Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Mutation profile of SARS-CoV-2 genome in a sample from the first year of the pandemic in Colombia

Jubby Marcela Gálvez [a], Henry Mauricio Chaparro-Solano [a,b,e], Ángela María Pinzón-Rondón [b], Ludwig L. Albornoz [c], Juan Mauricio Pardo-Oviedo [b,e], Fabio Andrés Zapata-Gómez [a], Andrés Felipe Patiño-Aldana [b], Andrea del Pila Hernández-Rodríguez [d], Mateo Díaz-Quiroz [d], Ángela María Ruiz-Sternberg [b,*]

[a] *Genuino Research Group, Gencell Pharma, Colombia*
[b] *Clinical Investigation Group, Universidad del Rosario, Colombia*
[c] *Fundación Valle del Lili, Colombia*
[d] *Universidad del Rosario, Colombia*
[e] *Hospital Universitario Mayor - Méderi, Colombia*

ARTICLE INFO

Keywords:
Covid-19
SARS-CoV-2
Whole genome sequencing
Genetic variation
High-throughput nucleotide sequencing
SARS-CoV-2 variants

ABSTRACT

The severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) is the etiopathogenic agent of COVID-19, a condition that has led to a formally recognized pandemic by March 2020 (World Health Organization –WHO). The SARS-CoV-2 genome is constituted of 29,903 base pairs, that code for four structural proteins (N, M, S, and E) and more than 20 non-structural proteins. Mutations in any of these regions, especially in those that encode for the structural proteins, have allowed the identification of diverse lineages around the world, some of them named as Variants of Concern (VOC) and Variants of Interest (VOI), according to the WHO and CDC. In this study, by using Next Generation Sequencing (NGS) technology, we sequenced the SARS-CoV-2 genome of 422 samples from Colombian residents, all of them collected between April 2020 and January 2021. We obtained genetic information from 386 samples, leading us to the identification of 14 new lineages circulating in Colombia, 13 of which were identified for the first time in South America. GH was the predominant GISAID clade in our sample. Most mutations were either missense (53.6%) or synonymous mutations (37.4%), and most genetic changes were located in the *ORF1ab* gene (63.9%), followed by the S gene (12.9%). In the latter, we identified mutations E484K, L18F, and D614G. Recent evidence suggests that these mutations concede important particularities to the virus, compromising host immunity, the diagnostic test performance, and the effectiveness of some vaccines. Some important lineages containing these mutations are the Alpha, Beta, and Gamma (WHO Label). Further genomic surveillance is important for the understanding of emerging genomic variants and their correlation with disease severity.

1. Introduction

In December 2019 a disease outbreak that caused respiratory failure in Wuhan, China, eventually developed into a global pandemic. The severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) was identified as the etiopathogenic agent of the respiratory disease called COVID-19, which after 2 months of its appearance, had already been established around the world (Sharma et al., 2020). On January 30, 2020, the World Health Organization (WHO) declared a public health emergency of international concern. By June 5, 2021, 172,712,949 confirmed cases and 3,716,918 deaths had been reported in more than 221 countries worldwide. In Colombia, by the same date 3,518,046 confirmed cases and 90,890 deaths had been reported, being the third most affected country in Latin America and 12th worldwide. By the moment of the submission of this manuscript, Colombia has faced three COVID-19 waves (July–August 2020; January 2021; and April 2021-present) (COVID-19 Map, 2021; Sharma et al., 2020).

The impact on health and the world economy associated with the emergence of this virus has generated interest in understanding its origin, spread, and evolution in a genomic path. Generation and analysis

[*] Corresponding author.
E-mail address: angela.ruiz@urosario.edu.co (Á.M. Ruiz-Sternberg).

https://doi.org/10.1016/j.meegid.2021.105192
Received 15 June 2021; Received in revised form 18 November 2021; Accepted 14 December 2021
Available online 18 December 2021

of virus genomic data has been a key component of this research and promises to provide critical insights into the emergence and spread of SARS-CoV-2 (COVID-19 Map, 2021).

The analysis of genomic data allowed to identify that the virus belongs to the beta-coronavirus 2b lineage. The SARS-CoV-2 genome consists of a positive polarity single-stranded RNA (+ ssRNA) of 29,903 base pairs. This RNA chain structurally resembles messenger RNA (mRNA) of eukaryotic cells, since it has a methylated cap (cap) at the 5′ end and a polyadenylated tail (poly-A) at the 3′ end. However, unlike eukaryotic mRNAs, this viral genome contains at least six open reading frames (ORFs) (Pastrian-Soto, 2020; Sharma et al., 2020; van Dorp et al., 2020).

Currently, the mechanisms underlying the variation in clinical susceptibility to COVID-19 and the disease presentation are yet to be fully characterized, raising doubts as to the importance of viral and host genetic mutations as likely factors that influence both disease severity and individual immune response (Pastrian-Soto, 2020; van Dorp et al., 2020). Viral and host genetic studies are essential to understand the pathophysiology of SARS-CoV-2 and inform the basis for the design of new vaccines and antiviral therapies. Additionally, such studies can help determine whether emerging viral strains are related to more severe clinical outcomes or whether individuals harboring certain alleles are susceptible to disease (Ovsyannikova et al., 2020; Pastrian-Soto, 2020; Sharma et al., 2020; van Dorp et al., 2020).

In December 2020, evidence began to emerge that a novel SARS-CoV-2 variant, Variant of Concern 202012/01 (lineage B.1.1.7, henceforth VOC 202012/01, and later named Alpha), was rapidly outcompeting preexisting variants in southeast England. Its incidence rose during a national lockdown in November 2020, in response to a previous and unrelated surge in COVID-19 cases, and continued to spread following the lockdown despite ongoing restrictions in many of the most affected areas. Concern over this variant led the UK government to enact stronger restrictions in these regions on December 20, 2020, and eventually to impose a third national lockdown on January 5, 2021. As of February 15, 2021, lineage Alpha comprised roughly 95% of new SARS-CoV-2 infections in England and has now been identified in at least 82 countries (Davies et al., 2021). There have been other strains identified in different countries, including South Africa, namely variant 501Y.V2, later termed Beta, likely emerging from the first wave of COVID-19 back in May 2020. The Beta strain has been found by now in several countries and is considered to be more transmissible (Ramanathan et al., 2021). In addition, the P.1 (20J/501Y.V3, labeled as Gamma) variant is a strain that was found in Manaus, Brazil, which generated international apprehension due to its high transmissibility (de Souza et al., 2021; Tang et al., 2021).

The global and uncontrolled scale of this pandemic created a scenario where genomic epidemiology was put into practice *en masse*, from the rapid sequencing of SARS-CoV-2 to the identification of new lineages by means of active surveillance throughout the world. Prior to the COVID-19 pandemic, the availability of genomic data on circulating pathogens in several Latin American and Caribbean countries was scarce or nil (Álvarez-Díaz et al., 2020). With the arrival of SARS-CoV-2, this scenario slightly changed, although available information remains scarce and, in countries such as Colombia, Brazil, Argentina, and Chile, genomic information of SARS-CoV-2 is obtained mainly by research groups in genomic epidemiology rather than by public health surveillance policy or programs. This indicates the need to reinforce public health policies aimed at implementing genomic epidemiology as a tool to strengthen surveillance and early warning systems against threats to public health in the region (Álvarez-Díaz et al., 2020; Islam et al., 2020).

In this study we aimed to describe the genomic sequence obtained from 422 Colombian resident patients diagnosed in Bogotá and Cali, between April 2020 and January of 2021, carrying out a molecular epidemiological characterization, and describing the frequency of genetic mutations.

2. Materials and methods

2.1. Sampling and nucleic acid inactivation and extraction

Nasopharyngeal swab / aspirate samples were obtained from patients confirmed to have COVID-19 in two participating tertiary care hospitals and a molecular diagnosis laboratory located in two main cities in Colombia.

Prospective samples were evaluated by RT-PCR to detect SARS-CoV-2 presence; alternatively, SARS-CoV-2 antigen positive tests were performed followed by a nasopharyngeal swab/ aspirate for confirmation by RT-PCR, within 72 h. RT-PCR negative samples were discharged and those patients were not included in the study. Retrospective specimen collection nasopharyngeal RT-PCR positive samples, either RNA eluate or primary swab/ aspirate samples, were also obtained.

Automated RNA extraction was performed using either the ExiPrep™ 96 Viral DNA/RNA Kit, on ExiPrep™ 96 Lite instrument (Bioneer Corp., Daejeon, Republic of Korea), the MGIEasy Nucleic Acid Extraction Kit on the MGISP-960 (MGI Tech Co. Ltd., Shenzhen, PRC), the NucliSENS® Nucleic Acid Extraction Reagents on NucliSENS® easyMAG® (bioMérieux SA, Marcy l'Etoile, France), or MagNA Pure® Compact Nucleic Acid Isolation Kit I on MagNA Pure® Compact (Roche Diagnostics GmbH, Mannheim, Germany).

For RNA extraction, a 200 μL aliquot of viral transport media containing the primary swab sample, or a 1000 μL aliquot in sterile saline isotonic solution in case of primary aspirate sample, was inactivated by exposure to 56 °C for 30 min, in a class II, type B2 biosafety cabinet.

Inactivated samples were transferred to any of the above described automated RNA extraction instruments. Briefly, RNA automated extraction consists of cell lysis, followed by bead binding to the magnetic rods, RNA binding to the beads, washing, and finally, RNA elution, obtaining approximately 100 μL of RNA (Kessler et al., 2001; Petrich et al., 2006).

2.2. SARS-CoV-2 genome sequencing

Whole-genome amplification of the SARS-CoV-2 was performed using the CleanPlex® for MGI SARS-CoV-2 Panel (Paragon Genomics Inc., Hayward, CA, USA). Briefly, cDNA was generated from previously extracted RNA followed by purification from purified RNA samples. Then, a multiplex PCR reaction using target-specific primers to amplify the entire SARS-CoV-2 genome was performed, using a 2-pool design, followed by a second PCR reaction to amplify and add sample-level indexes to the generated libraries. This second PCR introduced an indexed PCR primer for the specific sequencer. Finally, the library was purified using CleanMag® Magnetic Beads (Paragon Genomics Inc.) (Li et al., 2020). Libraries were considered for sequencing when a fragment size between 170 and 300 bp was obtained and the final concentration was above 2.0 ng/μl, measured by a fluorometric method Quantus® (Promega Corp., Fitchburg, WI, USA).

In order to be sequenced, libraries were converted to DNBs (DNA nanoballs). DNBs are generated by Rolling Circle Replication (RCR/RCA), a process that includes primer annealing, extension, displacement of DNA strand and continued extension, always from the original template, which allows avoiding error accumulation and low amplification bias. Then, circularized libraries were sequenced using either a DNBSeq-G50RS or a DNBSeq-G400RS sequencer (MGI Tech Co. Ltd., Shenzhen, PRC). These instruments use DNBSEQ™ technology consisting of DNBs pumping and loaded onto an array chip by the fluidics system. Sequencing primer is then added and hybridized to the adaptor region of the DNB. The sequencing reaction starts by pumping sequencing reagents containing fluorescently labeled dNTP probes and DNA polymerase. Images are taken after the fluorescently labeled probes on the DNB are excited with lasers. The images are then converted into a digital signal. This information is then used to determine the DNA sequence of the sample (Drmanac et al., 2010).

Reference-based genome assembly and sequence analysis were performed using SOPHiA DDM software for SARS-CoV-2 (SOPHiA Genetics Inc., Boston, MA, USA). Paired-end reads were aligned to the reference SARS-CoV-2 genome (NC_045512.2) using bwa mem (version 0.7.12-r1039) (2) and the following command: bwa mem -M reference.fasta read1.fastq read2.fastq. Aligned reads underwent adaptor trimming by removing all read 3′ bases extending beyond the 5′ starting position of its paired-end mate. Read alignment was inspected to identify and remove amplicons where at least one primer aligned to an unexpected genomic position (mispriming events). Mispriming events include any softclipped sequence longer than 10 bp matching another primer sequence better than the sequence at the current position. The remaining softclipped read fragments longer than 10 bp were aligned to the genome in the closest vicinity (up to a distance of 400 bp) to the read mapped position to detect putative insertions or deletions. Smith–Waterman alignment score was calculated with the following penalty scores: base match: +1, base mismatch: −1, gap open: −2, gap extend: −1. If the fragment aligned to the alternative location with a score of ≥10, the softclipped part of the read was flagged as part of an indel. Finally, sequences overlapping primer positions were trimmed, if found at most 5 bp from the beginning or end of the read. Read fragments shorter than 21 bp were discarded from the final alignment

A pileup file was generated from the final post-processed alignment and only base pairs with a Phred score higher than 20 were considered. To prevent miscalls due to background noise, at each genomic position a one-sided Fisher's test was performed on a 2 × 2 matrix containing (i) the average number of reads with reference alleles calculated at each covered genomic position, (ii) the number of reads with reference alleles at current position, (iii) the average number of reads with alternative alleles calculated at each genomic position, (iv) the number of reads with alternative alleles at current position. If the test *p*-value was higher than 10^{-5} the mutation was not reported at the position tested.

Miscalls due to strand-biased detection were removed in a similar way. Fisher's test was used to analyze the proportion of reads supporting the alternative or reference allele derived from either the plus and minus strands. If the test p-value was lower than 10^{-7} the mutation was filtered out as strand-biased.

The minimal total read coverage considered for mutation reporting was 10 reads while the minimal read coverage supporting a mutation was 3 reads. Nucleotide substitutions at first position of the amplicon were not reported unless they were also detected in adjacent, overlapping amplicons (Kubik et al., 2021).

2.3. Phylogenetic analysis

All SARS-CoV-2 genomes were downloaded from SOPHiA™ DDM® bioinformatics software (SOPHiA Genetics Inc.). We aligned these sequences along the reference genome, NC_045512, using MAFFT v7 software (Katoh et al., 2002). Then, the mutation was estimated using jMoldelTest v2.1.10 model (Posada, 2008). Then, a maximum likelihood tree was constructed using IQ-TREE 2 (Minh et al., 2020) and the GTR + Γ model, using 1000 bootstrap replicates. Finally, each sample had a lineage assigned using PANGOLIN's web server (Rambaut et al., 2020) and the online version of CoVsurver (CoVsurver - CoronaVirus Surveillance Server, *2021*) was used for GISAID clade assignation of the samples.

Samples analyzed were obtained from participants that accepted and consented to participate in the study. The research project was approved by the IRB of Universidad del Rosario and the participant hospitals in Colombia, and it accomplished all the international and national bioethical principles and regulations for research in human subjects.

2.4. Statistical analysis

To synthesize demographic characteristics of the sample we present descriptive statistics. We used absolute and relative frequencies for categorical variables. For quantitative data, we present median and interquartile range. We used R v 4.0.2 software for statistical analysis.

3. Results

To inform our findings we used the terminology proposed in outbreak.info, an open-data source supported by the National Institutes of Health (NIH) and Center for Data to Health (CD2H). Briefly, a mutation is an error in viral genome established during replication. A variant is an aggregation of several mutations in a single genome. A lineage refers to "the descendants of a branch of a phylogenetic tree". Finally, a strain is a viral genome with differential phenotypic characteristics and implicates a specific immune response by the hostage (Outbreak.Info, *2021*).

We recruited 422 Colombian resident patients with RT-PCR confirmed SARS-CoV-2 infection whose symptom onset ranged from April 2020 to January 2021. Sociodemographic characteristics from the 386 patients whose samples were included in the genetic analysis showed that 53.5% were male and the median age was 45.7 years (31.6–62.5). The geographic origin of the sequences was 73.2% from Bogotá and 26.8% from Cali, Colombia. Around half of the patients (40.9%) received in-hospital attention, the others were ambulatory.

3.1. Sequenced SARS-CoV-2 specimens

From the 422 samples, we were able to obtain genetic data from 386 cases. Variation analysis of these 386 sequences revealed 1716 mutations detected across the length of the SARS-CoV-2 genome. Most mutations were either missense (53.6%) or synonymous (37.4%). Loss-of-function (LoF) mutations were rare; only 26 (1.51%) nonsense and 47 (2.73%) frameshift mutations were identified. Most of the genetic changes found (63.9%) were located in the *ORF1ab* gene, followed by the *S* gene (12.9%), whereas the *ORF10* had the lowest number of changes (0.64%). Nevertheless, when corrected by gene length, the mutation rate per kb was higher in ORF8 (202.07 mutations per kb), followed by N (115.64 mutations per kb), while the lowest rate was found in the ORF1ab gene (51.48 mutations per kb) (Table 1).

Additionally, we found four recurrent mutations which have been previously described as hotspots since they have a frequency higher than 0.10 in the population. We identified Q57H on *ORF3a* in 184 out of 386 (47.6%) samples; we also found and R203K, R203R, G204R on *N* gene, in 74 (19.1%), 72 (18.7%), and 72 (18.7%) out of 386 cases, respectively.

Active domains of each SARS-CoV-2 protein with well-established function and their respective amino acid lengths were characterized using the bioinformatics tool UniProt. All missense and LoF mutations that compromise some of these domains were deeply analyzed hypothesizing that mutations in these regions will impact the biological processes of the virus. Important for the characterization and identification of new lineages with potential for increased transmissibility and/or lethality, also known as variants of concern (VOC) or variants of interest (VOI), we identified on the spike glycoprotein 223 mutations, from which 135 were missense (60.53%), 82 synonymous (36.77%), 5 frameshift (2.24%) and 1 (0.44%) intergenic. Also, we identified 10 mutations within the aminoacidic residues 319 and 541, which constitute the receptor-binding domain (RBD).

The mutations p.Lys1191Asp (K1191D), p.Glu484Lys (E484K), p.Leu18Phe (L18F), p.Pro26Ser (P26S), p.His655Tyr (H655Y), p.Asp614Gly (D614G), that along with other nucleotide substitutions constitute some of the VOC and VOI, were also identified, but not necessarily defining a complete lineage of interest. The spike glycoprotein mutation D614G was found in the totality of the sequences obtained (Table 2).

3.2. Phylogenetic analysis

The maximum likelihood tree inferred for the 386 samples revealed a

Table 1
SARS-CoV-2 variants description. Identification of the number and type of variants, and the corrected rate of variants per kilobase (Kb) per region or gene.

Coding consequence / Gene	3'UTR	5'UTR	E	M	N	ORF10	ORF1ab	ORF3a	ORF6	ORF7a	ORF7b	ORF8	S	Total
3'UTR	32													**32**
5'UTR		22												**22**
Frameshift				2	2		29	1		3		5	5	**47**
Inframe					1		4					1		**6**
Intergenic			3	2	1	6						1	1	**14**
Missense			10	14	66	3	580	59	9	17	9	18	135	**920**
Nonsense				1	3		16	2	1			3		**26**
Synonymus			2	20	32	2	467	11	4	10	2	10	82	**642**
Start Loss				1							2	1		**4**
Stop Loss										2	1			**3**
Total number of variants	**32**	**22**	**15**	**40**	**105**	**11**	**1096**	**73**	**14**	**32**	**14**	**39**	**223**	**1716**
Variants rate per Kb	**NA**	**NA**	**65,79**	**59,79**	**115,64**	**94,02**	**51,45**	**88,16**	**75,27**	**87,43**	**106,06**	**202,07**	**58,35**	

Table 2
Identified Spike Protein Substitutions present on a Variant of Concern (VOC) or Variant of Interest (VOI).

Substitution	Substitution present on a VOC or VOI		n
	VOC or VOI WHO Label	VOC or VOI First Detected	
p.Asp614Gly (D614G)	All	Multiple countries	386
p.Leu18Phe (L18F)	Gamma	Brazil	16
p.Lys1191Asp (K1191D)	Alpha	United Kingdom	8
p.Glu484Lys (E484K)	Alpha, Beta, Eta, Gamma, Iota	United Kingdom, South Africa, Brazil, USA	1
p.Pro26Ser (P26S)	Gamma	Brazil	1
p.His655Tyr (H655Y)	Gamma	Brazil	1

topology of two major genetic groups (Fig. 1), with one sample not included in these groups because it is closely related to the original Wuhan strain. The remaining samples were clustered into two different branches: The first one comprises 186 samples, the second one clustering 194 samples. The remaining five samples failed the chi-square test performed by Iqtree and therefore, they aren't present on the tree. However, this particular topology should be interpreted carefully, given the low bootstrap values (not shown) due to aligning sequences with high similarity percentages.

We identified 30 different PANGO lineages. The most frequent lineage was B.1 followed by B.1.111 and B.1.1.348 (Table 3). The distribution of PANGO lineages by sampling date is shown in Fig. 2.

The GH GISAID clade was the most frequent ($n = 167$; 43.3%), followed by the G clade ($n = 122$; 31.6%) and GR clade ($n = 66$; 17.1%); 31 (8.0%) samples were classified as other.

4. Discussion

Hereby we inform 30 PANGO lineages found in 386 SARS-CoV-2 sequences, including 14 newly reported PANGO lineages circulating in Colombia, 13 of which are also newly reported in South America as reviewed by June 5, 2021 (https://github.com/cov-lineages/pangolin).

According to GISAID data, as of January 2021, GR was the most common clade worldwide, followed by GV and GH. On our sample, the GH clade was predominant, followed by the G and GR clades, similar to what has been described for North America. Interestingly, this distribution differs from what has been described for South America. This is probably due to the larger contribution of sequences from Brazil, where the P.1 VOC, part of the GR clade, has become predominant (Hamed et al., 2021).

The frequency of the most predominant PANGO lineages identified here is consistent with the data of the 1085 sequences reported by the National Institute of Health in Colombia by May 21st of 2021 (Coronavirus Colombia, 2021). B.1 and B.1.111 lineages represent 55.4%, confirming them as the first and second most frequent lineages identified in Colombia. By February 2021 the B.1.111 lineage in GISAID had a significant representation in samples from Colombia. Recently, new evidence from Colombian National Institute of Health, has shown that the mutation E484K is becoming associated with this lineage. In our study we found the E484K mutation in a single sample taken in December 2020 from a case with the B.1 lineage, parental lineage of B.1.111 (Laiton-Donato et al., 2021).

The finding of E484K is relevant since this mutation has been described in several VOC and VOI, which WHO labeled as Alpha, Beta, Eta, Gamma, and Iota. According to the literature, single amino acid mutations in the viral Spike glycoprotein, could modulate ACE2 binding, alter B-cell epitopes to promote immune escape or render monoclonal antibodies ineffective (Pastrian-Soto, 2020; van Dorp et al., 2020). For example, the mutation in the 484 amino acid of the S glycoprotein, being part of the binding domain to the ACE2 human receptor, could have important implications related to the severity of the disease (Ramanathan et al., 2021). Recent studies have found *in vitro* evidence that the presence of this mutation is associated with a reduction in the activity of monoclonal antibodies and convalescent serum authorized for COVID-19 emergency use (Huang and Wang, 2021; Jangra et al., 2021). Interestingly we found other 10 mutations within the aminoacidic residues 319 and 541, which constitute the RBD that should be followed within the genomic surveillance program stablished by the country.

According to several studies, the L18F mutation, also located in the S glycoprotein and present in 16 of our samples, apparently gives a replicative advantage to the Alpha, Beta, and Gamma lineages. Preliminary evidence suggests that this mutation is associated with a decreased immunological response because it compromises susceptibility of the RBD to neutralizing antibodies (Cele et al., 2021; Grabowski et al., 2021).

All of our samples contain the Spike glycoprotein mutation D614G. This nucleotide substitution (Laiton-Donato et al., 2021) rapidly turned into a worldwide prevalent mutation. Nevertheless, despite its high

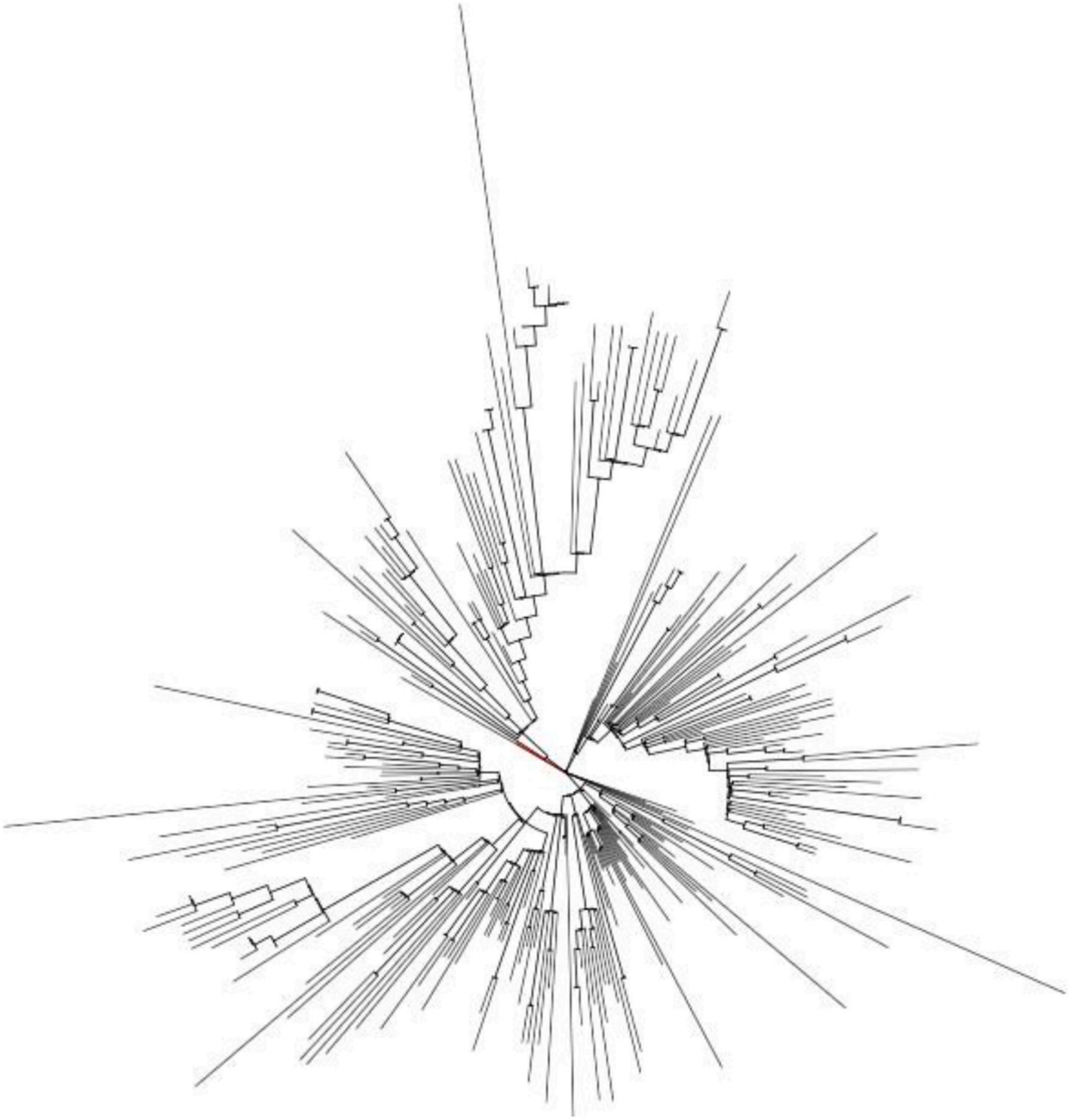

Fig. 1. Phylogenetic tree inferred with the approximate maximum likelihood method showing the relationship between 386 Colombian SARS-CoV-2 genome sequences and revealing two main genetic clusters.

speed of transmission, recent evidence demonstrated that this amino acidic change on its own, is not associated with a particular phenotype (Grubaugh et al., 2020; Isabel et al., 2020).

Mutations are an important mechanism for evolution, fitness, and survival of viruses, and SARS-CoV-2 is a virus still adapting to humans. A mutation rate of 9.9 × 10–4 to 2.2 × 10–3 nucleotides/genome per year for SARS-CoV-2 has been estimated, according to L. Vázquez-Sirvent, B. Martínez-González, M.E. Soria, and C. Perales, unpublished results, which is similar to other active RNA viruses (Domingo et al., 2021). In this study we described the genetic changes in our SARS-CoV-2 samples. We found that missense and synonymous mutations were the most prevalent genetic variations, similar to previous reports (Chan et al., 2020a; Wang et al., 2020). As expected, LoF, loss of start, and loss of stop mutations are the least frequent changes. Interestingly, we found several LoF mutations (both nonsense and frameshift) as well as a loss of initiation codon mutation that compromises the functional domains of the ORF8 protein, even when corrected for mutations per kb (Supplementary Material Table 1). It has been hypothesized that *ORF8* may play a role in modulating host immune response by blocking host IL-17 cytokine by its interaction with host IL17RA (Chan et al., 2020b). Therefore, it would be interesting to study the host immune response when LoF mutations in the *ORF8* gene are present.

In relation to genetic variation among SARS-CoV-2 genes, similar to previous reports, we found that *ORF1ab* and *S* genes have the highest number of mutations (Chan et al., 2020a; Wen et al., 2020). *ORF10*, previously reported as highly conserved, presented a low number of mutations (the lowest in our sample, with a total of 11 genetic variations) (Tsai et al., 2020). When corrected for mutations per kb, *ORF8* and *N* genes had the highest rate of mutations (202,1 and 115,6 mutations per kb, respectively), while *S* and *ORF1ab* genes had the lowest (58,3 and 51,5 mutations per kb, respectively). *ORF8* has been described as one of the most hypervariable and rapidly evolving genes of SARS-CoV-2, explaining its high mutation rate (Alkhansa et al., 2021; Zinzula, 2021).

Several hotspot mutations have been described in different regions of the SARS-CoV-2 genome. Of those, we reported here the following four: Q57H on *ORF3a*, and R203K, R203R, G204R on *N* gene. Three of them, R203K, R203R, and G204R, co-existed in 72 samples, similar to previous

Table 3
PANGO lineages detected and their frequencies.

PANGO Lineage	n	Percentage (%)
B.1	150	38,9
B.1.111	64	16,6
B.1.1.348	39	10,1
B.1.420	28	7,3
B.1.153	20	5,2
B.1.1	16	4,1
B	11	2,8
A	10	2,6
B.1.383	9	2,3
B.1.523	6	1,6
B.1.293	5	1,3
B.1.1.388	4	1,0
B.1.1.28	3	0,8
B.1.416	2	0,5
B.1.177.86	2	0,5
B.1.411	2	0,5
B.1.1.434	2	0,5
B.1.389	1	0,3
B.1.165	1	0,3
B.1.36.10	1	0,3
B.1.319	1	0,3
B.59	1	0,3
B.1.485	1	0,3
B.1.1.1	1	0,3
B.1.1.100	1	0,3
B.1.1.413	1	0,3
B.1.456	1	0,3
B.1.1.409	1	0,3
B.1.1.37	1	0,3
B.1.505	1	0,3

reports around the world, suggesting a linkage disequilibrium, which makes functional inferences difficult (Laamarti et al., 2020; Ogawa et al., 2020).

The frequency of Q57H hotspot mutation on *ORF3a* has been rated among the top three in the United States (top six around the world). Additionally, R203K is within the top five in the United States (top three in the world); G204R ranks in the top six in the United States (top four in the world)(Wang et al., 2021).

The ORF3a protein is the largest accessory protein in the SARS-CoV-2 genome: it is expressed in intracellular and plasma membranes, inducing apoptosis and inflammatory response in the infected cells. Q57H mutation makes the protein unstable, altering apoptosis and, as a result, increasing the viral load in the host cell (Ren et al., 2020; Wang et al., 2021).

N gene codifies for the nucleocapsid protein, which plays a key role in RNA packaging and the release of viral particles (Zeng et al., 2020). The amino acid positions 203 and 204 are highly conserved among other coronaviruses. As previously described, R203K mutation may not affect protein function. However, mutation G204R seems to alter protein folding and eventually lead to a functional impact (R. Wang et al., 2021).

Even though South America has been severely affected by the COVID-19 pandemic, the amount of genetic sequences available from the region is still scarce. The present sample represents a contribution to the genomic characterization of SARS-CoV-2 in Colombia, the 12th affected country worldwide, in terms of an absolute number of cases with 3,593,016 cases and the 20th worldwide (5th in South America) in relation to mortality rate per 100,000 by June 7, 2021. Additionally, this study comprises the first ten months of the COVID-19 pandemic in Colombia and involves the first two waves, which allows a modest approach to the virus genetic characterization through the study interval.

The study has some limitations: it only includes samples from two main cities in Colombia, including the capital city, leading to an underrepresentation of other regions in the country. Also, due to the descriptive nature of the study, it does not evaluate the functional impact of the mutations described.

5. Conclusions

In conclusion, we identified 14 new lineages in Colombia, 13 of which are newly reported in South America. Additionally, similar to previous reports, we found that missense and synonymous mutations are the most frequent, and *ORF1ab* and *S* genes have the highest number of mutations. Also, we found important mutations such as E484K, L18F, and D614G, which impact on viral transmissibility, host immunity, and disease severity.

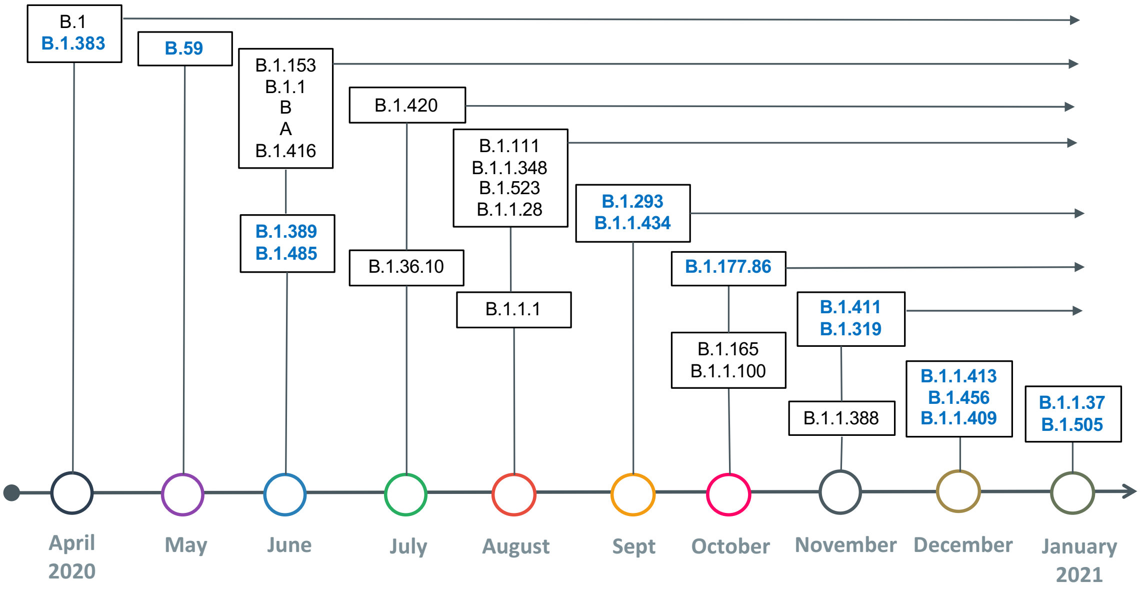

Fig. 2. Distribution of identified PANGO lineages collected from April 2020 to January 2021 during the first year of the pandemic in Colombia. PANGO lineages in blue represent novel variants in Colombia, while black ones were previously identified. Boxes with arrows represent the continuity of the PANGO lineages along time after they were first identified, while the others were only found in a single month. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Funding

This work was supported by governmental funding from the Colombian Ministry of Science, Technology and Innovation -Ministerio de Ciencia, Tecnología e Innovación- (Grant number 366-2020).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Lina Marcela Méndez Castillo, Andrés Felipe Torres Gómez and Danyela Faisury Valero Rubio for supporting the laboratory procedures, and Diego Andrés Otero Rodríguez for his guidance in the phylogenetic analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2021.105192.

References

Alkhansa, A., Lakkis, G., El Zein, L., 2021. Mutational analysis of SARS-CoV-2 ORF8 during six months of COVID-19 pandemic. Gene Rep. 23, 101024 https://doi.org/10.1016/j.genrep.2021.101024.

Álvarez-Díaz, D.A., Laiton-Donato, K., Franco-Muñoz, C., Mercado-Reyes, M., 2020. SARS-CoV-2 sequencing: the technological initiative to strengthen early warning systems for public health emergencies in Latin America and the Caribbean. Biomed.: Revista Inst. Nacional Salud 40 (Supl. 2), 188–197. https://doi.org/10.7705/biomedica.5841.

Cele, S., Gazy, I., Jackson, L., Hwa, S.-H., Tegally, H., Lustig, G., Giandhari, J., Pillay, S., Wilkinson, E., Naidoo, Y., Karim, F., Ganga, Y., Khan, K., Balazs, A.B., Gosnell, B.I., Hanekom, W., Moosa, M.-Y.S., NGS-SA, Team, C.-K, Sigal, A, 2021. Escape of SARS-CoV-2 501Y.V2 variants from neutralization by convalescent plasma. MedRxiv. https://doi.org/10.1101/2021.01.26.21250224, 2021.01.26.21250224.

Chan, J.F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K.K.-W., Yuan, S., Yuen, K.-Y., 2020a. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microb. Infect. 9 (1), 221–236. https://doi.org/10.1080/22221751.2020.1719902.

Chan, J.F.-W., Yuan, S., Kok, K.-H., To, K.K.-W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C. C.-Y., Poon, R.W.-S., Tsoi, H.-W., Lo, S.K.-F., Chan, K.-H., Poon, V.K.-M., Chan, W.-M., Ip, J.D., Cai, J.-P., Cheng, V.C.-C., Chen, H., Yuen, K.-Y., 2020b. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet (London, England) 395 (10223), 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9.

Coronavirus Colombia, 2021. Recuperado el 11 de junio de 2021, de. https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx.

COVID-19 Map, 2021. Johns Hopkins Coronavirus Resource Center. Recuperado el 11. de junio de 2021, de. https://coronavirus.jhu.edu/map.html.

CoVsurver—CoronaVirus Surveillance Server, 2021. Recuperado el 11 de junio de 2021, de. https://mendel3.bii.a-star.edu.sg/METHODS/corona/beta/indexAnno2.html.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A.B., Russell, T.W., Tully, D.C., Washburne, A.D., Wenseleers, T., Gimma, A., Waites, W., Wong, K.L.M., van Zandvoort, K., Silverman, J.D., CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Diaz-Ordaz, K., Edmunds, W.J., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science (New York, N.Y.) *372* (6538). https://doi.org/10.1126/science.abg3055.

de Souza, W.M., Amorim, M.R., Sesti-Costa, R., Coimbra, L.D., de Toledo-Teixeira, D.A., Parise, P.L., Barbosa, P.P., Bispo-dos-Santos, K., Mofatto, L.S., Simeoni, C.L., Brunetti, N.S., Claro, I.M., Duarte, A.S.S., Coletti, T.M., Zangirolami, A.B., Costa-Lima, C., Gomes, A.B.S.P., Buscaratti, L.I., Sales, F.C., Proenca-Modena, J.L., 2021. *Levels of SARS-CoV-2 Lineage P.1 Neutralization by Antibodies Elicited after Natural Infection and Vaccination* (SSRN Scholarly Paper ID 3793486). In: Social Science Research Network. https://doi.org/10.2139/ssrn.3793486.

Domingo, E., García-Crespo, C., Lobo-Vega, R., Perales, C., 2021. Mutation rates, mutation frequencies, and proofreading-repair activities in RNA virus genetics. Viruses 13 (9), 1882. https://doi.org/10.3390/v13091882.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K.P., Baccash, J., Borcherding, A.P., Brownley, A., Cedeno, R., Chen, L., Reid, C.A., 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science (New York, N.Y.) 327 (5961), 78–81. https://doi.org/10.1126/science.1181498.

Grabowski, F., Kochańczyk, M., Lipniacki, T., 2021. L18F substrain of SARS-CoV-2 VOC-202012/01 is rapidly spreading in England. MedRxiv. https://doi.org/10.1101/2021.02.07.21251262, 2021.02.07.21251262.

Grubaugh, N.D., Hanage, W.P., Rasmussen, A.L., 2020. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell 182 (4), 794–795. https://doi.org/10.1016/j.cell.2020.06.040.

Hamed, S.M., Elkhatib, W.F., Khairalla, A.S., Noreddin, A.M., 2021. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. Sci. Rep. 11 (1), 8435. https://doi.org/10.1038/s41598-021-87713-x.

Huang, S.-W., Wang, S.-F., 2021. SARS-CoV-2 entry related viral and host genetic variations: implications on COVID-19 severity, immune escape, and infectivity. Int. J. Mol. Sci. 22 (6) https://doi.org/10.3390/ijms22063060.

Isabel, S., Graña-Miraglia, L., Gutierrez, J.M., Bundalovic-Torma, C., Groves, H.E., Isabel, M.R., Eshaghi, A., Patel, S.N., Gubbay, J.B., Poutanen, T., Guttman, D.S., Poutanen, S.M., 2020. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. Sci. Rep. 10 (1), 14031. https://doi.org/10.1038/s41598-020-70827-z.

Islam, M.R., Hoque, M.N., Rahman, M.S., Alam, A.S.M.R.U., Akther, M., Puspo, J.A., Akter, S., Sultana, M., Crandall, K.A., Hossain, M.A., 2020. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci. Rep. 10 (1), 14004. https://doi.org/10.1038/s41598-020-70812-6.

Jangra, S., Ye, C., Rathnasinghe, R., Stadlbauer, D., Personalized Virology Initiative study group, Krammer, F., Simon, V., Martinez-Sobrido, L., García-Sastre, A., Schotsaert, M., 2021. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. Lancet. Microbe. https://doi.org/10.1016/S2666-5247(21)00068-9.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30 (14), 3059–3066. https://doi.org/10.1093/nar/gkf436.

Kessler, H.H., Mühlbauer, G., Stelzl, E., Daghofer, E., Santner, B.I., Marth, E., 2001. Fully automated nucleic acid extraction: MagNA pure LC. Clin. Chem. 47 (6), 1124–1126.

Kubik, S., Marques, A.C., Xing, X., Silvery, J., Bertelli, C., De Maio, F., Pournaras, S., Burr, T., Duffourd, Y., Siemens, H., Alloui, C., Song, L., Wenger, Y., Saitta, A., Macheret, M., Smith, E.W., Menu, P., Brayer, M., Steinmetz, L.M., Xu, Z., 2021. Recommendations for accurate genotyping of SARS-CoV-2 using amplicon-based sequencing of clinical samples. Clin. Microbiol. Infect. https://doi.org/10.1016/j.cmi.2021.03.029.

Laamarti, M., Alouane, T., Kartti, S., Chemao-Elfihri, M.W., Hakmi, M., Essabbar, A., Laamarti, M., Hlali, H., Bendani, H., Boumajdi, N., Benhrif, O., Allam, L., El Hafidi, N., El Jaoudi, R., Allali, I., Marchoudi, N., Fekkak, J., Benrahma, H., Nejjari, C., Ibrahimi, A., 2020. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. PLoS One 15 (11), e0240345. https://doi.org/10.1371/journal.pone.0240345.

Laiton-Donato, K., Usme-Ciro, J.A., Franco-Muñoz, C., Álvarez-Díaz, D.A., Ruiz-Moreno, H.A., Reales-González, J., Prada, D.A., Corchuelo, S., Herrera-Sepúlveda, M. T., Naizaque, J., Santamaría, G., Wiesner, M., Walteros, D.M., Ospina Martínez, M.L., Mercado-Reyes, M., 2021. Novel highly divergent SARS-CoV-2 lineage with the spike substitutions L249S and E484K. Front. Med. 8, 697605 https://doi.org/10.3389/fmed.2021.697605.

Li, C., Debruyne, D.N., Spencer, J., Kapoor, V., Liu, L.Y., Zhou, B., Pandey, U., Bootwalla, M., Ostrow, D., Maglinte, D.T., Ruble, D., Ryutov, A., Shen, L., Lee, L., Feigelman, R., Burdon, G., Liu, J., Oliva, A., Borcherding, A., Liu, Z., 2020. Highly sensitive and full-genome interrogation of SARS-CoV-2 using multiplexed PCR enrichment followed by next-generation sequencing. BioRxiv. https://doi.org/10.1101/2020.03.12.988246, 2020.03.12.988246.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37 (5), 1530–1534. https://doi.org/10.1093/molbev/msaa015.

Ogawa, J., Zhu, W., Tonnu, N., Singer, O., Hunter, T., Ryan, A.L., Pao, G.M., 2020. The D614G mutation in the SARS-CoV2 Spike protein increases infectivity in an ACE2 receptor dependent manner. BioRxiv: Preprint Server Biol. https://doi.org/10.1101/2020.07.21.214932.

Outbreak.info, 2021. Outbreak.Info. Recuperado el 18 de noviembre de 2021, de. https://outbreak.info/.

Ovsyannikova, I.G., Haralambieva, I.H., Crooke, S.N., Poland, G.A., Kennedy, R.B., 2020. The role of host genetics in the immune response to SARS-CoV-2 and COVID-19 susceptibility and severity. Immunol. Rev. 296 (1), 205–219. https://doi.org/10.1111/imr.12897.

Pastrian-Soto, G., 2020. Genetic and molecular basis of COVID-19 (SARS-CoV-2) mechanisms of pathogenesis and immune. Int. J. Odontostomatol. 14 (3), 331–337. https://doi.org/10.4067/S0718-381X2020000300331.

Petrich, A., Mahony, J., Chong, S., Broukhanski, G., Gharabaghi, F., Johnson, G., Louie, L., Luinstra, K., Willey, B., Akhaven, P., Chui, L., Jamieson, F., Louie, M., Mazzulli, T., Tellier, R., Smieja, M., Cai, W., Chernesky, M., Richardson, S.E., Ontario Laboratory Working Group for the Rapid Diagnosis of Emerging Infections, 2006. Multicenter comparison of nucleic acid extraction methods for detection of severe acute respiratory syndrome coronavirus RNA in stool specimens. J. Clin. Microbiol. 44 (8), 2681–2688. https://doi.org/10.1128/JCM.02460-05.

Posada, D., 2008. jModelTest: Phylogenetic model averaging. Mol. Biol. Evol. 25 (7), 1253–1256. https://doi.org/10.1093/molbev/msn083.

Ramanathan, M., Ferguson, I.D., Miao, W., Khavari, P.A., 2021. SARS-CoV-2 B.1.1.7 and B.1.351 spike variants bind human ACE2 with increased affinity. Lancet Infect. Dis. 21 (8), 1070. https://doi.org/10.1016/S1473-3099(21)00262-0.

Rambaut, A., Holmes, E.C., Hill, V., O'Toole, Á., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. BioRxiv. https://doi.org/10.1101/2020.04.17.046086, 2020.04.17.046086.

Ren, Y., Shu, T., Wu, D., Mu, J., Wang, C., Huang, M., Han, Y., Zhang, X.-Y., Zhou, W., Qiu, Y., Zhou, X., 2020. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. Cell. Mol. Immunol. 17 (8), 881–883. https://doi.org/10.1038/s41423-020-0485-9.

Sharma, A., Tiwari, S., Deb, M.K., Marty, J.L., 2020. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. Int. J. Antimicrob. Agents 56 (2), 106054. https://doi.org/10.1016/j.ijantimicag.2020.106054.

Tang, J.W., Toovey, O.T.R., Harvey, K.N., Hui, D.D.S., 2021. Introduction of the south African SARS-CoV-2 variant 501Y.V2 into the UK. J. Infect. 82 (4), e8–e10. https://doi.org/10.1016/j.jinf.2021.01.007.

Tsai, P.-H., Wang, M.-L., Yang, D.-M., Liang, K.-H., Chou, S.-J., Chiou, S.-H., Lin, T.-H., Wang, C.-T., Chang, T.-J., 2020. Genomic variance of open Reading frames (ORFs) and spike protein in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). J. Chinese Med. Assoc. 83 (8), 725–732. https://doi.org/10.1097/JCMA.0000000000000387.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., Ortiz, A.T., Balloux, F., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genetics Evol. J. Mol. Epidemiol. Evol. Genetics Infect. Dis. 83, 104351 https://doi.org/10.1016/j.meegid.2020.104351.

Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., Zhang, Z., 2020. The establishment of reference sequence for SARS-CoV-2 and variation analysis. J. Med. Virol. 92 (6), 667–674. https://doi.org/10.1002/jmv.25762.

Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., Wei, G.-W., 2021. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun. Biol. 4 (1), 228. https://doi.org/10.1038/s42003-021-01754-6.

Wen, F., Yu, H., Guo, J., Li, Y., Luo, K., Huang, S., 2020. Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2. J. Infect. 80 (6), 671–693. https://doi.org/10.1016/j.jinf.2020.02.027.

Zeng, W., Liu, G., Ma, H., Zhao, D., Yang, Y., Liu, M., Mohammed, A., Zhao, C., Yang, Y., Xie, J., Ding, C., Ma, X., Weng, J., Gao, Y., He, H., Jin, T., 2020. Biochemical characterization of SARS-CoV-2 nucleocapsid protein. Biochem. Biophys. Res. Commun. 527 (3), 618–623. https://doi.org/10.1016/j.bbrc.2020.04.136.

Zinzula, L., 2021. Lost in deletion: the enigmatic ORF8 protein of SARS-CoV-2. Biochem. Biophys. Res. Commun. 538, 116–124. https://doi.org/10.1016/j.bbrc.2020.10.045.