
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.

Virus Research 311 (2022) 198688

Contents lists available at ScienceDirect

Virus Research

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

# 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 kg 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

tail (CT) (Walls et al., 2020) (Walls et al., 2020). The HR1 is vital for the S protein stability, maintaining the correct protein fold in the closed pre-fusion conformation (Juraszek et al., 2020). As in all members of the family *Coronaviridae*, the S protein is cleaved by host proteases at the S1/S2 junction; cleavage has been suggested to activate the protein for host membrane fusion through irreversible conformational changes. There is a second cleavage site, S2', located 130 residues from the N terminus of the S2 subunit, which is highly conserved among coronaviruses. Cleavage at the S2' site by host cell proteases is vital for successful infection by coronaviruses (Belouzard, Chu and Whittaker, 2009; Belouzard et al., 2009). SARS-CoV-2 S protein interacts directly with the host cell receptor angiotensin-converting enzyme 2 (ACE2) (Hoffmann et al., 2020). ACE2 is a protease responsible for blood pressure and volume regulation and it is widely expressed on the cell membranes of the lung, heart, kidneys and gastrointestinal tract (Samavati and Uhal, 2020) (Samavati and Uhal, 2020).

The RBD of S protein is immunodominant and the target of 90% of the neutralizing activity present in SARS-CoV-2 immune sera (Piccoli, Park and Tortorici, 2020; Piccoli et al., 2020). Antibodies directed to S protein NTD and to RBD can neutralize with high potency (less than 0.01 μg/mL IC50) (Cerutti et al., 2021). While RBD shows many non-overlapping antigenic sites (Barnes et al., 2020), NTD appears to contain only a single site of vulnerability to neutralization, identified as an antigenic supersite (McCallum, De Marco and Lempp, 2021a; McCallum et al., 2021).

Clinically applied monoclonal antibodies (Weinreich et al., 2021; Weinreich et al., 2021) and vaccinations (Wang and Cheng, 2021) (Wang et al., 2021) have shown a significant success in virus neutralization. However, recent studies on SARS-CoV-2 evolution revealed the emergence of variants with substitutions in the S protein that appears to be more transmissible, increasing the affinity with ACE-2 or are partly responsible for reduced or escaping to anti-SARS-CoV-2 antibodies (Davies et al., 2021; Davies et al., 2021). These variants are known as variants of concern (VOC), i.e.: VOC Alpha (B.1.1.7, first identified in the United Kingdom); VOC Beta (B.1.351, first identified in South Africa); VOC Gamma (P.1, first identified in Brazil) and VOC Delta (B.1.617.2, first isolated in India). Very recently, by November 24th, 2021, a new VOC named Omicron (B.1.1.529) identified in South Africa (Wang and Cheng, 2021; Wang and Chen, 2021). All VOCs have substitutions in the S NTD and/or RBD domains (Chakraborty, Bhattacharya and Sharma, 2021; Chakraborty et al., 2021) (see Supplementary Material Table 1). Besides VOCs, there are variants that are being closely followed because of its prevalence or other phenotypic characteristics. These variants are known as variants of interest (VOI): VOI Eta (B.1.52, first identified in Nigeria); VOI Iota (B.1.53, first identified in the United States); VOI Kappa (B.1.617.1, first identified in India); VOI Lambda (C.37, first identified in Peru), VOI Mu (B.1.621, first identified in Colombia) and VOI Zeta (P.2, first identified in Brazil). These VOIs also have substitutions in S NTD and/or RBD domains (see Supplementary Material Table 1).

Therefore, there is a concern that VOCs and/or VOIs may reduce the efficacy of the anti-SARS-CoV-2 induced-antibodies or eventually evade them completely (Leach et al., 2021; Leach et al., 2021; Salleh, Derrick and Deris, 2021; Salleh et al., 2021).

In order to better understand the emergence, spread and evolution of SARS-CoV-2 variants circulating in the South American region, a comprehensive phylogenetic study of SARS-CoV-2 strains circulating in this region was performed.

## 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 2.

### 2.2. Sequence alignment

Sequences were aligned using MAFFT version 7 program (Katoh, Rozewicki and Yamada, 2019) (Kato 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 análisis

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 (Trifinopoulos, Nguyen, von Haeseler and Minh, 2016; Trifinopoulos et al., 2016). Bayesian information criterion (BIC), Akaike information criterion (AIC), and the log of the likelihood (LnL), indicated that the GTR+┌ + I 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

credibility trees were generated by means of the use of the Tree Annotator program from the BEAST package. Visualization of the annotated trees was done using the FigTree program v1.4.2 (available at: http://tree.bio.ed.ac.uk). A Bayesian Skyline was constructed using TRACER 1.6 software.

### 2.5. Epidemiology data

Daily number of cases and deaths due to COVID-19 in the South American region from Marh 12th, 2020 through May 28th, 2021, was obtained from Our World in Data (Dong, Du and Gardner, 2020; Dong et al., 2020).

### 2.6. SARS-CoV-2 genotype assignment

In order to capture local and global patterns of virus genetic diversity in a timely and coherent manner, we employed Pangolin COVID-19 genetic lineage strain assignment and nomenclature (Rambaut et al., 2020; Rambaut et al., 2020). In an ongoing and rapidly changing pandemic such as the one caused by SARS-CoV-2, this nomenclature system can facilitate real-time epidemiology by providing commonly agreed labels to refer to viruses circulating in different parts of the world (Rambaut et al., 2020). According to this, two main genetic lineages of SARS-CoV-2, named A and B, are still circulating in many countries around the world, reflecting the exportation of viruses from China to elsewhere before strict travel restrictions and quarantine measures were imposed there. Each descending lineage from either A or B is assigned by a numerical value (for example, lineage A.1 or lineage B.2). This iterative procedure refers to a maximum of three sublevels (for example, A.1.1.1) after which new descendant lineages are given a letter (i.e. A.1.1.1.1 would become C.1; A.1.1.1.2 would become C.2).

### 2.7. Prediction of N- and O-linked glycosylation sites in Spike protein

Potential N-linked glycosylation sites in S protein were predicted using the NetNGlyc 1.0 Server (Gupta and Brunak, 2002; Gupta and Brunak, 2002). The NetNglyc server predicts N-Glycosylation sites in proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons. A threshold value of $>0.5$ average potential score was set to predict glycosylated sites. By the same approach, potential O-linked glycosylation sites were predicted using the NetOGlyc-4.0 (Steentoft et al., 2013; Steentoft et al., 2013).

### 2.8. Amino acid sequence profiles

In order to construct and visualize amino acid sequence profiles in the S protein of SARS-CoV-2 variants isolated in South America we used SeqLogo-2.0 (Thomsen and Nielsen, 2012; Thomsen and Nielsen, 2012). This method allows a graphical representation of the information content stored in a multiple sequence alignment (MSA) and provide a highly intuitive representation of the position-specific amino acid composition. Moreover, this method includes sequence weighting to correct for data redundancy and pseudo counts to correct for low number of observations. We employed SeqLogo-2.0 using Kullback-Leibler logo type and 200 pseudo counts (Thomsen and Nielsen, 2012) (Thomsen and Nielsen, 2012).

### 2.9. Mapping of amino acid substitutions in a 3D structure of the receptor binding domain of Spike protein

Amino acid substitutions found in the receptor binding domain of S proteins from SARS-CoV-2 isolated in South America were mapped in the 3D structure of receptor binding domain complexed with its acceptor ACE2, available at the Protein Data Bank (PDB) under accession number 6LZG. Visualization was done using Jmol-14.0.4 software (available at: http://www.jmol.org/).

## 3. Results

### 3.1. Diversification and sharp increase of virus population size

To determine which SARS-CoV-2 genotypes circulate in the South American region, 933 genomes from strains circulating in this region were genotyped. 62 different genotypes were found to circulate in this region (see Fig. 1). Four VOCs circulate in the South American region. VOC Gamma (P.1) was found to circulate in almost all South American countries enrolled in these studies (Fig. 1). VOI Lambda (C.37) was found to circulate in Argentina, Chile and Peru. VOI Zeta (P.2) was found to circulate in Brazil, Ecuador, Paraguay, Suriname and Uruguay. VOI Mu (B.1.621) was found to circulate in Colombia.

Then, in order to gain insight into the mode of evolution of SARS-CoV-2 variants isolated in South America, a first analysis using 145 available and comparable complete genomes sequences from major SARS-CoV-2 genotypes isolated in South America was performed using a Bayesian MCMC approach (Bouckaert et al., 2019; Bouckaert et al., 2019) (for variants included in these analyses see Supplementary Material Table 2). The results shown in Table 1 are the outcome of 40 million steps of the MCMC, using the GTR+Γ+*I* model, a relaxed molecular clock and the Bayesian Skyline model.

The results of these studies suggest that the SARS-CoV-2 variants isolated in the South American region evolved from ancestors that existed around November 26th, 2019. This is in agreement with recent estimations that point to all sequences sharing a common ancestor towards the end of 2019, supporting this as the period when SARS-CoV-2 jumped into its human host (van Dorp, Acman and Richard, 2020; Leung et al., 2021; Leung et al., 2021).

A skyline plot is a graphical representation of historical effective population sizes as a function of time. Past population sizes for these plots are estimated from genetic data, without *a priori* assumptions on the mathematical function defining the shape of the demographic trajectory. Due to these facts, skyline plots can provide realistic descriptions of the complex demographic scenarios that occur in natural populations (Navascués, Leblois and Burgarella, 2017; Navascues et al., 2017). At present, most of the methods to estimate demography from genetic data are based on the coalescent. The coalescent is a mathematical model that describes the rate at which genetic lineages coalesce (i.e., join in a common ancestor) towards the past, forming the genealogy of the sample (Kuhner, 2009; Kuhner, 2009). In order to reconstruct the demographic history of SARS-CoV-2 variants detected in South America enrolled in these studies, a Bayesian skyline plot was constructed (Drummond, Rambaut, Shapiro and Pybus, 2005). The results of these studies are shown in Fig. 2.

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 to 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 relations 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)

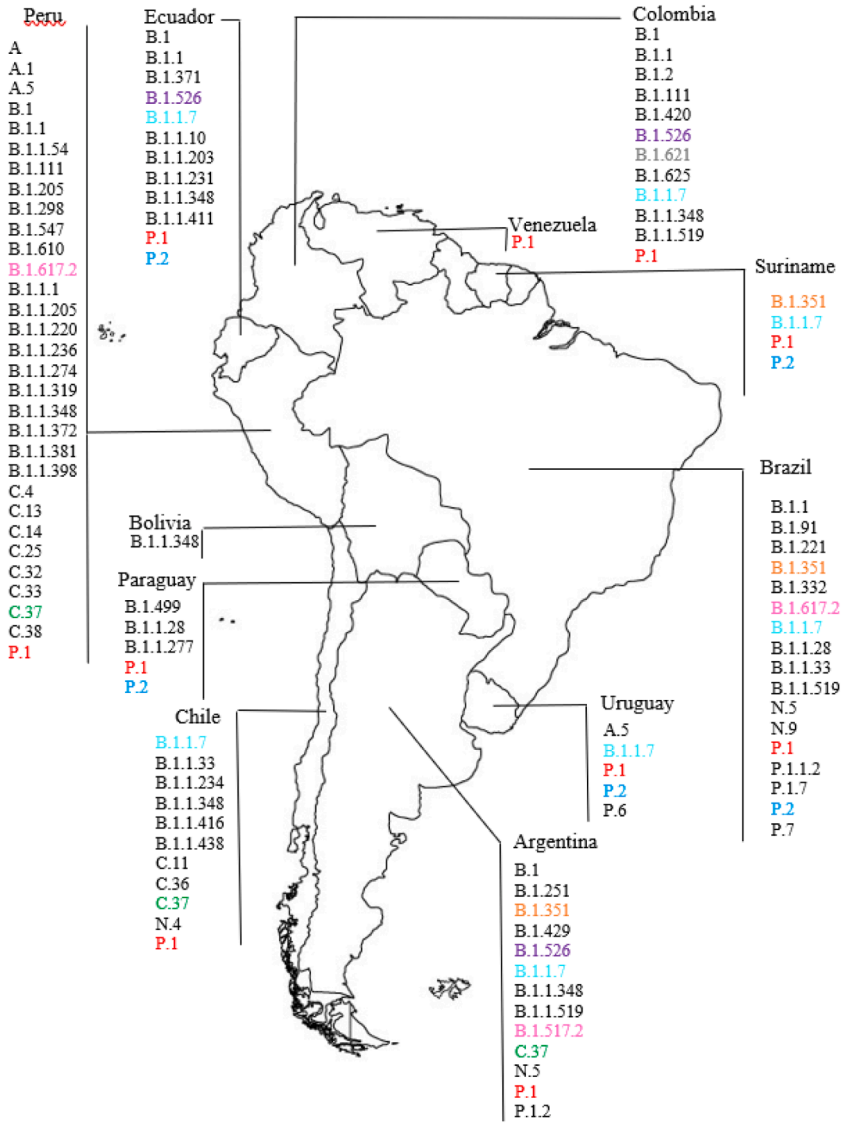


**Fig. 1.** Map of South America. Genotypes isolated in each country are shown under country names. VOCs Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1) and Delta (B.1.617.2) are shown in cyan, orange, red and fuchsia, respectively. VOIs Lambda (C.37), Mu (B.1.621) and Zeta (P.2) are shown in green, gray and blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
Bayesian coalescent inference of SARS-CoV-2 strains isolated in South America.

| Group[a] | Parameter | Value [b] | HPD[c] | ESS[d] |
|---|---|---|---|---|
| SARS-CoV-2 full-length sequences | Tree Likelihood | −52,353.03 | −53,279.00 to −53,228.42 | 1189.15 |
| | tMRCA[e] | 1.564 11/26/2019 | 1.305 to 1.943 | 273.36 |

[a] See Supplementary Material Table 2 for strains included in this analysis.
[b] In all cases, the mean values are shown.
[c] HPD, high probability density values.
[d] ESS, effective sample size.
[e] 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

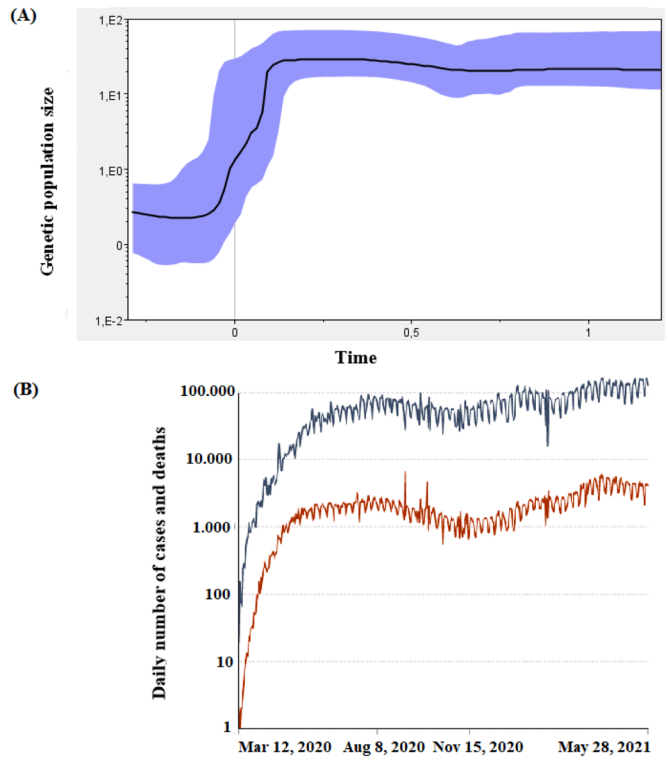


**Fig. 2.** Bayesian Skyline plot depicting the population history of SARS-CoV-2 strains isolated in South America. In (A) a Bayesian Skyline is shown. The thick solid black line represents the median estimate, and the blue area shows the 95% highest probability density (HPD) values (Bouckaert et al., 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.)

acid sequences of the S protein of 933 SARS-CoV-2 variants circulating in the South America region were aligned with corresponding sequences of reference strain Wuhan-Hu-1 (GenBank: NC_045512) and the S protein substitutions observed (see Fig. 3) (for sequences enrolled in these studies, see Supplementary Material Table 2).

From the 933 South American isolates enrolled in these studies 98.82% has a D614G substitution. This substitution is present in all the B.1 lineages established in South America (Martin, VanInsberghe and Koelle, 2021; Martin et al., 2021).

On the other hand, substitutions were found in all S protein domains. 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).

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

In addition to the adaptive S protein substitutions, there are adaptative substitutions in other viral components that also contribute to the spread, fitness and infectivity of the virus. That is the case of the Nucleocapsid protein (N), where recent studies revealed that variants having substitutions R203K and G204R in N protein have replication advantages over the original virus, as well as show increased infectivity in human lung cells, contributing to an increased transmission and virulence of these variants (Wu et al., 2021). For these reasons, the same studies outlined above for S protein were performed for the N proteins of the same 933 strains. The results of these studies are shown in Fig. 5.

The crystal structure of N-protein revealed two distinct domains (Kang, Yang and Hong, 2020; Kang et al., 2020). One domain is present towards the N terminus of the protein and is also known as the RNA-binding domain (RBD). The C terminal side of the protein harbors a dimerization domain, which interacts with other N protein to make a dimer. Apart from these two domains there are three intrinsically disordered regions (IDRs) at N- and C-terminal ends as well as between the RBD and dimerization domain (Kang et al., 2020) (see Fig. 5). We have observed that R203K and G204R are the most frequently mutated residues of the N-protein. 89% of the SARS-CoV-2 variants isolated in South America and enrolled in these studies have these substitutions. Moreover, 70 substitutions were found by comparison with reference strain Wuhan-Hu-1 (see Fig. 5). 61% (43 out of 70) of these substitutions reside in the IDR regions. Recent studies have shown that positions S197, S202, R203 and G204 are important sites of phosphorylation by Aurora kinase A/B, GSK-3 as well as for its interactions with 14–3–3 protein (Tung and Limtung, 2020) (Tung and Limtung, 2020). Substitutions were found in two of these four positions. Interestingly, substitution S194L was observed in strains isolated in Peru and Colombia (see Fig. 5). Recent studies revealed an association of strains having this substitution and symptomatic patients (Barona-Gomez, Delaye and Diaz-Valenzuela, 2021; Barona-Gomez et al., 2021). Another recent study revealed that one of the most important B and T cell epitope of the N protein lies between residues 305–340 (Ahmed, Quadeer and McKay, 2020). Our study identified four substitutions in those positions (Fig. 5).

### *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 (Ilmjärv, Abdul and Acosta-Gutiérrez, 2021; Ilmjärv 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 domain (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

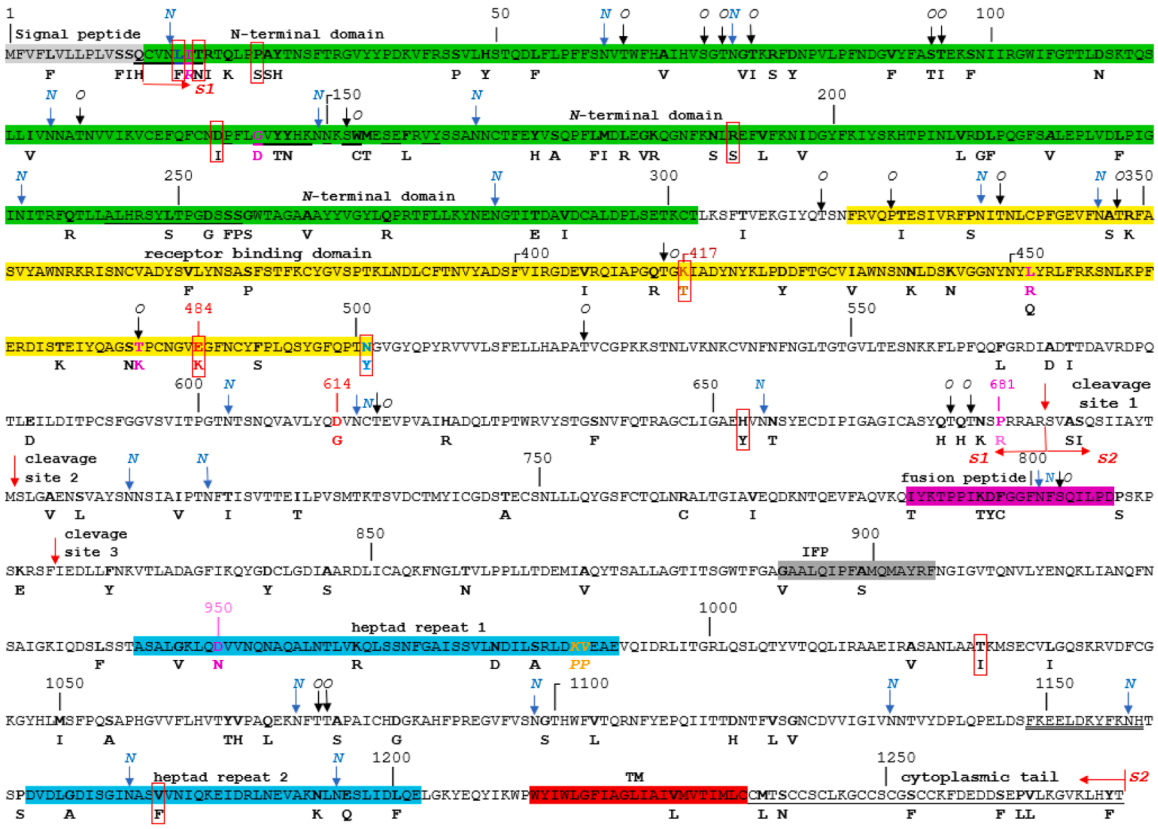


**Fig. 3.** Substitutions found in the Spike protein of SARS-CoV-2 viruses circulating in South America. The amino acid sequence of reference strain Wuhan-Hu-1 strain (GenBank: NC_045512) is shown. Domains of the Spike protein are shown at the top of the amino acid sequence. IFP, internal fusion peptide; TM, transmembrane domain (Xia, 2021). The cytoplasmic tail is shown underlined. Boundaries of S1 and S2 sub-units are shown in red and italics. Numbers at the top of the sequence denote amino acid position (relative to the reference strain). Substitution D614G, present in all strains B.1 strains and substitution E484K, present in all VOI Zeta (P.2) as well of VOCs Beta and Gamma are shown in red. Substitution N501Y, found in all VOC Gama (P.1) strains, is shown in bold blue. Substitution K417T, characteristic of VOC Beta (B.1.351) is shown in brown. Substitutions L452R,T478K, P681R and D950N characteristic of VOC Delta (B.1.1617.2) are shown in fuchsia. Positions where substitutions were found in strains isolated in South America are highlighted in bold and the corresponding substitution is shown below each position. Proline substitutions in Pfizer/BioTech and Moderna vaccine to stabilize the Spike protein at the prefusion conformation are shown in orange (Xia, 2021; Xia, 2021). Positions in the antigenic supersite of the *N*-terminal domain (NTD) are underlined. The *N*-glycosylation sites are shown by a blue arrow while the *O*-glycosylation sites are shown by a black arrow and they are indicated *N* or *O* next to the arrow. Sites where significant polymorphic sites were found are shown in red squares. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

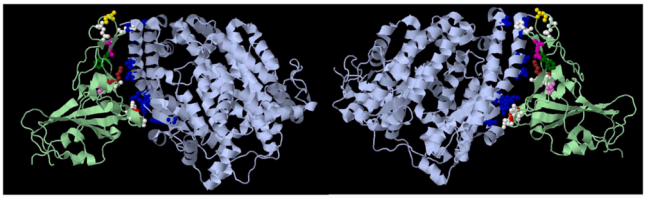

**Fig. 4.** 3D structure of Spike protein receptor-binding domain complexed with its receptor ACE2. Receptor-binding domain of Spike protein is shown in light green, whereas ACE2 protein is shown in gray. ACE2 amino acids interacting with Spike receptor-binding domain (Gln24, Asp30, His34, Tyr41, Gln42, Met82, Lys353 and Arg357) (according to Yi et al., 2020) are shown in blue. Substitutions K417N, L452R, T478K, E484K, F490S and N501Y are shown in brown, rose, yellow, fuchsia, green and red, respectively. Positions of other amino acids in the Spike receptor-binding domain also shown to interact with ACE2 (Gln498, Thr500, Tyr453, Gln474 and Phe 486) (also according to Yan et al., 2020) are shown in white. The 3D structure was obtained from the PDB Database under accession number 6LZG. Two views rotated 180° in the X-axis are shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 epidemiological highly successful and replaced the original D614/P323 variants (Ilmjärv, Abdul and Acosta-Gutiérrez, 2021; Ilmjärv 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) no correlated among themselves. The

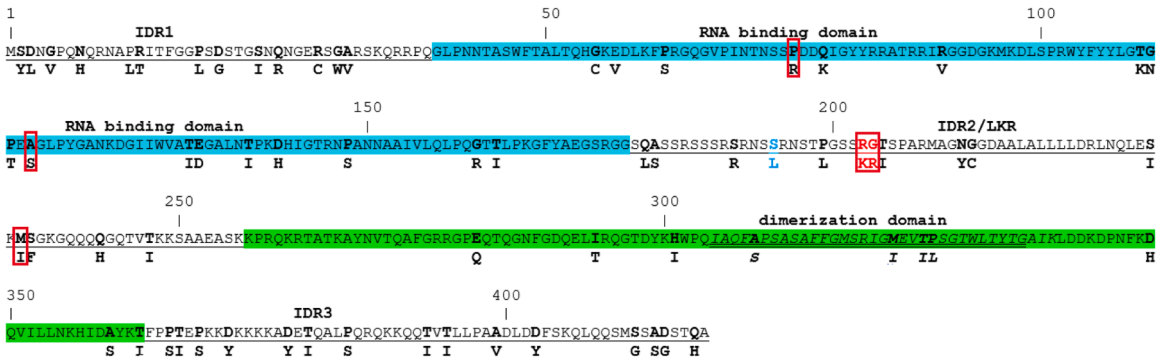


**Fig. 5.** Substitutions found in the Nucleocapsid (N) protein of SARS-CoV-2 viruses circulating in South America. The amino acid sequence of reference strain Wuhan-Hu-1 strain (GenBank: NC_045512) is shown. Domains of the N protein are shown at the top of the amino acid sequence. Intrinsically disordered regions (IDR) are shown underlined. LKR, liker region. Residues conforming B/T cell epitope are shown in italics and double underlined. Substitutions R203K and G204R, found in 89% of the strains enrolled in these studies are shown in red. Substitution S194S is shown in blue. Numbers at the top of the sequence denote amino acid position (relative to the reference strain). Sites where significant polymorphic sites were found are shown in red squares.

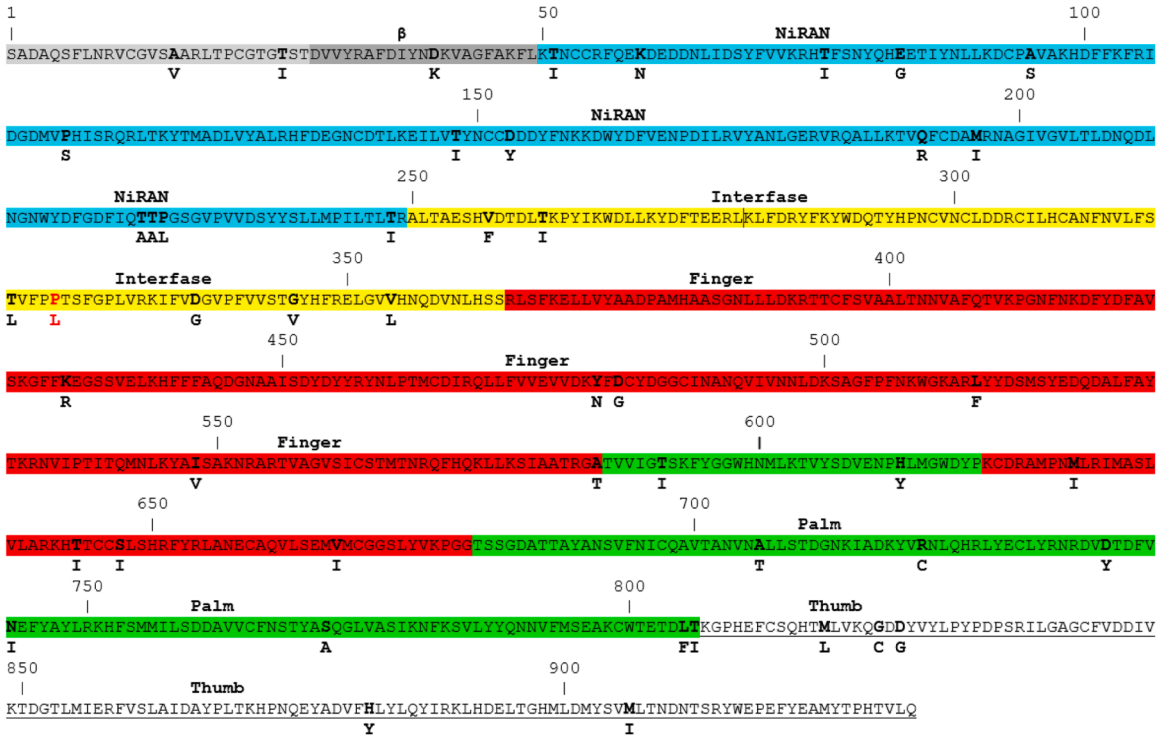


**Fig. 6.** Substitutions found in the RNA-dependent RNA polymerase (RdRp) protein of SARS-CoV-2 viruses circulating in South America. The amino acid sequence of reference strain Wuhan-Hu-1 strain (GenBank: NC_045512) is shown. Domains of the RdRp protein are shown at the top of the amino acid sequence. Numbers at the top of the sequence denote amino acid position (relative to the reference strain). Substitution P323L, found in 98% of the strains enrolled in these studies is shown in red. NiRAN domain is shown in cyan. Interfase domain is shown in yellow. RdRp domains Finger, Palm and Thumb are shown in red, blue and underlined, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 GCG, AGC 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, Gammaa 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

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

| 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 | UGA | *** |
| | UUG | 1.12 | | UCG | 0.12 | | UAG | *** | Trp | UGG | 1.00 |
| | CUU | 1.98 | Pro | CCU | 1.99 | His | CAU | 1.51 | Arg | CGU | 1.28 |
| | CUC | 0.67 | | CCC | 0.28 | | CAC | 0.49 | | CGC | 0.15 |
| | CUA | 0.50 | | CCA | 1.73 | Gln | CAA | 1.49 | | CGA | 0.00 |
| | CUG | 0.17 | | CCG | 0.00 | | CAG | 0.51 | | CGG | 0.29 |
| Ile | AUU | 1.74 | Thr | ACU | 1.81 | Asn | AAU | 1.22 | Ser | AGU | 1.04 |
| | AUC | 0.55 | | ACC | 0.40 | | AAC | 0.78 | | AGC | 0.30 |
| | AUA | 0.71 | | ACA | 1.65 | Lys | AAA | 1.26 | Arg | AGA | 2.87 |
| Met | AUG | 1.00 | | ACG | 0.13 | | AAG | 0.74 | | AGG | 1.41 |
| Val | GUU | 1.97 | Ala | GCU | 2.12 | Asp | GAU | 1.37 | Gly | GGU | 2.31 |
| | GUC | 0.87 | | GCC | 0.41 | | GAC | 0.63 | | GGC | 0.72 |
| | GUA | 0.62 | | GCA | 1.37 | Glu | GAA | 1.41 | | GGA | 0.82 |
| | GUG | 0.54 | | GCG | 0.10 | | GAG | 0.59 | | GGG | 0.15 |

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

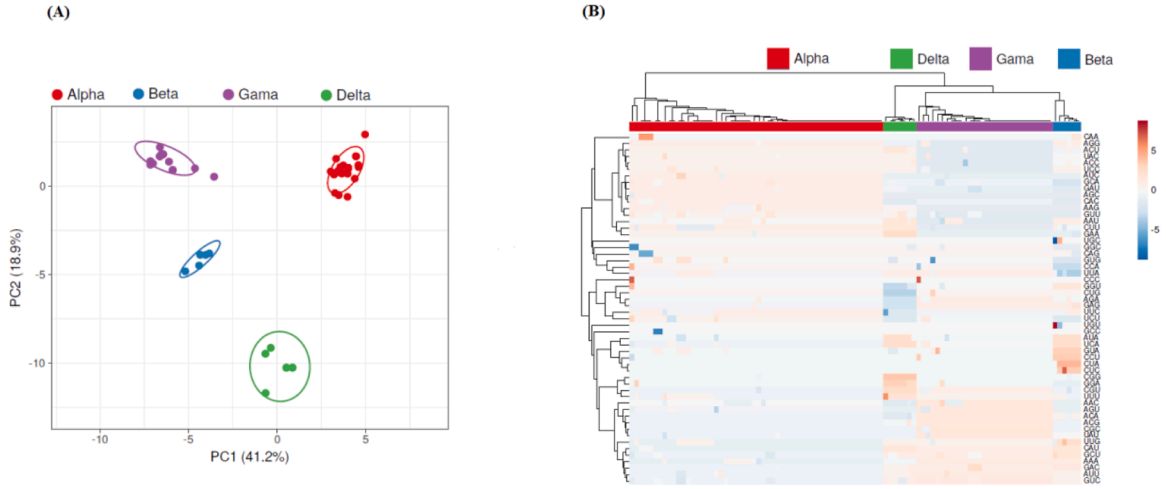


**Fig. 7.** PCA of codon usage in Spike proteins from VOC's SARS-CoV-2 strains. In (A) the position of the Spike proteins in the plane conformed by the first two major components of PCA is shown. SVD was used to calculate principal components and unit variance was applied. The proportion of variance explained by each axis is shown between parentheses. Prediction ellipses are such that with probability 0.95, a new observation from the same group will fall inside the ellipse. Genotypes are indicated at the top of the figure. *N*= 86 data points. In (B) Heatmaps of codon usage in Spike proteins are shown. Unit variance scaling was applied. Each column corresponds to a different Spike protein from SARS-CoV-2 VOCs strains, who's genotype is shown at the top of the figure. Both rows and columns are clustered using correlation distance and average linkage.

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. 7A 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). These results suggest codon usage plays an important role in shaping the evolution of Spike proteins from SARS-CoV-2 VOC's. Moreover, when codon usage frequencies are analyzed using 494 S genes from 10 different genotypes found to circulate in South America, a relation among genotype and codon usage is also found (see Supplementary Material Fig. 6).

## 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). Phylodynamic 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 who 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 et al., 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 T95I, 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 American 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. 3). 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 Rodríguez-Gómez, 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; Berkout 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. **Alicia Costábile:** Data curation, Visualization, Investigation. **Mercedes Paz:** 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 Grupos $I+$ 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.

## References

Afgan, E., Baker, D., Batut, 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 affinity to ACE2 due to N501Y mutant. Med. Drug Discov. 10, 100086 https://doi.org/10.1016/j.medidd.2021.100086.

Ahmed, S.F., Quadeer, A.A., McKay, M.R., 2020. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses 12, 254. https://doi.org/10.3390/v12030254.

Barona-Gomez, F., Delaye, L., Diaz-Valenzuela, E., et al., 2021. Phylogenomics and population genomics of SARS-CoV-2 in Mexico during the pre-vaccination stage reveals variants of interest B.1.1.28.4 and B.1.1.222 or B.1.1.519 and the nucleocapsid mutation S194L associated with symptoms. Microb Genom 7, 000684. https://doi.org/10.1099/mgen.0.000684.

Belouzard, S., Chu, V.C., Whittaker, G.R., 2009. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. Proc. Natl. Acad. Sci. USA 106, 5871–5876. https://doi.org/10.1073/pnas.0809524106.

Berkhout, B., van Hemert, F., 2015. On the biased nucleotide composition of the human coronavirus RNA genome. Virus Res 202, 41–47. https://doi.org/10.1016/j.virusres.2014.11.031.

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al., 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 15, e1006650 https://doi.org/10.1371/journal.pcbi.1006650.

Candido, D.S., Claro, I.M., de Jesus, J.G., et al., 2020. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science 369, 1255–1260. https://doi.org/10.1126/science.abd2161.

Chakraborty, C., Bhattacharya, M., Sharma, A.R., 2021 Jun 27. Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: their significant mutations in S-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity. Rev Med Virol e2270. https://doi.org/10.1002/rmv.2270.

Chan, J.F., Kok, K.H., Zhu, Z., et al., 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes Infec. 9, 21–236. https://doi.org/10.1080/22221751.2020.1719902.

Cerutti, G., Guo, Y., Zhou, T., et al., 2021. Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. Cell Host Microbe 29 (5), 819–833.e7. https://doi.org/10.1016/j.chom.2021.03.005.

Chen, Y., Liu, Q., Guo, D., 2020. Emerging coronaviruses: genome structure, replication, and pathogenesis. J. Med. Virol. 92, 418–423. https://doi.org/10.1002/jmv.25681.

Collier, D.A., De Marco, A., Ferreira, I.A., Meng, B., Datir, R.P., Walls, A.C., Kemp, S.A., et al., 2021. Sensitivity of SARS-CoV-2 B1.1.7 to mRNA vaccine-elicited antibodies. Nature 593, 136–141. https://doi.org/10.1038/s41586-021-03412-7.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Edmunds, W.J., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372. https://doi.org/10.1126/science.abg3055.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis. 20 (5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.

Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22, 1185–1192. https://doi.org/10.1093/molbev/msi103.

Faria, N.R., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D., Mishra, S., Crispim, M. A., Sales, F.C., Hawryluk, I., McCrone, J.T., et al., 2021. Genomics and epidemiology of a novel SARS-CoV-2 lineage in manaus, brazil. Science 372, 815–821. https://doi.org/10.1126/science.abh2644.

Flores-Alanis, A., Cruz-Rangel, A., Rodríguez-Gómez, F., et al., 2021. Molecular epidemiology surveillance of SARS-CoV-2: mutations and genetic diversity one year after emerging. Pathogens 10, 184. https://doi.org/10.3390/pathogens10020184.

Gao, Y., Yan, L., Huang, Y., et al., 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 368, 779–782. https://doi.org/10.1126/science.abb7498.

Gorbalenya, A.E., Baker, S.C., Baric, R.S., et al., 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. 5, 536–544. https://doi.org/10.1038/s41564-020-0695-z.

Greenacre, M., 1994. Theory and Applications of Correspondence Analysis. Academic Press, London.

Gupta, R., Brunak, S., 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. Pac. Symp. Biocomput. 310–322. https://services.healthtech.dtu.dk/service.php?NetNGlyc-1.0.

Hou, Y.X., Chiba, S., Halfmann, P., et al., 2020. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science 370, 1464–1468. https://doi.org/10.1126/science.abe8499.

Ilmjärv, S., Abdul, F., Acosta-Gutiérrez, S., et al., 2021. Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant. Sci Rep. 11, 13705. https://doi.org/10.1038/s41598-021-91662-w.

Kang, S., Yang, M., Hong, Z., et al., 2020. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharmaceutica Sinica B 10, 1228–1238. https://doi.org/10.1016/j.apsb.2020.04.009.

Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinformatics 4, 1160–1166. https://doi.org/10.1093/bib/bbx108.

Korber, B., Fischer, W.M., Gnanakaran, S., et al., 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 182, 812–827. https://doi.org/10.1016/j.cell.2020.06.043.

Kuhner, M.K., 2009. Coalescent genealogy samplers: windows into population history. Trends Ecol Evol 24, 86–93. https://doi.org/10.1016/j.tree.2008.09.007.

Leach, A., Ilca, F.T., Akbar, Z., Ferrari, M., Bentley, E.M., Mattiuzzo, G., Onuoha, S., Miller, A., Ali, H., Rabbitts, T.H., 2021. A tetrameric ACE2 protein broadly neutralizes SARS-CoV-2 spike variants of concern with elevated potency. Antiviral Res 194, 105147. https://doi.org/10.1016/j.antiviral.2021.105147.

Leung, K.S., Ng, T.T., Wu, A.K., Yau, M.C., Lao, H.Y., et al., 2021. Territory wide study of early coronavirus disease outbreak, Hong Kong, China. Emerg. Infect. Dis. 27, 196–204. https://doi.org/10.3201/eid2701.201543.

Li, Q., Guan, X., Wu, P., et al., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N. Engl. J. Med. 26, 1199–1207. https://doi.org/10.1056/NEJMoa2001316.

Lo Presti, A., Rezza, G., Stefanelli, P., 2020. Selective pressure on SARS-CoV-2 protein coding genes and glycosylation site prediction. Heliyon. 6, e05001. https://doi.org/10.1016/j.heliyon.2020.e05001.

Luan, B., Wang, H., Huynh, T., 2021. Enhanced binding of the N501Y-mutated SARSCoV-2 spike protein to the human ACE2 receptor: insights from molecular dynamics simulations. FEBS Lett. https://doi.org/10.1002/1873-3468.14076.

Martin, M.A., VanInsberghe, D., Koelle, K., 2021. Insights from SARS-CoV-2 sequences. Science 371, 466–467. https://doi.org/10.1126/science.abf3995.
McCallum, M., De Marco, A., Lempp, F.A., et al., 2021a. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. Cell 184, 2332–2347.e16. https://doi.org/10.1016/j.cell.2021.03.028.
McCallum, M., Bassi, J., De A., Marco, et al., 2021b. SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. Science 373, 648–654. https://doi.org/10.1126/science.abi7994.
Metsalu, T., Vilo, J., 2015. Clustvis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res. 43 (W1), W566–W570. https://doi.org/10.1093/nar/gkv468.
Navascués, M., Leblois, R., Burgarella, C., 2017. Demographic inference through approximate-Bayesian-computation skyline plots. Peer J 5, e3530. https://doi.org/10.7717/peerj.3530.
Omotuyi, I.O., Nash, O., Ajiboye, O.B., Iwegbulam, C.G., Oyinloye, E.B., Oyedeji, O.A., Kashim, Z.A., Okaiyeto, K., 2021. Atomistic simulation reveals structural mechanisms underlying D614G spike glycoprotein-enhanced fitness in SARS-COV-2. J. Comput. Chem. 41, 2158–2161. https://doi.org/10.1002/jcc.26383, 2020.
Pekar, J., Worobey, M., Moshiri, N., Scheffler, K., Wertheim, J.O., 2021. Timing the SARS-CoV-2 index case in Hubei province. Science 372, 412–417. https://doi.org/10.1126/science.abf8003.
Piccoli, L., Park, Y.J., Tortorici, M.A., et al., 2020. Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. Cell 183, 1024–1042.e21. https://doi.org/10.1016/j.cell.2020.09.037.
Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5.
Samavati, L., Uhal, B.D., 2020. ACE2, much more than just a receptor for SARS-COV-2. Front. Cell. Infect. Microbiol. 10, 317. https://doi.org/10.3389/fcimb.2020.00317.
Salleh, M.Z., Derrick, J.P., Deris, Z.Z., 2021. Structural evaluation of the spike glycoprotein variants on SARS-CoV-2 transmission and immune evasion. Int. J. Mol. Sci. 22, 7425. https://doi.org/10.3390/ijms22147425.
Shajahan, A., Supekar, N.T., Gleinich, A.S., Azadi, P., 2020. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. Glycobiology 30, 981–988. https://doi.org/10.1093/glycob/cwaa042.
Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38. https://doi.org/10.1007/BF02099948.
Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., et al., 2013. Precision mapping of the human O-GalNAc glycoproteome through Simple Cell technology. EMBO J 32, 1478–1488. https://doi.org/10.1038/emboj.2013.79.
Trifinopoulos, J., Nguyen, L.T., von Haeseler, A., Minh, B.Q., 2016. Nucl. Acids Res. 44 (W1), W232–W235. https://doi.org/10.1093/nar/gkw256.
Thomsen, M.C.F., Nielsen, M., 2012. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 40, W281–W287. https://doi.org/10.1093/nar/gks469.
Thomson, E.C., Rosen, L.E., Shepherd, J.G., Spreafico, R., da Silva, F., Wojcechowskyj, J. A., Davis, C., et al., 2021. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell 184, 1171–1187.e20. https://doi.org/10.1016/j.cell.2021.01.037.
Tung, H.Y.L., Limtung, P., 2020. Mutations in the phosphorylation sites of SARS-CoV-2 encoded nucleocapsid protein and structure model of sequestration by protein 14-3-3. Biochem. Biophys. Res. Commun. 532, 134–138. https://doi.org/10.1016/j.bbrc.2020.08.024.
Uriu, K., Kimura, I., Shirakawam, K., et al., 2021. Neutralization of the SARS-CoV-2 Mu Variant by Convalescent and Vaccine Serum. N. Engl. J. Med. 385, 2397–2399. https://doi.org/10.1056/NEJMc2114706.
van Dorp, L., Acman, M., Richard, D., et al., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 283, 104351 https://doi.org/10.1016/j.meegid.2020.104351.
Volz, E., Hill, V., McCrone, J.T., et al., 2021. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell 1, 64–75. https://doi.org/10.1016/j.cell.2020.11.020.
Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., Veesler, D., 2020. Structure, function, and antigenicity of the SARS-CoV-2 Spike glycoprotein. Cell 181, 281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058.
Walls, A.C., Tortorici, M.A., Frenz, B., Snijder, J., Li, W., Rey, F.A., DiMaio, F., Bosch, B. J., Veesler, D., 2016. Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy. Nat. Struct. Mol. Biol. 23, 899–905. https://doi.org/10.1038/nsmb.3293.
Wang, L., Cheng, G., 2021. Sequence analysis of the Emerging Sars-CoV-2 Variant Omicron in South Africa. J. Med. Virol. https://doi.org/10.1002/jmv.27516, 2021 Dec 12.
Wang, Z., Schmidt, F., Weisblum, Y., Muecksch, F., Barnes, C.O., Finkin, S., Nussenzweig, M.C., 2021. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. Nature. https://doi.org/10.1038/s41586-021-03324-6.
Weinreich, D.M., Sivapalasingam, S., Norton, T., Ali, S., Gao, H., Bhore, R., Trial, I., 2021. REGN-COV2, a neutralizing antibody cocktail, in outpatients with covid-19. N. Engl. J. Med. https://doi.org/10.1056/NEJMoa2035002.
Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J.C., Muecksch, F., et al., 2020. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. Elife. 9, e61312. https://doi.org/10.7554/eLife.61312.
Winger, A., Caspari, T., 2021. The Spike of Concern-The Novel Variants of SARS-CoV-2. Viruses 13, 1002. https://doi.org/10.3390/v13061002.
World Health Organization. 2020a. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). https://www.who.int (accessed August12th, 2021).
World Health Organization. 2020b. Coronavirus disease 2019 (COVID-19) Weekly epidemiological update on COVID-19 - 14 December 2021). https://www-who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed December 14th, 2021).
Wu, H., Xing, N., Meng, K., et al., 2021. Nucleocapsid mutation R203K/G294R increase the infectivity, fitness, and virulence of SARS-CoV-2. Cell Host Microbe 29, 1788–1801. https://doi.org/10.1016/j.chom.2021.11.005.
Xia, X., 2021. Domains and functions of Spike protein in SARS-Cov-2 in the context of vaccine design. viruses 13, 109. https://doi.org/10.3390/v13010109/v13010109.
Hoffmann, M., Kleine-Weber, H., Pöhlmann, S., 2020. A multibasic cleavage site in the Spike protein of SARS-CoV-2 Is essential for infection of human lung cells. Mol. Cell. 78, 779–784.e5. https://doi.org/10.1016/j.molcel.2020.04.022.
Yi, C., Sun, X., JYe, J.1., Ding, L., Liu, M., et al., 2020. Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. Cell. Mol. Immunol. 17, 621–630. https://doi.org/10.1038/s41423-020-0458-z.
Zhang, J., Cai, Y., Xiao, T., Lu, J., Peng, H., Sterling, S.M., Walsh, R.M., Rits-Volloch, S., Zhu, H., Woosley, A.N., et al., 2021. Structural impact on SARS-CoV-2 spike protein by D614G substitution. Science 372, 525–530. https://doi.org/10.1126/science.abf2303, 2021.