Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Phylogenetic analysis of SARS-CoV-2 viruses circulating in the South American region: Genetic relations and vaccine strain match

Paula Perbolianachis, Diego Ferla, Rodrigo Arce, Irene Ferreiro, Alicia Costábile, Mercedes Paz, Diego Simón, Pilar Moreno, Juan Cristina*

Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay

ARTICLE INFO

Keywords: Coronavirus Evolution SARS-CoV-2 South America COVID-19

ABSTRACT

The pandemic of coronavirus disease 2019 (COVID-19) is caused by a novel member of the family Coronaviridae, now known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Recent studies revealed the emergence of virus variants with substitutions in the spike and/or nucleocapsid and RNA-dependent RNA polymerase proteins that are partly responsible for enhanced transmission and reduced or escaped anti-SARS-CoV-2 antibodies that may reduce the efficacy of antibodies and vaccines against the first identified SARS-CoV-2 strains. In order to gain insight into the emergence and evolution of SARS-CoV-2 variants circulating in the South American region, a comprehensive phylogenetic study of SARS-CoV-2 variants circulating in this region was performed. The results of these studies revealed sharp increase in virus effective population size from March to April of 2020. At least 62 different genotypes were found to circulate in this region. Variants of concern (VOCs) Alpha, Beta, Gamma and Delta co-circulate in the region, together with variants of interest (VOIs) Lambda, Mu and Zeta. Most of SARS-CoV-2 variants circulating in the South American region belongs to B.1 genotypes and have substitutions in the spike and/or nucleocapsid and polymerase proteins that confer high transmissibility and/or immune resistance. 148 amino acid positions of the spike protein and 70 positions of the nucleocapsid were found to have substitutions in different variants isolated in the region by comparison with reference strain Wuhan-Hu-1. Significant differences in codon usage among spike genes of SARS-CoV-2 strains circulating in South America was found, which can be linked to SARS-CoV-2 genotypes.

1. Introduction

The pandemic of coronavirus disease 2019 (COVID-19) started in China in December of 2019 (Li, Guan and Wu, 2020) (Li et al., 2020). This severe respiratory pneumonia is caused by a novel member of the family Coronaviridae, now known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Gorbalenya, Baker and Baric, 2020; Gorbalenya et al., 2020). The World Health Organization declare this SARS-CoV-2 pandemic as a public health emergency of international concern on January 30th, 2020 (World Health Organization 2020a; WHO, 2020a). As December 14th, 2021, there have been more than 270 million confirmed cases worldwide and the global deaths of SARS-CoV-2 disease surpasses 5 million people (World Health Organization 2020b) (WHO, 2020b).

As all members of the family Coronaviridae, SARS-CoV-2 possess a single stranded, positive-sense RNA genome of approximately 30 kb bases in length, which encodes for multiple structural and non-structural proteins. The structural proteins include the spike (S) protein, the envelope (E) protein, the membrane (M) protein, and the nucleocapsid (N) protein (Chen, Liu and Guo, 2020; Chen et al., 2020).

The spike (S) glycoprotein of SARS-CoV-2 facilitates coronavirus entry into host cells. The S protein forms a homotrimeric complex protruding from the viral surface and consists of two functional subunits, S1 and S2, which are responsible for host cell receptor binding and the viral fusion to the host cellular membranes (Xia, 2021) (Xia et al., 2021). The smaller S1 subunit consists of an N-terminal domain (NTD) and three C-terminal domains (CTD1–3), of which CTD1 forms the receptor-binding domain (RBD) and contributes to the stabilization of the membrane-anchored S2 subunit. The larger S2 subunit contains the machinery for viral fusion and comprises a hydrophobic fusion peptide (FP), heptad repeat 1 (HR1), central helix (CH), connector domain (CD), heptad repeat 2 (HR2), transmembrane domain (TM) and cytoplasmic

* Correspondence author.
E-mail address: cristina@cin.edu.uy (J. Cristina).

https://doi.org/10.1016/j.virusres.2022.198688
Received 27 October 2021; Received in revised form 19 January 2022; Accepted 20 January 2022
Available online 21 January 2022
0168-1702/© 2022 Elsevier B.V. All rights reserved.
2. Material and methods

2.1. Sequences

Available and comparable complete genome sequences of 933 SARS-CoV-2 variants isolated in South America from March 12th, 2020 to May 28th, 2021, were used throughout these studies (including Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Suriname, Uruguay and Venezuela). In the selection of the sequences to be included in the dataset, we carefully selected only those ones with a minimum of N in the full-length genome sequence. Sequences were obtained from the Global Initiative on Sharing Avian Influenza Data (GISAID) database.

For accession numbers, country of origin and date of isolation, see Supplementary Material Table 1.

2.2. Sequence alignment

Sequences were aligned using MAFFT version 7 program (Katoh, Rozewicki and Yamada, 2019) (Katoh et al., 2019).

2.3. Data analysis

Nucleotide frequencies, codon and amino acid usage and relative synonymous codon usage (RSCU) (Sharp and Li, 1986; Sharp and Lee, 1986) of S proteins from SARS-CoV-2 variants isolated in South America were calculated using the program CodonW (written by John Peden) as implemented in the Galaxy server version 1.4.4 (Afgan, Baker and Batut, 2018; Afgan et al., 2018). The relationship between compositional variables and samples was obtained using Principal Component Analysis (PCA). Singular value decomposition (SVD) method was used to calculate the PCA method. The unit variance was used as the scaling method. This means that all variables are scaled so that they will be equally important (variance = 1) when finding the components. By the same approach, Heatmaps were also constructed, which is a data matrix for visualizing values in the dataset by the use of a color gradient. Rows and/or columns of the matrix are clustered so that sets of rows or columns rather than individual ones can be interpreted. PCA and Heatmaps analysis were done using the ClustVis program (Metsalu and Vilo, 2015; Metsalu and Vilo, 2015).

Correspondence analysis (COA) is another multivariate statistical analysis. This method was used to analyze the RSCU of the S genes of SARS-CoV-2 variants enrolled in these studies. COA allows a geometrical representation of the sets of rows and columns in a dataset. Each ORF is represented as a 59-dimensional vector and each dimension corresponds to the RSCU value of one codon (excluding AUG, UGG, and stop codons). Major trends within a dataset can be determined using measures of relative inertia and genes ordered according to their position along the different axes (Greenacre, 1994; Greenacre, 1994). COA was performed on the RSCU values using the CodonW program (Afgan et al., 2018).

2.4. Bayesian Markov chain Monte Carlo analysis

To investigate the evolutionary patterns of SARS-CoV-2 variants circulating in the South American region, a Bayesian Markov Chain Monte Carlo (MCMC) approach was used as implemented in the BEAST package v2.5.2 (Bouckaert et al., 2019; Bouckaert et al., 2019). First, the evolutionary model that best fit the sequence dataset was determined using software from the IQ-TREE program (Trifonopoulos, Nguyen, von Haeseler and Minh, 2016; Trifonopoulos et al., 2016). Bayesian information criterion (BIC), Akaike information criterion (AIC), and the log of the likelihood (LnL), indicated that the GTR+I+Γ model was the most suitable model (BIC = 16,824.83; AIC = 13,848.63; LnL = −6346.31). Both strict and relaxed molecular clock models were used to test different dynamic models (constant population size, exponential population growth, expansion population growth, logistic population growth and Bayesian Skyline). Statistical uncertainty in the data was reflected by the 95% highest probability density (HPD) values. Results were examined using the TRACER v1.6 program (available from http://beast.bio.ed.ac.uk/Tracer). Convergence was assessed by effective sample sizes (ESS) above 200. Models were compared by AICM from the likelihood output of each of the models using TRACER v1.6 program. Lower AICM values indicate better model fit. The Bayesian Skyline model was the best model to analyze the data. Maximum clade
proteins using artificial neural networks that examine the sequence using the NetNGlyc 1.0 Server (Gupta and Brunak, 2002; Gupta and approach, potential O-linked glycosylation sites were predicted using potential score was set to predict glycosylated sites. By the same context of Asn-Xaa-Ser/Thr sequons. A threshold value of 2.7. Prediction of N- and O-linked glycosylation sites in Spike protein A.1.1.1.1 would become C.1; A.1.1.1.2 would become C.2).

6LZG. Visualization was done using Jmol-14.0.4 software (available at:
the 3D structure of receptor binding domain complexed with its acceptor
2.9. Mapping of amino acid substitutions in a 3D structure of the receptor binding domain of Spike protein

A sharp increase in effective population size from March to April of 2020 was observed. Then, a small descent is observed towards the end of 2020 and later recovery in the first months of 2021 to remained constant towards the end of the period covered by these studies (see Fig. 2A). This is in agreement with the epidemiology observed in the region, where a sharp increase in the number of cases and deaths was observed in the region at the beginning of the pandemic period in South America, a small descent towards the end of 2020 and a new increase in the months of 2021 covered by these studies (Dong et al., 2020) (see Fig. 2B).

To study the phylogenetic relationships among SARS-CoV-2 strains isolated in South America, maximum clade credibility trees were generated using software from the BEAST package (Rambaut et al., 2020). The results of these studies are shown in Supplementary Material Fig. 1. The results of these studies revealed that both main SARS-CoV-2 genetic lineages (A and B) have been circulating in the South American region (see Supplementary Material Fig. 1). Lineage A was found circulating in Chile, Peru and Uruguay at the beginning of the pandemic period. This lineage is considered at the root of the pandemic, like Wuhan/WH04/2020 (EPI_ISL_406801), and share two nucleotide positions in SARS-CoV-2 genome (positions 8782 in ORF1ab and 28,144 in ORF8) with the closest known relative being a bat virus (RaTG13).
This is in agreement with ongoing studies showing that the lineage B (particularly B.1) have spread and replaced the lineage A in several different countries (Korber, Fischer and Gnanakaran, 2020; Korber et al., 2020). An extensive co-circulation of different genetic lineages is observed in several countries in the region (Supplementary Material Fig. 1). Most of SARS-CoV-2 variants circulating in the South American region belongs to B.1 genotypes. These genotypes have diversified from its entry in the region. VOCs Alpha, Beta, Gamma and Delta were found to co-circulate in the region during the period covered by these studies (March 12th, 2020 to May 28th, 2021) (see Supplementary Material Fig. 1). Since most of the lineages found to circulate in the South American region were B.1 lineages, they carry an amino acid substitution at position 614 of the S protein (D614G). Variants having this substitution have shown to be more transmissible (Hou et al., 2020; Volz, Hill and McCrone, 2021; Volz et al., 2021).

### Table 1

Bayesian coalescent inference of SARS-CoV-2 strains isolated in South America.

| Group             | Parameter | Value      | HPD       | ESS    |
|-------------------|-----------|------------|-----------|--------|
| SARS-CoV-2 full-length sequences | Tree Likelihood | -52,353.03 | -53,279.00 to -53,228.42 | 1189.15 |
|                   | tMRCA    | 1.564      | 1.305 to 1.943 | 273.36 |

*See Supplementary Material Table 2 for strains included in this analysis.

**In all cases, the mean values are shown.

* HPD, high probability density values.

* ESS, effective sample size.

* tMRCA, time of the most common recent ancestor is shown in years. The date estimated for the tMRCA is indicated in bold.

(Chan, Kok and Zhu, 2020). This is in agreement with ongoing studies showing that the lineage B (particularly B.1) have spread and replaced the lineage A in several different countries (Korber, Fischer and Gnanakaran, 2020; Korber et al., 2020). An extensive co-circulation of different genetic lineages is observed in several countries in the region (Supplementary Material Fig. 1). Most of SARS-CoV-2 variants circulating in the South American region belongs to B.1 genotypes. These genotypes have diversified from its entry in the region. VOCs Alpha, Beta, Gamma and Delta were found to co-circulate in the region during the period covered by these studies (March 12th, 2020 to May 28th, 2021) (see Supplementary Material Fig. 1). Since most of the lineages found to circulate in the South American region were B.1 lineages, they carry an amino acid substitution at position 614 of the S protein (D614G). Variants having this substitution have shown to be more transmissible (Hou et al., 2020; Volz, Hill and McCrone, 2021; Volz et al., 2021).

#### 3.2. Substitutions were found in all domains of the S protein

SARS-CoV-2 S protein plays a key role in virus biology, epidemiology and adaptation of virus to its human host. Moreover, almost all vaccine candidates against SARS-CoV-2 are based on the S protein (Xia, 2021).

In order to contribute to a better understanding of the results found in these phylogenetic analyses and to understand the relation among SARS-CoV-2 circulating in the region and vaccines, the complete amino
The thick solid black line represents the median estimate, and the blue area shows the 95% highest probability density (HPD) values (Bouckaert et al., 2019). Time is shown in the x-axis in years. In (B) the daily number of cases and deaths due to COVID-19 in South America are shown in blue and red, respectively. Time is shown in the x-axis as dates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Nucleocapsid substitutions that increase infectivity and fitness were found in strains isolated in South America

Although S amino acid sequence resulted to be roughly conserved at population level in strains isolated in South America, 148 amino acid positions of the S protein were found to have substitutions in different variants isolated in South America by comparison with reference strain Wuhan-Hu-1 (see Fig. 3).

In order to establish which of them were found significantly represented in the population studied, we employed SeqLogo (Thomsen and Nielsen, 2012) (see Fig. 3 and Supplementary Material Figure 2). Although S amino acid sequence resulted to be roughly conserved at population level in strains isolated in South America, we detected 11 sites in S protein with significant polymorphisms. Five of these polymorphisms were found in the N-terminal domain (substitutions L18F, T20N, P26S, D138I, R190S); three in the receptor binding domain (K417T, E484K and N501Y) and substitution H653Y in the S1 sub-unit of S protein. Substitution T1027I and V1176F were found in the S2 sub-unit, being the last one in the heptad repeat 2. Substitution L18F have been detected in SARS-CoV-2 clinical isolates, mainly from VOCS Alpha (B.1.1.7), Beta (B.1.351) and Gamma (P.1) and it is known to affect loop 1 of the antigenic supersite in the NTD domain (McCallum, De Marco and Lempp, 2021a; McCallum et al., 2021a). The finding that multiple circulating SARS-CoV-2 variants map to the NTD, several of them in the antigenic supersite (site I), suggests that the NTD is subject to a strong selective pressure from the host humoral immune response (McCallum, Bassi and De, 2021b; McCallum et al., 2021b).

Due to the fact that the receptor-binding domain (RBD) of S protein is vital for virus attaching to the host receptor and triggering a conformational change in the protein that results in fusion with the host cell membrane, we mapped the substitutions sites where significantly polymorphisms was found in the RBD of S protein in the 3D structure of RBD complexed with its acceptor ACE2. The results of these studies are shown in Fig. 4. Substitutions K417T, E484K and N501Y map in the RBD region that interacts with ACE2 protein, in agreement with recent results (Winger and Caspari, 2021) (Winger and Caspari, 2021).

3.4. RNA-dependent RNA polymerase substitutions conferring epidemiological advance over Wuhan strains were found in strains circulating in South America

RNA-dependent RNA polymerase (RdRp or nsp12) is a key player in the synthesis of viral RNA (Ilmijärvi, Abdul and Acosta-Gutierrez, 2021; Ilmijärvi et al., 2021). The structure of the SARS-CoV-2 nsp12 contains a cupped, right-handed RdRp domain linked to a nidovirus RdRp-associated nucleotidyl-transferase domain (NiRAN) via an interface face (see Fig. 6). The RdRp domain adopts the conserved architecture of the viral polymerase family 4 and is composed of three domains: a fingers domain, a palm domain and a thumb domain (Fig. 6). The crystal structure of SARS-CoV-2 nsp12 in complex with its non-structural protein 7 and 8 (nsp7 and nsp8) co-factors underlines the
central role these nsp in the replication and transcription of the virus (Gao, Yan and Huang, 2020; Gao et al., 2020). Recent studies have revealed that substitution P323L in RdRp together with D614G in the Spike protein have shown that G614/L323 variants are epidemiologically highly successful and replaced the original D614/P323 variants (Ilmarv, Abdul and Acosta-Gutiérrez, 2021; Ilmarv et al., 2021). For these reasons, the same studies performed for S and N proteins were performed for the RdRp proteins for the same strains enrolled in S and N studies. The results of these studies are shown in Fig. 6. We have found that 98% of SARS-CoV-2 strains circulating in South America and enrolled in these studies carried the P323L in the RdRp protein (see Fig. 6). This substitution is located on the interface domain. We have observed 14 substitutions in the NiRAN domain, 7 substitutions in the interface domain, and 12, 9, and 5 substitutions in finger, palm and thumb domains of the RdRp proteins in strains isolated in South America (Fig. 6). The RdRp protein were found to be significantly conserved among variants and not polymorphic sites were observed among the RdRp of strains isolated in South America and enrolled in these studies.

3.5. A significant bias in nucleotide frequencies, codon and amino acid usage among S proteins

PCA is a statistic technique used to describe a collection of data in terms of variables (components) not correlated among themselves. The
components are ordered according to the original variation that they describe, and for this reason it is a very useful technique to reduce the dimensionality of the collection of data while preserving as much of the data’s variation as possible.

In order to gain insight into the trends of evolution of the S protein, a composition analysis among the 933 S proteins enrolled in these studies was performed. For this purpose, the nucleotide frequencies for first, second and third codon positions were established for S genes from 933 SARS-CoV-2 variants circulating in South America and PCA was performed (see Supplementary Material Fig. 3).

A significant bias in nucleotide frequencies was found in the S gene sequences from SARS-CoV-2 variants enrolled in these studies. In fact, PC1 component (that accounts for the 85% of the total variation observed) has a strong positive correlation with A and U frequencies at the third codon position and a strong negative correlation with C and G frequencies at that position (see Supplementary Material Fig. 3). This bias in preferences for A and U ended codons accounts for highly preferred CCU, CUU, GAU, GGU, GCU, UUU and CGU codons, and an underrepresentation use of UCG, CUG, CCG, AGC, CGC and GGG codons (see Table 2).

To gain insight into the trends of the variation observed among VOCs circulating in the South American region, codon usage frequencies of 86 S genes from SARS-CoV-2 strains belonging to VOCs Alpha, Beta, Gamma and Delta were determined and PCA and Heatmap analysis was performed (for strains included in these analyses, see Supplementary Material Table 2) (see Fig. 7). This analysis revealed significant
differences in codon usage among the S genes of SARS-CoV-2 VOCs circulating in South America, which can be linked to SARS-CoV-2 genotypes (see Fig. 7 A and B). When the same analysis is done for amino acid usage frequencies, the same results are also found (see Supplementary Material Figure 4).

Since codon usage by its very nature is multivariate, it is necessary to analyze the data using different and complementary approaches. To confirm the results outlined above, we performed a COA analysis on the RSCU values for the same 86 S genes from VOCs strains and examined the distribution of the genes along the plane determined by the first two principal axes of COA. The results of these studies are shown in Supplementary Material Fig. 5. The distribution of the genes in the plane defined by the first two major axes of COA showed the same results, revealing that different S genes are located at different places, which again are linked to VOCs genotypes (see Supplementary Material Fig. 5).

Table 2
Codon usage in SARS-CoV-2 S genes from strains isolated in South America*.

| AA  | Cod | Frequency | AA  | Cod | Frequency | AA  | Cod | Frequency | AA  | Cod | Frequency |
|-----|-----|-----------|-----|-----|-----------|-----|-----|-----------|-----|-----|-----------|
| Phe | UUU | 1.54      | Ser | UCU | 2.25      | Tyr | UAU | 1.49      | Cys | UGU | 1.40 |
|     | UUC | 0.46      |     | UCC | 0.72      |     | UAC | 0.51      |     | UGC | 0.60 |
| Leu | UUA | 1.56      |     | UCA | 1.57      | TER | UAA | ***       |     | TER | *** |
|     | UUG | 1.12      |     | UCG | 0.12      |     | UAG | ***       |     | Trp | UGG |
|     | CUG | 1.98      |     | CCC | 0.28      |     | CAU | 1.51      |     | Arg | CGU |
|     | CUA | 0.50      |     | CCA | 1.73      |     | CAG | 0.51      |     | CGA | 0.00 |
|     | CUG | 0.17      |     | CCG | 0.00      |     |     |           |     |     |           |
| Ile | AUU | 1.74      |     | ACC | 1.81      |     | ACA | 1.65      | Lys | AAA | 1.26 |
|     | AUG | 0.71      |     | ACA | 0.40      |     |     |           |     | Arg | AGA |
|     | AUA | 1.00      |     | ACG | 0.13      |     | AAG | 0.74      |     | AGG | 1.41 |
|     | Val | GUU | 1.97      |     | GCC | 2.12      | Asp | GAU | 1.37      | Gly | GGU |
|     | GUC | 0.87      |     | GCC | 0.41      |     | GAC | 0.63      |     | GGC | 0.72 |
|     | GUA | 0.54      |     | GCA | 1.37      | Glu | GAA | 1.41      |     | GGA |
|     | GUG | 0.54      |     | GCG | 0.10      |     | GAG | 0.59      |     | GGG | 0.15 |

* Average frequencies in 1:187,463 codons. AA, amino acid; Cod, codons; TER, termination codons. Preferred codons (∆ ≥ 0.30) are shown in bold. Underrepresented codons are shown in italics.

4. Discussion

SARS-CoV-2 has spread across the world, causing a health threat of international concern. As the virus circulation becomes widespread, phylodynamic analyses can give insight into how the virus spreads both spatially and temporally. Moreover, viruses from a given region can be placed in the context of those circulating globally, allowing for the number of independent virus introductions into a region to be estimated in these analyses (Rambaut et al., 2020; Rambaut et al., 2020). Phylo-
dynamic analyses can also be extremely useful to study viral adaptation, a particular concern since SARS-CoV-2 has recently spilled to humans.

The results of these studies suggest that SARS-CoV-2 variants circulating in South America evolved from ancestors that existed around November 26th, 2019 (Table 2). Similar studies carried out at the beginning of the pandemic on SARS-CoV-2 strains circulating in the Hubei province of China trace the case index to November 9th, 2019 (Pekar et al., 2021; Pekar et al., 2020). These results revealed a rapid transmission of SARS-CoV-2 strains from China to South America.
Emerging SARS-CoV-2 variants with an amino acid substitution at position 614 of the S protein (D614G) have shown to be more transmissible (Hou, Chiba and Halfmann, 2020; Hou et al., 2020; Faria et al., 2021; Volz, Hill and McCrone, 2021; Volz et al., 2021). Although G614 resides at a fair distance from the RBD, it affects the ACE2 binding site through an allosteric link with T500 (Zhang et al., 2021; Zhang et al., 2021; Omotuyi et al., 2021) Omotuyi et al., 2021). This is in agreement with the results of this work, since most of the lineages found to circulate in the South American region where B.1 lineages carry this substitution (see Figs. 1 and Supplementary Material Fig. 1). By the same token, most of the variants enrolled in these studies also carried substitutions R203K and G204R in the N protein (see Fig. 5), providing an enhanced replication advance by comparison with strains isolated early in the pandemic, as well as enhanced infectivity and disease severity in hamster model (Wu et al., 2021). Moreover, 98% of the strains enrolled in these studies carried the P232L substitution in the RdRp. Taking these results together, this may help to explain, at least in part, the sharp increase in population size observed at the beginning of the pandemic in the South American region (see Fig. 2).

An emergence of Zeta (P.2) variants from a parental cluster of B.1.28 strains was observed in these studies (see Table 1). Besides the D614G, these strains also carry the substitution E484K in the receptor binding domain (RBD) of S protein (see also Fig. 3). Recent studies suggest that the substitution E484K permits virus variants to be selected as an escape mutation in the presence of neutralizing antibodies or plasma from immune humans in vitro (Weisblum et al., 2020; Weisblum et al., 2020). Interestingly, studies carried out in England and Wales suggest the accumulation of substitution E484K in the Alpha (B.1.1.7) background is the result of the vaccination program (Collier et al., 2021; Collier et al., 2021). This substitution, which is situated at the RBD interface with the receptor ACE2 (see Figs. 3 and 4), may also play an important role in neutralization escape of VOI Zeta as well of VOCs Beta and Gamma, that share this substitution (Salleh et al., 2021).

This same parental cluster B.1.28 give rise to the emergence of VOC Gamma (P.1) (see Supplementary Material Fig. 1). This variant was first observed in Manaus, Brazil, in December of 2020, where a sharp increase in the total COVID-19 infections, followed by an increase in the number of hospital admissions was observed (Faria et al., 2021; Faria et al., 2021; Candido, Claro and de, 2020; Candido, 2020). This variant shares with VOC Beta (B.1.351) three important substitutions in the RBD of the S protein: K417T, E484K and N501Y (see Fig. 3). These substitutions have been shown to increase the binding affinity of the S protein to its receptor ACE2, particularly substitution N501Y significantly contribute to this increase in binding affinity (Luan, Wang and Huynh, 2021; Luan et al., 2021; Ali, Kasry and Amin, 2021; Ali et al., 2021). Moreover, recent studies revealed that N439K viruses have similar in vitro replication fitness as compared to wild type, while at the same time N439K substitutions confers resistance against several neutralizing monoclonal antibodies (Thomson et al., 2021; Thomson et al., 2021). This highlights the importance of for molecular surveillance in all regions of the world to guide development and usage of vaccines and therapeutics.

VOI Mu was detected in Colombia (see Fig. 1). This VOI was isolated for the first time in January 11th, 2021. VOI Mu have substitutions in the S protein, as T93I, Y144S, Y145N in the N-terminal domain; R346K, E484K and N501Y in the receptor-binding domain and D614G, P681H and D950N in other regions (Uriu, Kimura and Shirakawam, 2021; Uriu et al., 2021). In these studies, Colombian VOI Mu variants having substitution Y144T were also observed (see Fig. 3). Several of these substitutions have been identified in other VOCs: e.g., E484K in Beta and Gamma, N501Y in Alpha and Beta, P681H in Alpha, and D950N in Delta. Virus neutralization studies, performed with the use of serum samples obtained from persons who had recovered from Covid-19 and who were infected early in the pandemic (April through September 2020), showed that the VOI Mu was 10.6 times as resistant to neutralization as the B.1 lineage parental virus (Uriu, Kimura and Shirakawam, 2021; Uriu et al., 2021). Similar studies using sera from persons who had received the BNT162B2 vaccine showed that the VOI Mu variant was 9.1 as resistant as the parental virus (Uriu, Kimura and Shirakawam, 2021; Uriu et al., 2021). These studies highlight the importance of further studies regarding this variant.

Taking these results together it is possible to observe that most of the SARS-CoV-2 viruses circulating in South America have substitutions in the S and N proteins that confer high transmissibility and/or immune resistance (see Figs. 3 and 5). Therefore, it is extremely important to aim for the best possible vaccination campaign in South America in order to enhance protection against these and newly emerging SARS-CoV-2 variants. Vaccination campaigns in South American countries began in early 2021 (February-March) and most of the countries enrolled in these studies had less than 10% coverage of their population by the period covered by these studies (Dong et al., 2020).

Moreover, 148 amino acid positions in the S protein were found to have substitutions in variants from South America in the period covered by these studies (March 10th, 2020 to May 28th, 2021) (see Fig. 5). Continuous molecular surveillance of SARS-CoV-2 will be necessary to detect new variants of the virus with clinical relevance. This is extremely important to improve programs to control the virus (Flores-Alanis, Cruz-Rangel and Rodriguez-Gomez, 2021; Flores et al., 2021).

The SARS-CoV-2 S protein is covered by a shield of N-linked and O-linked glycans (Lo Presti, Rezza and Stefanelli, 2020; Lo Presti et al., 2020; Shajahan, Supekar, Gleinich and Azadi, 2020; Shajahan et al., 2020). An important fact in the development of effective subunit vaccines is the characterization of the glycosylation of key viral proteins, since they play a crucial role in immune recognition affecting vaccine designs. Glycosylation of viral surface proteins is extremely important for immune shielding and altering positions where glycosylation sites occur is a well-known immune evasion mechanism in viruses (Walls et al., 2016; Walls et al., 2016). From the predicted putative N- or O-glycosylated sites found in these studies (21 N- and 18 O-glycosylated sites), no substitutions were found in N- glycosylation sites and 4 O-glycosylation sites were found to have substitutions (Fig. 3). The results of these studies suggest that glycosylation sites are roughly conserved among S protein from SARS-CoV-2 viruses. More studies will be needed in order to address the effect of substitutions in glycosylation sites of this protein.

In these studies, a biased nucleotide composition was found in the S proteins of SARS-CoV-2 variants circulating in South America (Fig. 5). While S genes composition analysis revealed a positive correlation of U and A at the third codon positions, a negative correlation was found for C and G at these positions (see Fig. 5). This is in agreement with previous work carried out in other coronaviruses, where A/G bias is a relatively stable property shared in the family, while the C/U bias differs significantly per virus type (Berkhout and van Hemert, 2015; Berkhout and Hemert, 2015). These biases also have a major influence on derived parameters as codon usage (see Table 2). PCA and Heatmap analysis revealed correlation among codon usage and genotypes in the S protein from VOC’s strains (see Fig. 5). Interestingly, these results also demonstrate that S genes have suitable genetic information for clear assignment of emerging VOCs to its specific genotypes (see Fig. 6).

5. Conclusions

The results of these studies revealed that at least 62 different genotypes of SARS-CoV-2 were found to circulate in South America. Most of the variants circulating in this region belongs to B.1 genotypes. From the 933 South American isolates enrolled in these studies 98.82% has a D614G substitution in the spike protein, while 89% of them have substitutions R203K and G204R in the nucleocapsid protein and 98% have the substitution P323L. All VOCs (Alpha, Beta, Gamma and Delta) co-circulate in the region, together with VOI Lambda, VOI Mu and Zeta. Substitutions were found in all S protein domains in different variants from South America by comparison with reference strain Wuhan-Hu-1.
11 polymorphic sites were detected in the S protein of SARS-CoV-2 sequences from strains circulating in South America. Particularly, polymorphisms were found in the receptor binding domain with substitutions in positions that interact with ACE2 cellular receptor. Significant trends of variation in codon and amino acid usage were observed among VOCs circulating in the South American region. This variation can be linked to VOCs genotypes. These results of these studies suggest that codon usage plays an important role in shaping the evolution of S proteins from SARS-CoV-2 variants.

Funding

This work was supported by Agencia Nacional de Investigación e Innovación, PEDECIBA and Comisión Sectorial de Investigación Científica (Grupos I + D grant), Universidad de la República, Uruguay.

CRediT authorship contribution statement

Paula Perbolianachis: Data curation, Visualization, Investigation. Diego Ferla: Rodrigo Arce: Data curation, Visualization, Investigation. Irene Ferreiro: Data curation, Visualization, Investigation. Alichia Costabile: Data curation, Visualization, Investigation. Mercedes Par: Data curation, Visualization, Investigation. Diego Simón: Pilar Moreno: Writing – review & editing. Juan Cristina: Conceptualization, Methodology, Writing – original draft.

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.

Acknowledgments

This research was funded by Agencia Nacional de Investigación e Innovación and PEDECIBA, Uruguay. We acknowledge Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay, for support through Grants I + D grant. We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID. We thank Dr. Gonzalo Moratorio for critical reading of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.virusres.2022.198688.

Supplementary materials associated with this article can be found, in the online version, at doi: 10.1016/j.viruses.2022.198688.

References

Afgan, E., Baker, D., Batt, B., et al., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379.
Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding interest or personal relationships that could have appeared to influence the work reported in this paper.

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.

Acknowledgments

This research was funded by Agencia Nacional de Investigación e Innovación and PEDECIBA, Uruguay. We acknowledge Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay, for support through Grants I + D grant. We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID. We thank Dr. Gonzalo Moratorio for critical reading of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.virusres.2022.198688.

Supplementary materials associated with this article can be found, in the online version, at doi: 10.1016/j.viruses.2022.198688.

References

Afgan, E., Baker, D., Batt, B., et al., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379.
Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding interest or personal relationships that could have appeared to influence the work reported in this paper.

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.

Acknowledgments

This research was funded by Agencia Nacional de Investigación e Innovación and PEDECIBA, Uruguay. We acknowledge Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay, for support through Grants I + D grant. We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID. We thank Dr. Gonzalo Moratorio for critical reading of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.virusres.2022.198688.

Supplementary materials associated with this article can be found, in the online version, at doi: 10.1016/j.viruses.2022.198688.

References

Afgan, E., Baker, D., Batt, B., et al., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379.
Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding interest or personal relationships that could have appeared to influence the work reported in this paper.

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.

Acknowledgments

This research was funded by Agencia Nacional de Investigación e Innovación and PEDECIBA, Uruguay. We acknowledge Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay, for support through Grants I + D grant. We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID. We thank Dr. Gonzalo Moratorio for critical reading of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.virusres.2022.198688.

Supplementary materials associated with this article can be found, in the online version, at doi: 10.1016/j.viruses.2022.198688.

References

Afgan, E., Baker, D., Batt, B., et al., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379.
Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding interest or personal relationships that could have appeared to influence the work reported in this paper.

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.

Acknowledgments

This research was funded by Agencia Nacional de Investigación e Innovación and PEDECIBA, Uruguay. We acknowledge Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay, for support through Grants I + D grant. We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID. We thank Dr. Gonzalo Moratorio for critical reading of this manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.virusres.2022.198688.

Supplementary materials associated with this article can be found, in the online version, at doi: 10.1016/j.viruses.2022.198688.

References

Afgan, E., Baker, D., Batt, B., et al., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46, W537–W544. https://doi.org/10.1093/nar/gky379.
Ali, F., Kasry, A., Amin, M., 2021. The new SARS-CoV-2 strain shows a stronger binding interest or personal relationships that could have appeared to influence the work reported in this paper.

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.

Acknowledgments

This research was funded by Agencia Nacional de Investigación e Innovación and PEDECIBA, Uruguay. We acknowledge Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay, for support through Grants I + D grant. We gratefully acknowledge the Originating and Submitting Laboratories for sharing newly identified coronavirus sequences through GISAID. We thank Dr. Gonzalo Moratorio for critical reading of this manuscript.
