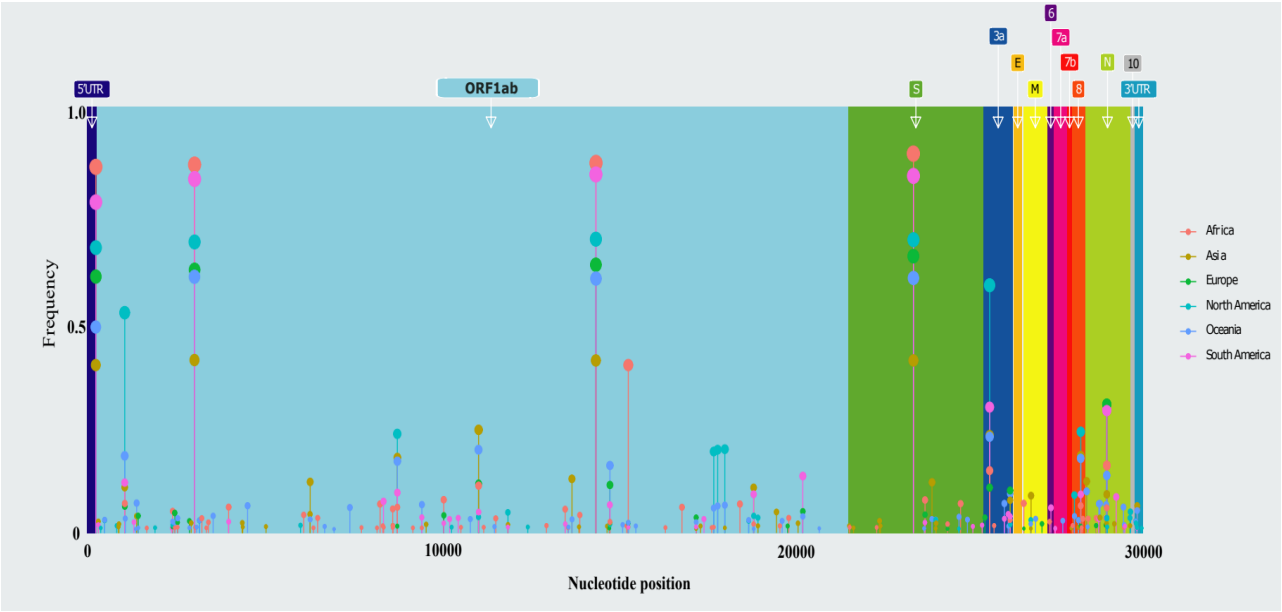

Figure 2: The frequency distribution of each recurrent mutation of SARS-CoV-2. Lollipo plots showing the location and frequency of recurrent mutations > 1% of the genomes in each geographic region. Areas of the SARS-CoV2 gene harboring recurrent mutations are reported in the legend. All types of mutations are included (Non-syninymous, synonymous and intergenic). Abbreviations: Spike (S), Envelope (E), Membrane (M), Nucleocapsid (N), ORF3a (3a), ORF6 (6), ORF7a (7a), ORF7b (7b), orf8 (8).

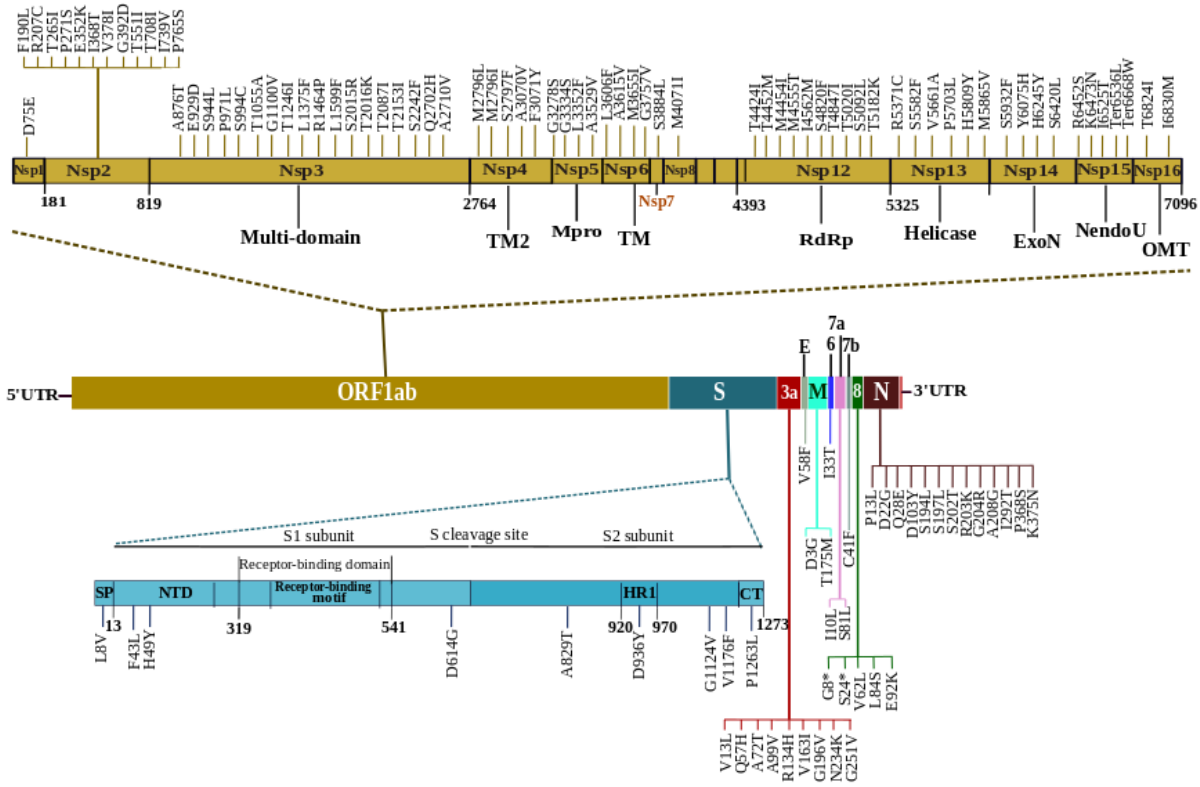

Figure 3: Schematic representation illustrating the distribution of non-synonymous recurrent mutations along the SARS-CoV-2 genome. 116 Recurrent non-synonymous mutations > 1% of the genomes are represented by vertical lines. Abbreviations: Non-Structural Protein (NSP) Spike (S), Envelope (E), Membrane (M), Nucleocapsid (N), ORF3a (3a), ORF6 (6), ORF7a (7a), ORF7b (7b), orf8 (8), peptide signal (SP), N-terminal domain (NTD), Heptad repeat 1 (HR1), cytoplasmic tail (CT).

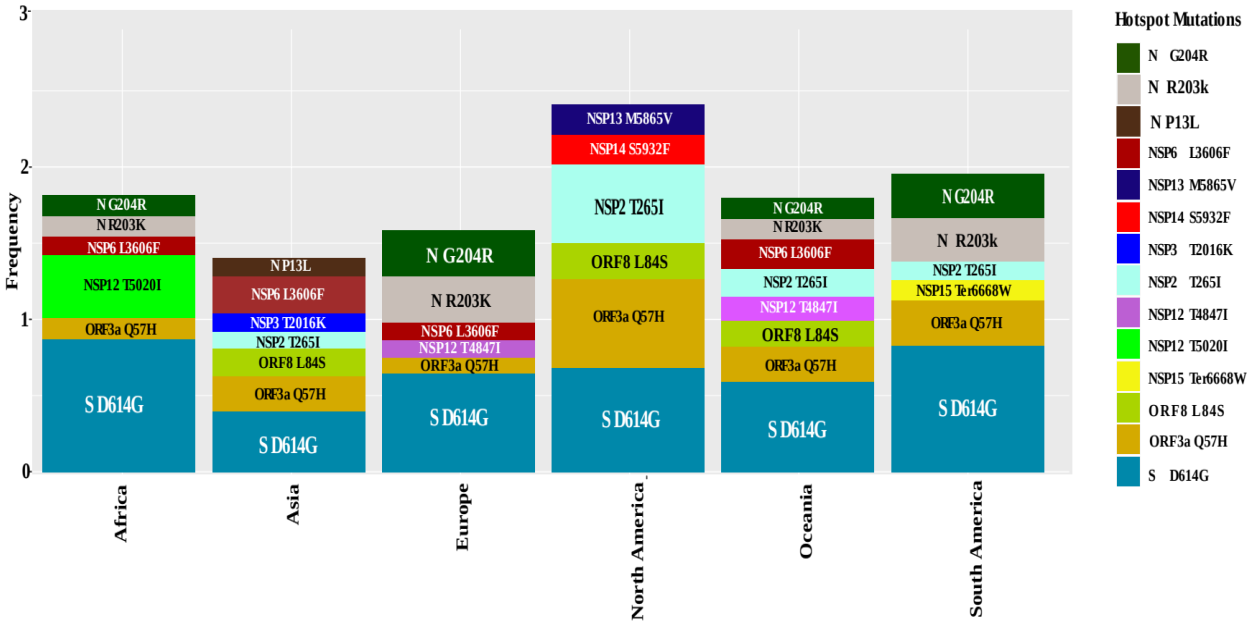

Figure 4: Frequency of hotspot mutation of SARS-CoV-2 in different geographic areas. Fourteen hotspot mutations were distributed in six geographic areas with a frequency> 10% of the genomes in each region; Africa (n = 6), Asia (n = 7), Europe (n = 6), North America (n = 6), Oceania (n = 8), South America (n = 6). The frequency of mutations was calculated for each of these mutations, by normalizing the number of genomes hosted per a given mutation in a geographic area, per the overall number of genomes recovered by geographic area. Abbreviations: Non-structural protein (NSP), Spike (S), Nucleocapsid (N).

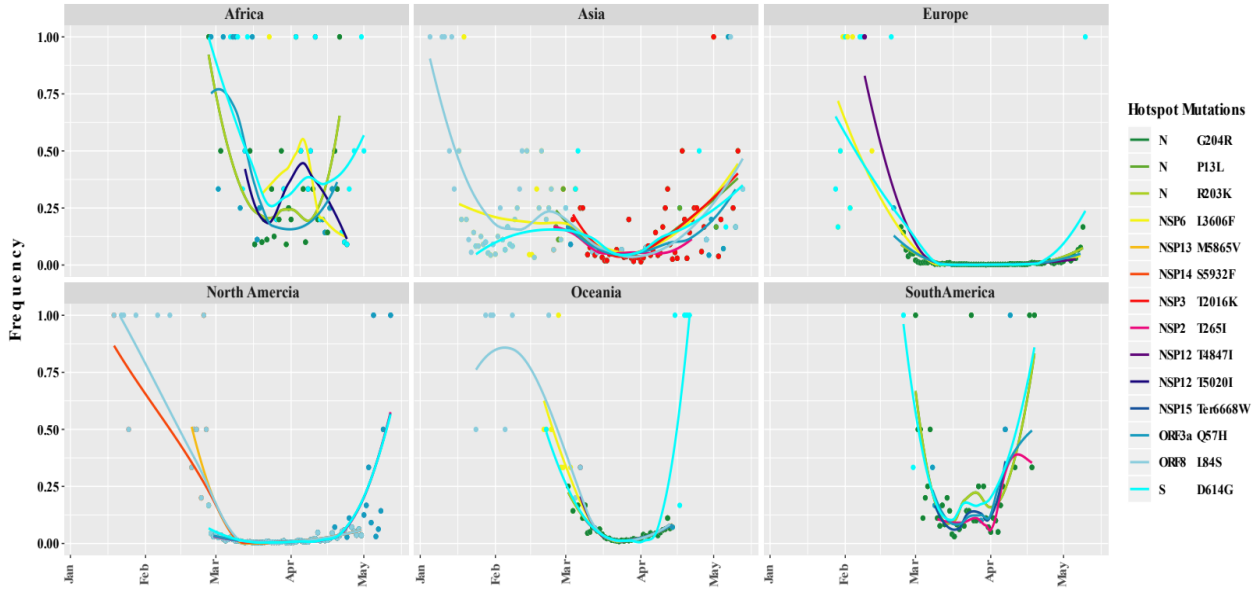

Figure 5: Tracking of hostpot mutations over time in each geographic area. We have represented the resulting frequencies in points and introduced estimates of the correlation curve based on the nearest neighbor using the regression loss function. Curves showing the time trends of the emergence and prevalence of each SARS-CoV-2 mutation hotspot over time. The axis of X represents the time measured in months and the Y axis represents the frequencies of the genomes hosting a given mutation.

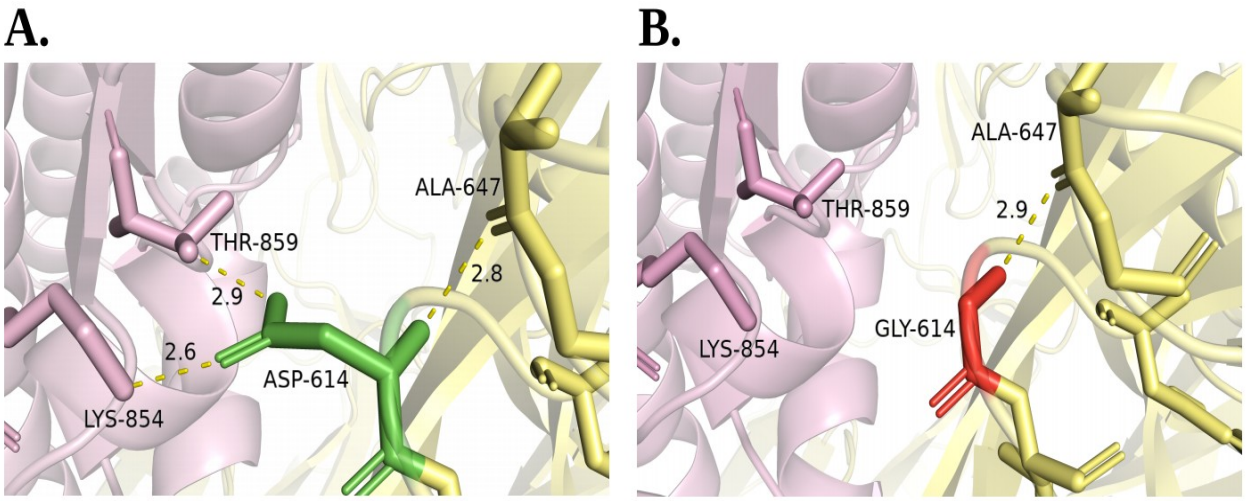

Figure 6: Comparison of spike wild type residue ASP-614 (A) and the mutated GLY-614 (B). ASP-614 (green sticks) in subunit S1 (yellow) is involved in two hydrogen bonds with THR-859 and LYS-854 from the S2 subunit (pink). The substitution of ASP by GLY at position 614 causes the loss of the two hydrogen bonds between S1 subunit and THR-859 and LYS-854 in the S2 subunit (pink).

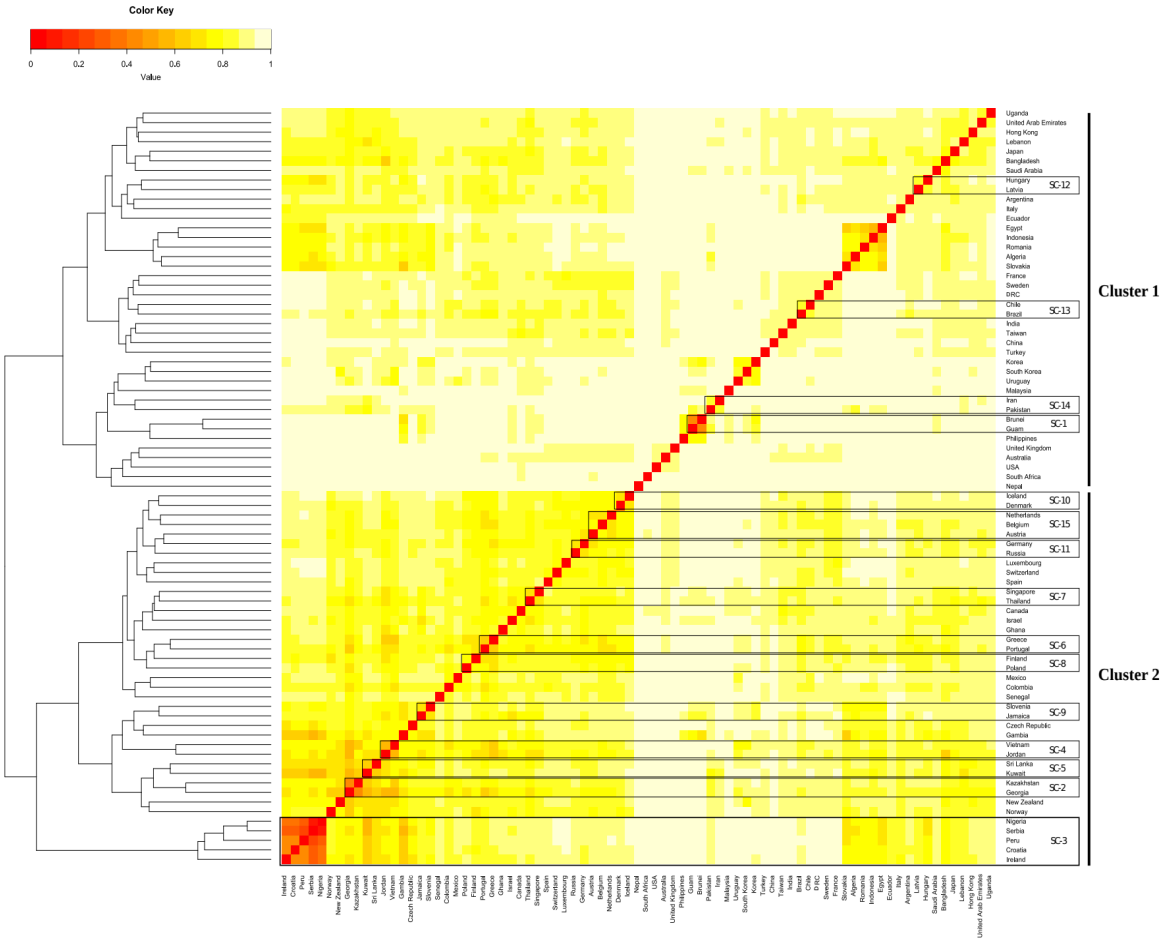

Figure 7: Correlation matrix heatmap showing the Jaccard distance between countries based on the SARS-CoV-2 mutation frequencies. The mutation frequencies have been standardized individually for each country. These frequencies were used to calculate the distance matrix between countries in a standard way using the Jaccard method. The resulting distance matrix has been plotted in the hierarchical grouping heat map to display the countries grouped with a significant distance. Countries with a distance <0.5 have been plotted in sub-cluster from SC-1 to SC-15.

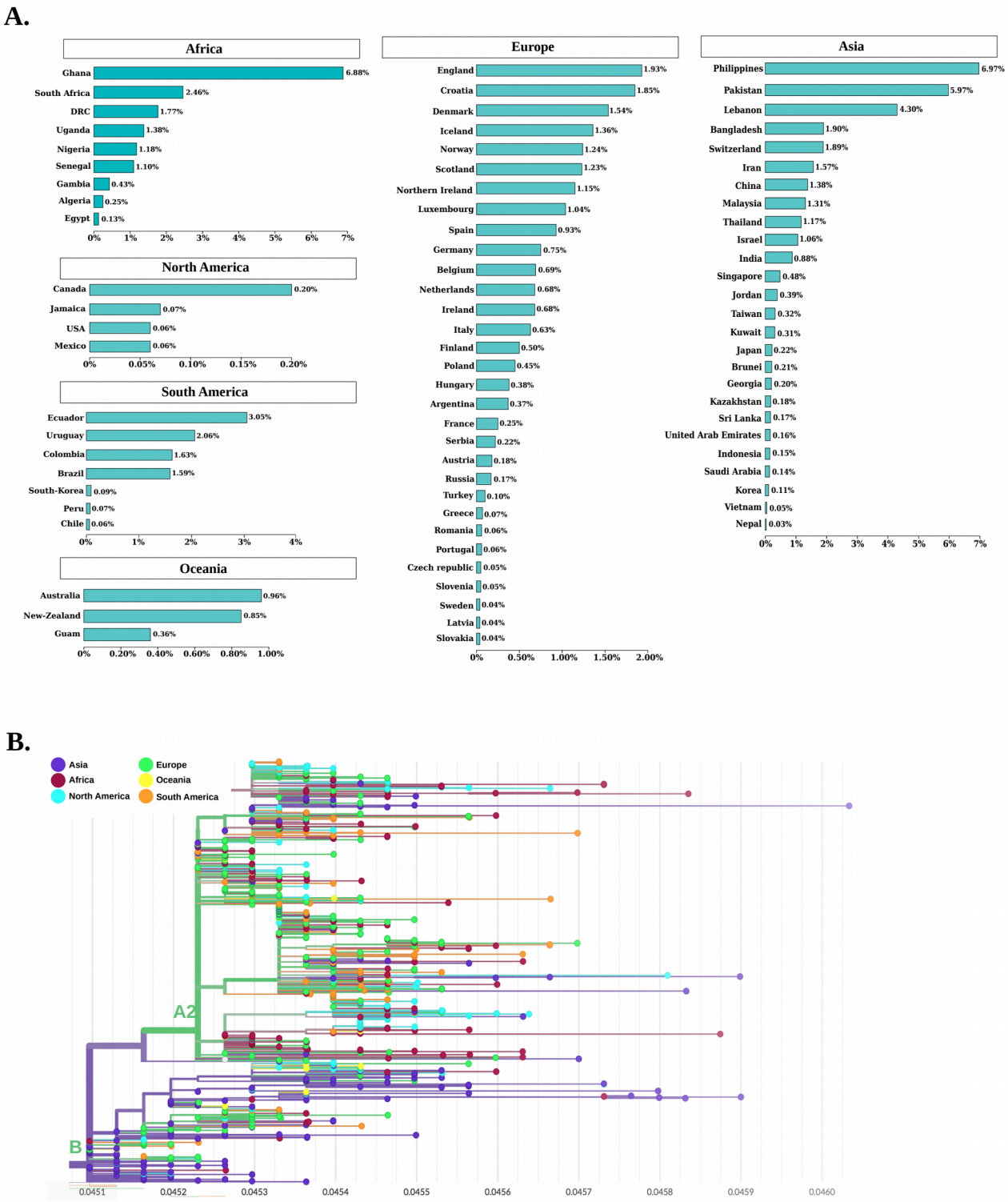

Figure 8: Divergence of the SARS-CoV-2 genomes in six different regions. (A) The bar graph showing the divergence of the SARS-CoV-2 genomes in each country compared to the reference Wuhan-Hu-1 genome. The length of the bar represents the percentage of divergence. **(B)** Phylogenetic Tree of SARS-CoV-2 whole genomes by genetic divergence ratio. Colored circles repre-

sent different genomes and branch length represent divergence. This tree is based on 1,025 genomes, filtred from 1,200 genomes with equal representation from each region.

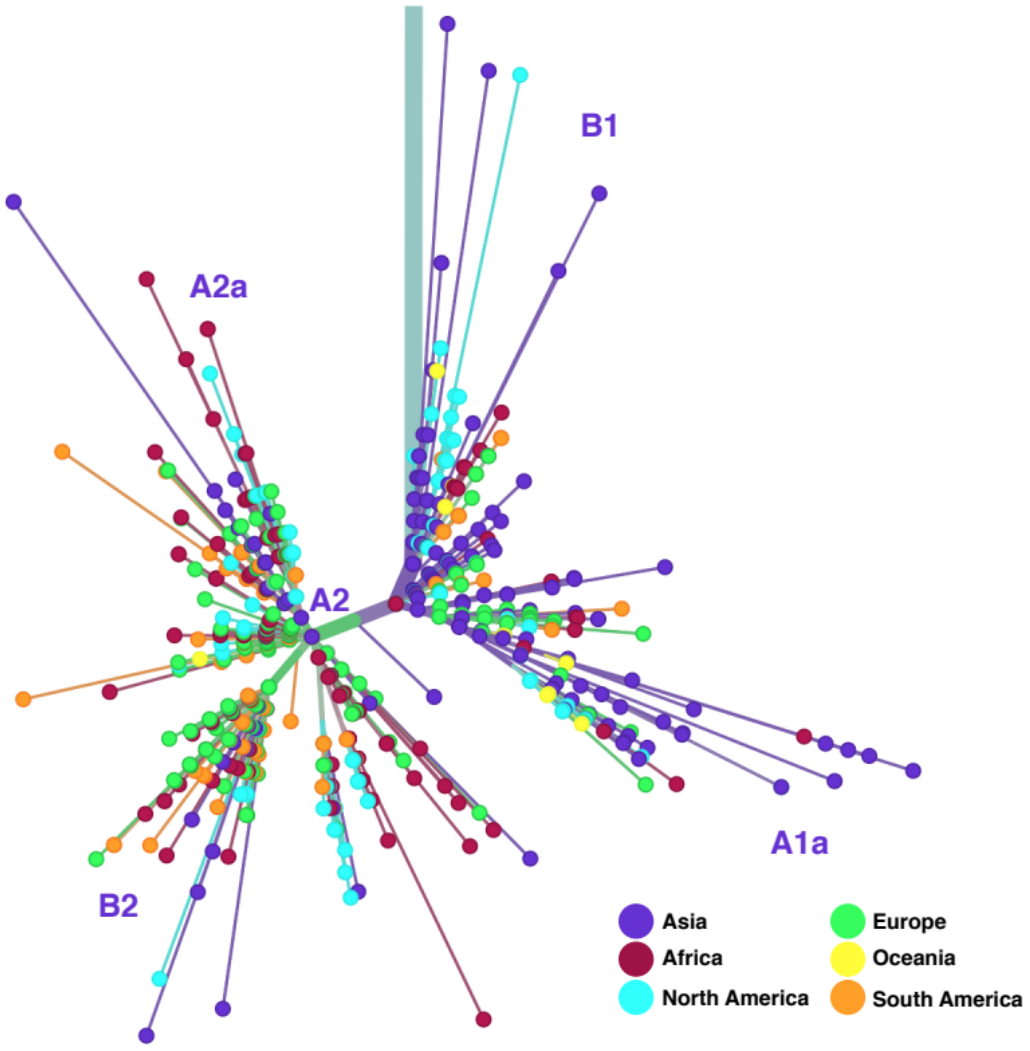

Figure 9: Phylogenetic tree created using whole genome alignment of 1026 SARS-CoV-2 outbreak isolates selected from GISAID bank. Colored circles represent different genomes grouped into six geographic areas.and branch length represent distance in time.

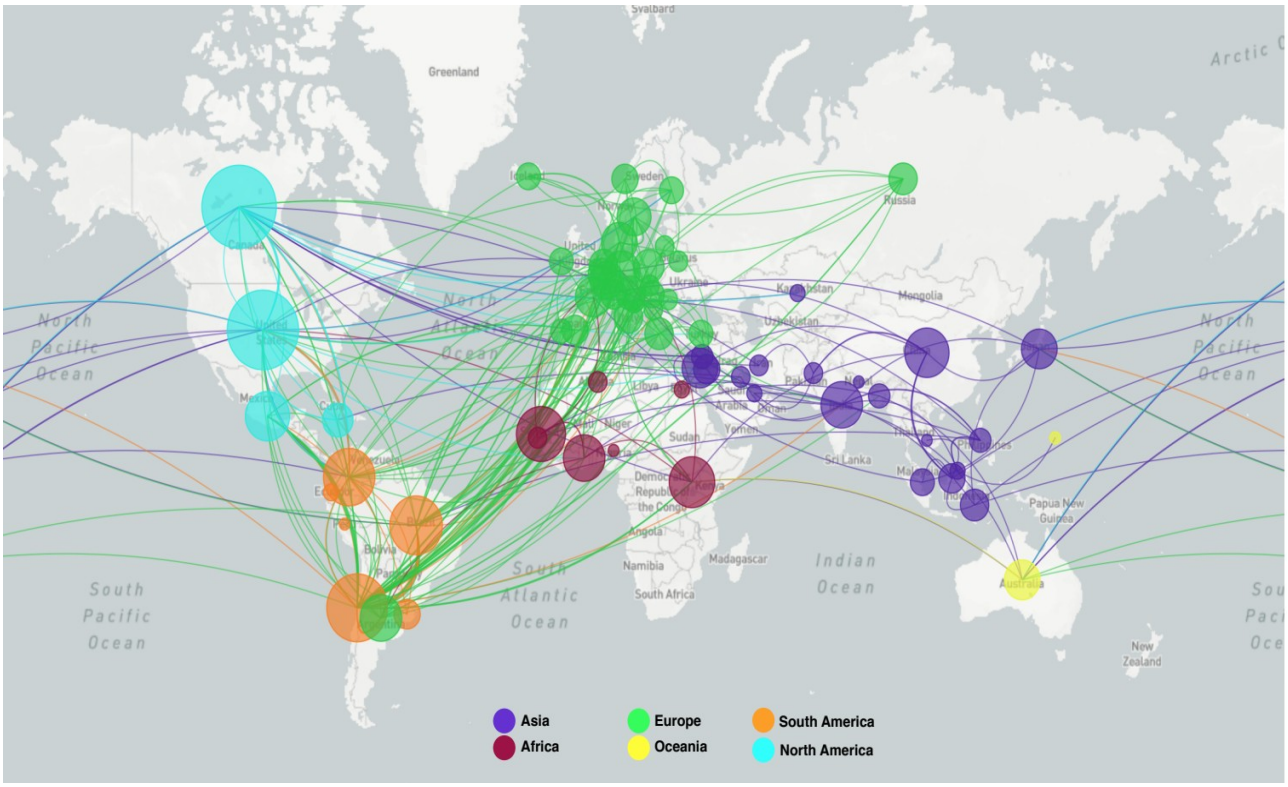

Figure S1: Phylodynamics analysis representing the spread and evolution of 1,025 SARS-Cov-2 across the continents. The continent of origin is color-coded.