Genomic characterization and molecular evolution of SARS-CoV-2 in Rio Grande do Sul State, Brazil

Amanda de Menezes Mayer[1], Patrícia Aline Gröhs Ferrareze[2], Luiz Felipe Valter de Oliveira[3], Tatiana Schäffer Gregianini[4], Carla Lucia Andretta Moreira Neves[2], Gabriel Dickin Caldana[2], Lívia Kmetzsch[1], Claudia Elizabeth Thompson[1, 2, 5*]

[1] Center of Biotechnology, Graduate Program in Cell and Molecular Biology (PPGBCM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

[2] Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil

[3] BiomeHub, Florianópolis, Brazil.

[4] Laboratório Central de Saúde Pública do Centro Estadual de Vigilância em Saúde da Secretaria de Saúde do Estado do Rio Grande do Sul (LACEN/CEVS/SES-RS), Porto Alegre, RS, Brazil

[5] Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, Brazil

*** Corresponding author**

Claudia Elizabeth Thompson

Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), 245/200C Sarmento Leite St, Porto Alegre, RS, Brazil. ZIP code: 90050-170. Phone: +55 (51) 3303 8889. E-mail: cthompson@ufcspa.edu.br, thompson.ufcspa@gmail.com

Running title: SARS-CoV-2 Genomics in Rio Grande do Sul State

Keywords: SARS-CoV-2, Genomics, COVID-19, molecular evolution, phylogenomics

Abstract

The SARS-CoV-2 is the virus responsible for the COVID-19 pandemic and is plaguing the world since the end of 2019. Different lineages have been discovered ever since and the Gamma lineage, which started the second wave of infections, was first described in Brazil, one of the most affected countries by pandemic. Describing the viral genome and how the virus behaves is essential to contain its propagation and to the development of medications and vaccines. Therefore, this study analyzed SARS-CoV-2 sequenced genomes from Esteio city in Rio Grande do Sul, Southern Brazil. We also comparatively analyzed genomes of the two first years of the pandemic from Rio Grande do Sul state for understanding their genomic and evolutionary patterns. The phylogenomic analysis showed monophyletic groups for Alpha, Gamma, Delta and Omicron, as well as for other circulating lineages in the state. Molecular evolutionary analysis identified several sites under adaptive selection in membrane and nucleocapsid proteins which could be related to a prevalent stabilizing effect on membrane protein structure, as well as majoritarily destabilizing effects on C-terminal nucleocapsid domain.

1 INTRODUCTION

After the first outbreak of COVID-19 (Coronavirus disease 2019) in Wuhan, Hubei Province, China in December 2019 (Zhu et al., 2020), the new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spread around the world. Starting a world pandemic, declared by the World Health Organization (WHO) in March 2020, COVID-19 already exceeds 649 million cases and 6.65 million deaths until December, 2022 (Dong et al., 2020, https://coronavirus.jhu.edu/map.html accessed on December 13, 2022). Currently (December, 2022), Brazil is the fifth country most

affected by SARS-CoV-2, reaching the mark of 35 million confirmed cases and more than 690 thousands deaths (Dong et al., 2020, https://coronavirus.jhu.edu/map.html accessed on December 13, 2022). Of these, 7.9% of the cases and 6% of the deaths are from Rio Grande do Sul (RS), the southernmost state of Brazilian territory and the fourth state in the ranking of COVID-19 cases (https://covid.saude.gov.br/, accessed on December 13, 2022).

Among several lineages, along these two years of pandemic, various Variants of Concern (VOCs), such as Alpha, Beta, Gamma, Delta, and Omicron, carrying signature aminoacid substitutions (especially in the spike protein) circulated in RS state (Gularte et al., 2022; Wink et al., 2022; Gräf et al., 2022). The sublineage P.1.2, a Gamma-like variant also considered as a Variant of Interest, was initially identified in RS state (Franceschi et al., 2021), despite other studies estimating its divergence between late 2020 and early 2021 in São Paulo state (Junior et al., 2021). As in Brazil, an increasing number of cases and deaths by Gamma (P.1 lineage) became evident in RS in the beginning of 2021, causing a second COVID-19 wave. The P.1 lineage harbors mutations in the Spike's receptor-binding domain (RBD) such as E484K, K417T, and N501Y, which promote evasion from antibody neutralization elicited by infection or vaccination (R. E. Chen et al., 2021; Chakraborty, 2022).

In December, 2020, the Delta variant was described initially in India. This variant is highly transmissible and spreads easily, causing new waves of infection around the world by the middle 2021 (Shiehzadegan et al., 2021). However, despite this lineage rapidly becoming dominant in Brazil (including RS state), in July/August, 2020, there was not a concurrent increase in reported cases or deaths (Giovanetti et al., 2022). By the end of 2021, the number of cases of COVID-19 were progressively decreasing, until the emergence of the Omicron variant, confirmed in November

2021, in South Africa (Wang & Powell, 2021). This new variant is a concern due to the mutations on RBD and cleavage sites that suggest higher transmissibility, up to three times more contagious than Delta variant (J. Chen et al., 2022).

Despite all VOC characterizations mostly focused on spike mutations, structural proteins E (Envelope), M (Membrane), and N (Nucleocapsid) are functionally important to the virus assembly and pathogenesis (Yadav et al., 2021). The N protein has been associated to the promotion of inflammatory processes by activation of cyclooxygenase-2 (COX-2), to interaction with p42 proteasome in order to avoid the degradation of viral proteins, and to inhibition of IFN-I in immune response (Satarker & Nampoothiri, 2020). Moreover, M protein is known to inhibit NFκB, to reduce levels of COX-2, to activate IFN-β, and to interact with PDK1/PKB proteins, leading to cell death or apoptosis.

In this way, this study aims to perform genomic sequencing and characterization of the SARS-CoV-2 genomes from RS state as well as to identify selection traits in E, M, and N protein sites, elucidating the molecular evolution processes that drive the diversification or conservation of the structural proteins from SARS-CoV-2.

2 RESULTS

2.1 Sample Characterization

Twelve respiratory secretion samples used for COVID-19 diagnostic purposes (RT-qPCR) were collected from patients from the municipality of Esteio, Rio Grande do Sul (RS) state, between April 9th and June 29th, 2021. The mean cycle threshold (Ct) value for the first RT-qPCR conducted at Laboratório Central de Saúde Pública do Estado do Rio Grande do Sul (LACEN) was 23.83 (median: 23.00; IQR: 4.5).

The sequence coverage for the twelve sequenced genomes ranged between 84.92 and 99.76% (mean: 98.26%) of the 29,903 bp of NC_045512.2 reference genome. The mean of sequencing depth was calculated to 292.23x, with a variation between 53.15 and 542.28x. Leastwise 48.81% of the sequence accomplished a coverage depth ≥51x (max: 98.27%, mean: 87.68%) (Supplementary File 1). According to Pango lineage assignment, 10 sequenced samples belong to P.1 lineage and 2 are from P.1.17 sublineage, both from Gamma variant clade.

2.2 Comparative Genomics of Esteio SARS-CoV-2 sequences

Sixty-eight different nucleotide substitutions were found in the twelve genomes from this study, being 36 of these non-synonymous. The mutations NSP13:E341D (ORF1b:E1264D), S:D614G and S:V1176F were predominantly found in these sequences. The most common missense substitutions (in at least 75% of the sequences) were NSP12:P323L (ORF1b:P314L), S:K417T, S:T1027I, ORF8:E92K, and N:P80R (in 11 genomes); 5'UTR:C241T and NSP3:K977Q (10 genomes) and H655Y (in 9 genomes). Only two samples carry substitutions S:E484K and S:N501Y, being both mutations in the same samples. The mutations S:E661D and S:S689I were present in one sample each.

A few mutations found in our samples are described for the first time in RS (Table 1). Most of them were already identified in the world and occurred in sequences from different VOCs such as Alpha, Beta, Gamma, Delta, and Omicron. Substitution D1208A (NSP3) was also found for the first time in Brazil. The non-synonymous mutation I136V, in the NSP12 gene, was not described on GISAID, being the first report about this mutation (Table 1).

TABLE 1 List of mutations firstly described in Rio Grande do Sul in the sequenced genomes from this study.

Mutation	Occurrences in the world	First occurrence in world	Occurrences in Brazil	First occurrence in Brazil	Date of our sample	VOCs
ORF3a: A72S	7,462	2020	85	2021-02-03	2021-05-01	A, B, G, D, O
NSP3: P1103L	17,219	2020	428	2020-06-16	2021-06-14	A, B, G, D, O
NSP3: D1208A	271	2020-03-13	not described	-	2021-06-14	A, G, D, O
NSP4: D144G	198	2020	2	2021-05-31	2021-06-14	A, G, D, O
NSP13: L325F	1,255	2020	15	2021-01-27	2021-06-14	A, B, G, D, O
NSP12: I136V	not described	-	not described	-	2021-05-08	G

The table describes the mutations and the number of occurrences in the world and Brazil, as well as their first occurrences in these locations. The table also indicates the VOC lineages where each mutation can be found (Access on GISAID: June 7, 2022). A: Alpha; B: Beta; G: Gamma; D: Delta; O: Omicron.

2.3 Comparative Genomics of Rio Grande do Sul SARS-CoV-2 sequences

A number of 4,706 sequences were downloaded from the GISAID platform ranging from March, 2020 up to May, 2022 (26 months) (Submission up to September 30, 2022). The genome sequencing count per month in RS state can be visualized in Figure 1. Most sequencing efforts are associated with COVID-19 waves

caused by the introduction of new lineages in RS state. This can be observed during Gamma / Delta waves (February up to October, 2021) and, recently, the Omicron wave, started in January, 2022. The total number of sequences in the state is low compared to the number of cases, representing around 0.19% of them (number of cases = 2,435,883 on May 31, 2022).

As shown in Figure 2, lineage B.1.1.33 predominates in 2020. Between November 2020 and January 2021, the Zeta (P.2) lineage became more frequent, followed by B.1.1.28 and P.7 lineages. From February until July 2021, P.1 and derivative lineages (mostly P.1 and P.1.2) became prevalent in the RS genomes, in accordance with the lineages of our samples, which were collected between April and June 2021, during the Gamma-related second wave of COVID-19 in the state. Approximately 27.6% of these SARS-CoV-2 genomes obtained on the GISAID database between March 2020 and May 2022 belong to the Gamma clade. Considering only 2021, 57.1% of the SARS-CoV-2 sequenced genomes were classified as Gamma. The Delta lineages (mostly AY.99.2 and AY.101) were initially identified in the state by sequencing in June 2021 and became prevalent from August up to the end of the year. Delta lineages were accountable for 32.30% of the sequenced genomes from RS state in 2021. The Omicron lineages (mostly BA.1.1 and BA.2) arised in RS state in December, 2021, achieving more than 90% of sequenced genomes between January and April, 2022.

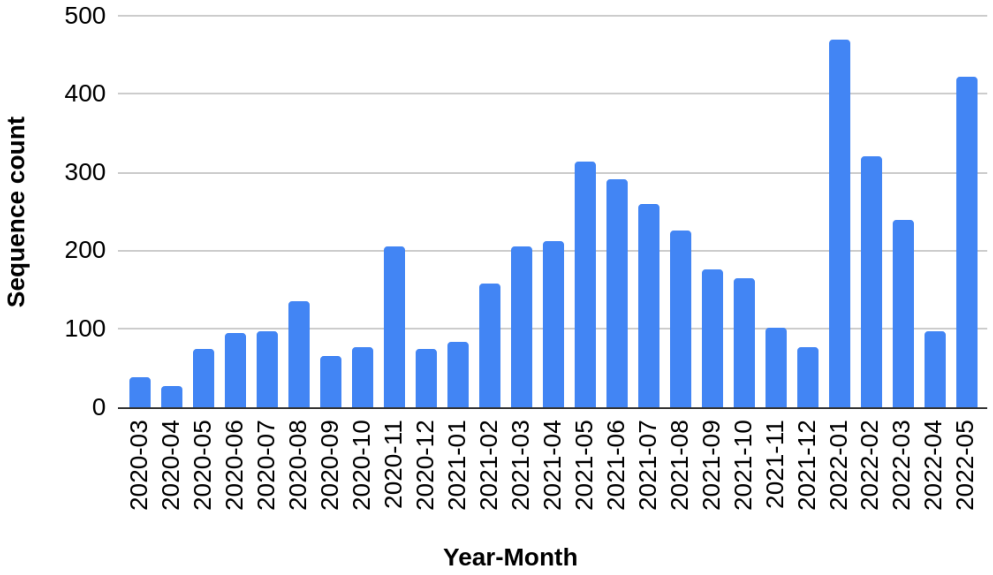

FIGURE 1 Counts of sequences per month in Rio Grande do Sul between March 2020 and May 2022.

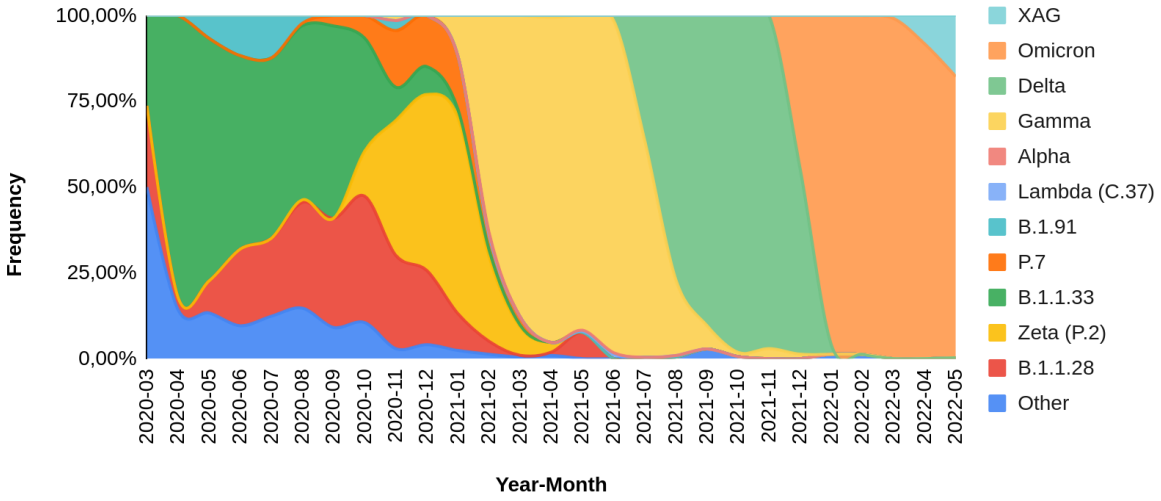

FIGURE 2 Lineage distribution in Rio Grande do Sul state between March 2020 and May 2022.

About the genome set from RS state, the missense variant NSP12:P323L (ORF1b:P314L) was the most prevalent, found in 98.70% (n = 4,645) of samples (Figure 3). The highly frequent mutations (present in at least 90% of the genomes) also include the extragenic substitution C241T (n = 4,484), the synonymous mutation NSP3:F106F (n = 4,509), and the non-synonymous mutation S:D614G (n = 4,504). Other non-synonymous mutations such as N:R203K/G204R (n = 3,856/3,854), S:N501Y (n = 2,762), S:H655Y (n = 2,842) were found in > 50% of samples.

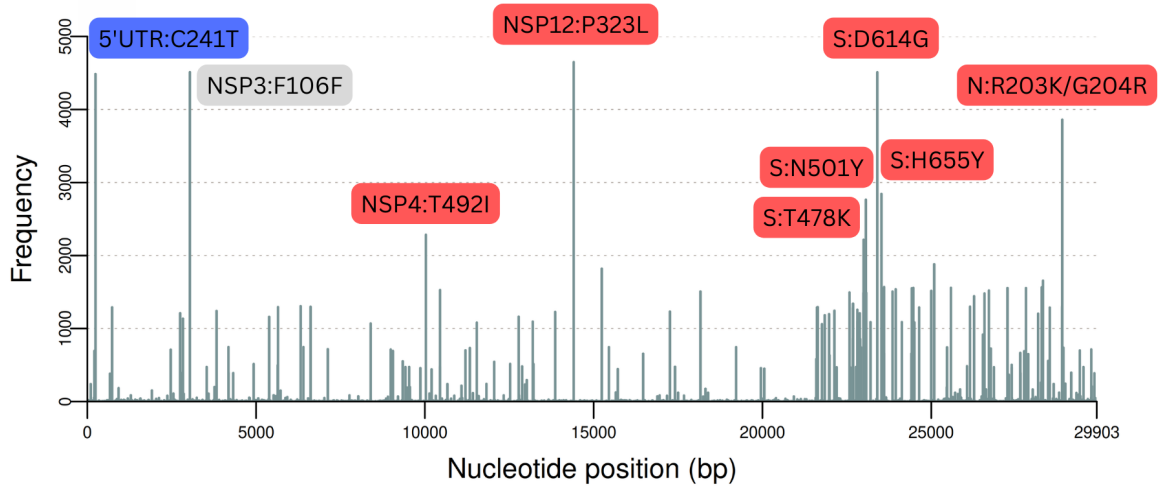

FIGURE 3 Amino acid and nucleotide substitutions associated with SARS-CoV-2 genomes from RS state. Synonymous and non-synonymous mutations are labeled with the amino acid residue. Substitutions in extragenic positions were labeled with the nucleotide alteration. Mutations occurring in more than 40% of the samples were labeled in different colors: red (non-synonymous mutations), gray (synonymous mutations), and blue (extragenic mutations).

2.4 Global Phylogenomics

In order to establish the evolutionary relationships of Esteio sequenced genomes with SARS-CoV-2 global dataset, the AudacityInstant tool from GISAID database was used to identify genetically related genomes. Four sequences could not be related to other sequences in the database, probably due to low sequencing quality. Considering the most related genomes (Table 2), four sequences were associated with samples from Brazil (one of them from Rio Grande do Sul state) and the other four were more closely related to genomes from Chile, Mexico, USA, and Canada.

TABLE 2 Closest related SARS-CoV-2 sequences from GISAID to the Esteio sequenced genomes from this study according to AudacityInstant search.

Sequence	Closest related genome				
	Distance	Match quality	Location	Collection date	Lineage
RS-44473	No related genomes found				
RS-44474	No related genomes found				
RS-44475	4	0.910	Brazil / São Paulo	2021-11-22	P.1
RS-44476	3	0.920	Canada	2021-05-26	P.1.17
RS-44477	No related genomes found				
RS-44478	4	0.900	Chile	2021-08-13	B.1.1
RS-44479	8	0.937	Brazil / Rio de Janeiro	2021-02-10	P.1
RS-44480	2	0.905	Brazil / São Paulo	2021-06-07	P.1

RS-44481	1	0.971	Brazil / Rio Grande do Sul / Guaíba	2021-06-24	P.1
RS-44482	5	0.901	USA	2021-03-17	P.1.13
RS-44483	No related genomes found				
RS-44484	4	0.903	Mexico	2021-06-05	P.1.17

Besides the most closely related genomes, other 424 unique sequences (638 genomes in total) were recovered as being related to the sequenced genomes from this study with a genetic distance of 9 or less according to AudacityInstant parameters (Figure 4). Sequences RS-44475, RS-44478, RS-44479, and RS-44481 were mostly associated with Brazilian sequences (47.4 up to 89% of the retrieved genomes). RS-44476, RS-44480, and RS-44484 were predominantly related to Mexico (50%) and USA (30.3 and 27.5% of the genomes), respectively.

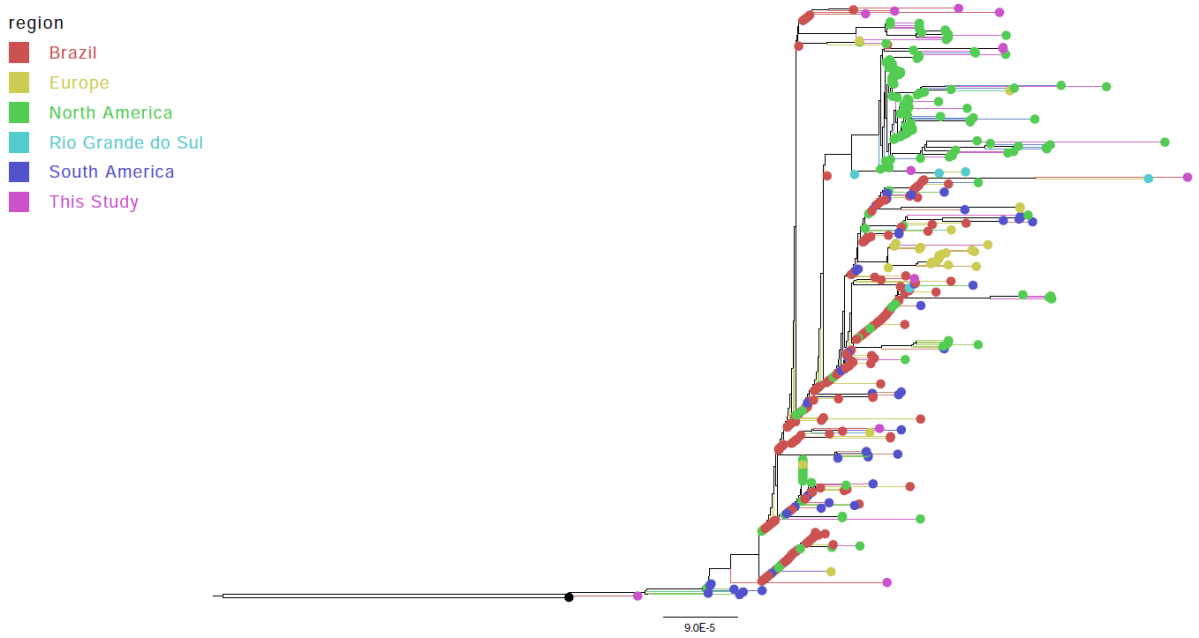

FIGURE 4 Global Maximum likelihood phylogenomic analysis with SARS-CoV-2 genomes closely related to eight sequenced genomes from this study according to Table 2.

2.5 Phylogenomic analysis of SARS-CoV-2 from Rio Grande do Sul state

The phylogenomic analysis of the SARS-CoV-2 genomes from Rio Grande do Sul state showed the formation of multiple monophyletic groups for the main VOCs (Figure 5). Alpha (B.1.1.7), Gamma (P.1 and derivative lineages), Delta (B.1.617.2 and derivative lineages), and Omicron (BA.2) clades were validated by SH-aLRT and aBayes tests with at least 97% of branch support (100/1, 99.9/1, 97.1/1, and 100/1 for Alpha, Gamma, Delta, and Omicron, respectively). For other lineages and former VOIs P.7 and Zeta (P.2), it was also observed the clustering in monophyletic groups (97.1/1, and 99/1 of statistical support for P.7 and P.2, respectively), as well as a larger clade including the P.1 (and its derivatives), P.2, and P.7 clades with B.1.1.28 sequences at basal branch (85.3/0.995 of statistical support for SH-aLRT and aBayes test). Interestingly, a clade with Alpha and Omicron genomes was formed with 86.2/0.996 of branch support.

As expected, all 12 genomes sequenced by this study clustered in the Gamma clade, which also presented subclades related to P.1 sublineages. P.1.2 (94/1 for SH-aLRT/aBayes), P.1.17 (85.3/0.997 for SH-aLRT/aBayes, including two genomes from this study at the basal branch), and P.1.7 (88.7/1 for SH-aLRT/aBayes) were found to form monophyletic groups. In the Delta group, sublineages AY.101 and AY.9.2 were supported as subclades by the statistical tests (96.3/1 and 100/1 for SH-aLRT/aBayes, respectively).

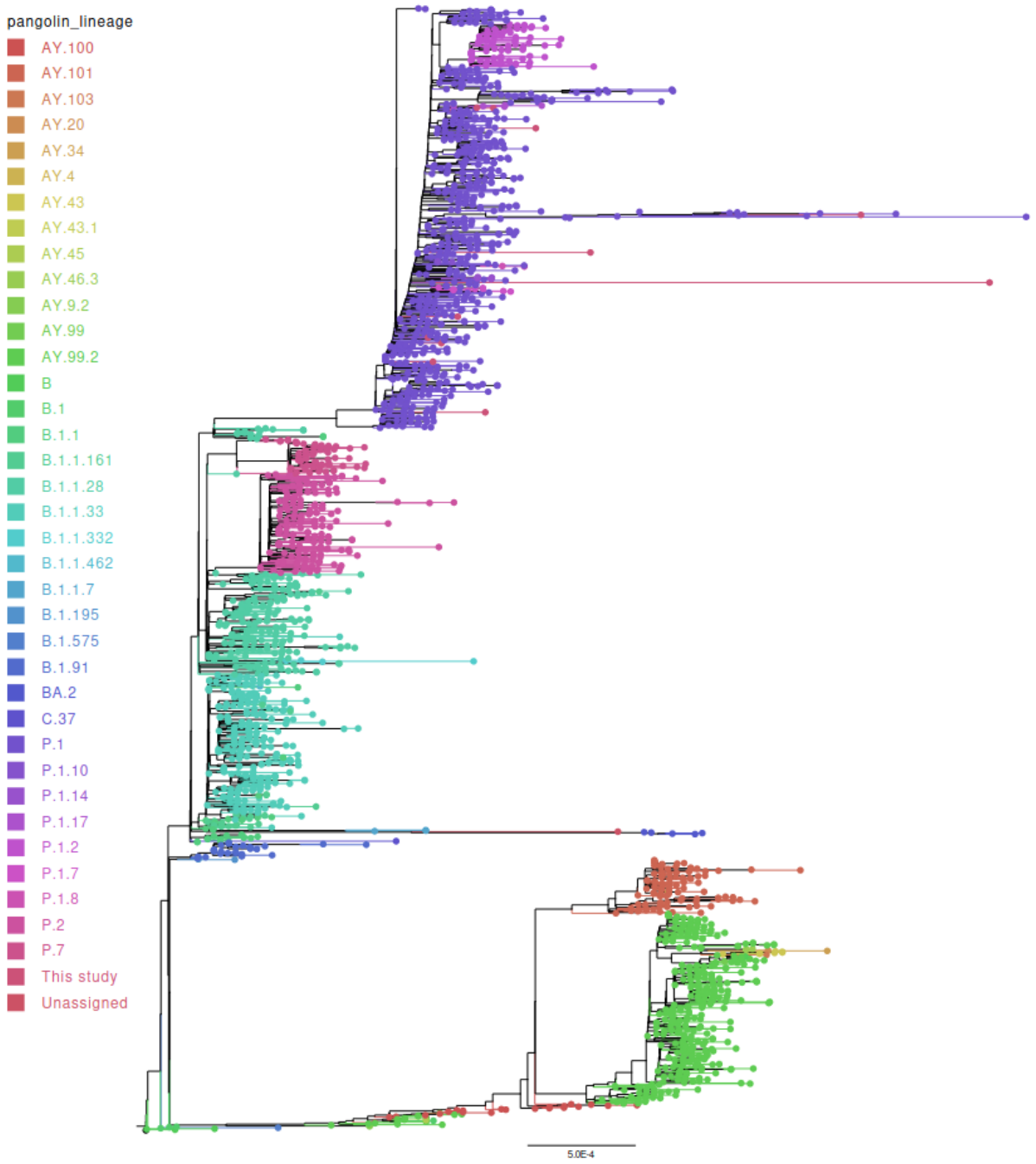

FIGURE 5 Maximum likelihood phylogenomic analysis of SARS-CoV-2 genomes from Rio Grande do Sul state.

2.6 Phylogenetics and Molecular Evolution of SARS-CoV-2 Structural Proteins

The molecular evolutionary analysis aimed to identify positively and negatively selected sites from SARS-CoV-2 structural proteins in the genome dataset from Rio Grande do Sul state. The E, M, and N proteins were tested with HyPhy FUBAR, FEL, and SLAC methods (Tables 3 - 5).

E protein

No sites were identified by FUBAR, FEL or SLAC methods to be under positive selective pressure in the Envelope protein. Two sites were identified by a negative selection test from the FEL method.

TABLE 3 Protein E sites submitted to negative selection according to the HyPhy methods.

Codon	FUBAR			FEL				SLAC		
	Alpha	Beta	Post. prob.	Alpha	Beta	LRT	Prob.	dS	dN	Prob
25	--	--	--	9.477	0.000	2.747	0.0974	--	--	--
63	--	--	--	11.265	0.000	2.756	0.0969	--	--	--

Post. prob.: Posterior probability. Prob: Probability. Gray rows indicate negative selection test results.

M protein

Two sites were identified to be under adaptive selection by FUBAR method in Membrane protein (Table 4). Twelve sites were found to be under purifying selection by FEL and/or SLAC methods, four of them by both methods.

TABLE 4 Protein M sites submitted to positive and negative selection according to the HyPhy methods.

Codon	FUBAR			FEL				SLAC		
	Alpha	Beta	Post. prob.	Alpha	Beta	LRT	Prob.	dS	dN	Prob
53	--	--	--	7.466	0.000	6.794	0.0091	3.447	0.000	0.024
94	0.764	6.559	0.9003	--	--	--	--	--	--	--
112	--	--	--	5.455	0.000	4.538	0.0332	2.306	0.000	0.084
114	--	--	--	15.269	0.000	7.834	0.0051	--	--	--
121	--	--	--	5.333	0.000	2.799	0.0943	2.294	0.000	0.084
125	1.006	10.637	0.9555	--	--	--	--	--	--	--
135	--	--	--	11.754	0.000	3.455	0.0631	--	--	--
138	--	--	--	4.521	0.000	3.761	0.0524	--	--	--
139	--	--	--	7.642	0.000	3.612	0.0574	--	--	--
147	--	--	--	7.642	0.000	3.292	0.0696	--	--	--
166	--	--	--	12.036	0.000	3.279	0.0702	--	--	--
195	--	--	--	7.642	0.000	3.873	0.0491	--	--	--
203	--	--	--	8.065	0.000	4.217	0.0400	3.443	0.000	0.024
208	--	--	--	7.764	0.000	3.503	0.0613	--	--	--

Post. prob.: Posterior probability. Prob: Probability. Gray rows indicate negative selection test results.

N protein

Fourteen sites were identified to be under positive selective pressure by FUBAR and/or FEL in Nucleocapsid protein, of which nine were identified by both methods (Table 5). Forty-six sites were found to be under negative selective pressure, twenty of them identified by FEL and SLAC methods.

TABLE 5 Protein N sites submitted to positive and negative selection according to HyPhy methods.

Codon	FUBAR			FEL				SLAC		
	Alpha	Beta	Post. prob.	Alpha	Beta	LRT	Prob.	dS	dN	Prob
9	1.163	6.688	0.9170	--	--	--	--	--	--	--
21	--	--	--	5.593	0.000	4.290	0.0383	--	--	--
30	--	--	--	5.530	0.000	3.301	0.0692	--	--	--
34	0.627	6.935	0.9841	0.000	4.547	5.308	0.0212	--	--	--
35	--	--	--	6.487	1.007	4.013	0.0452	5.000	0.500	0.018
40	--	--	--	2.586	0.000	3.399	0.0652	--	--	--
60	--	--	--	2.810	0.000	3.736	0.0532	--	--	--
63	0.626	6.645	0.9772	0.000	4.538	4.348	0.0371	--	--	--
70	--	--	--	11.034	0.000	3.401	0.0652	--	--	--
78	--	--	--	3.206	0.000	3.486	0.0619	2.379	0.000	0.079
100	--	--	--	11.034	0.000	3.763	0.0524	--	--	--

107	--	--	--	7.579	0.000	3.904	0.0482	--	--	--
110	--	--	--	6.426	0.000	11.948	0.0005	4.779	0.000	0.006
118	--	--	--	11.034	0.000	3.814	0.0508	--	--	--
149	--	--	--	3.211	0.000	3.613	0.0573	--	--	--
151	0.595	7.118	0.9879	0.000	5.412	6.086	0.0136	--	--	--
157	--	--	--	4.417	0.780	2.727	0.0987	3.213	0.484	0.092
170	--	--	--	--	--	--	--	3.000	0.000	0.037
172	--	--	--	5.021	0.000	6.599	0.0102	3.607	0.000	0.025
182	0.609	4.020	0.9256	0.000	3.128	3.258	0.0711	--	--	--
192				8.387	0.000	6.223	0.0126	5.968	0.000	0.002
208	0.604	4.018	0.9266	0.000	3.130	3.478	0.0622	--	--	--
210	--	--	--	--	--	--	--	58.466	1.127	0.008
215	0.620	3.320	0.9035	0.000	2.534	2.776	0.0957	--	--	--
221	--	--	--	1.500	0.000	2.827	0.0927	--	--	--
226	--	--	--	15.769	0.000	7.811	0.0052	2.425	0.000	0.085
227	--	--	--	3.098	0.000	5.676	0.0172	2.172	0.000	0.098
228	--	--	--	--	--	--	--	2.381	0.000	0.078
238	0.613	5.716	0.9707	0.000	3.731	3.992	0.0457	--	--	--
265	0.733	5.368	0.9120	0.000	4.711	3.278	0.0702	--	--	--
268	--	--	--	5.021	0.000	6.596	0.0102	3.573	0.000	0.026

274	--	--	--	10.000	0.000	17.898	0.000	7.165	0.000	0.000
289	1.191	6.872	0.9165	--	--	--	--	--	--	--
291	--	--	--	5.474	0.000	6.537	0.0106			
292	--	--	--	--	--	--	--	5.327	1.455	0.069
296	--	--	--	0.000	3.085	3.082	0.0791	--	--	--
298	--	--	--	4.982	0.000	6.561	0.0104	3.572	0.000	0.026
302	--	--	--	6.788	0.000	7.863	0.0050	5.000	0.000	0.004
309	--	--	--	2.810	0.000	3.332	0.0680	--	--	--
312	--	--	--	5.721	0.000	4.285	0.0384	--	--	--
313	--	--	--	2.810	0.000	3.283	0.0700	--	--	--
315	--	--	--	3.333	0.000	5.946	0.0148	2.379	0.000	0.079
318	--	--	--	4.025	0.000	6.183	0.0129	3.000	0.000	0.041
327	--	--	--	3.999	0.000	6.181	0.0129	3.000	0.000	0.041
329	--	--	--	4.000	0.000	3.733	0.0533	3.000	0.000	0.037
330	--	--	--	--	--	--	--	0.000	0.428	1.000
333	--	--	--	5.021	0.000	6.565	0.0104	3.576	0.000	0.026
337	--	--	--	3.000	0.000	4.235	0.0396	2.142	0.000	0.097
341	--	--	--	--	--	--	--	3.573	0.463	0.069
344	--	--	--	5.675	0.000	3.621	0.0571	--	--	--
346	--	--	--	4.977	0.000	8.931	0.0028	3.573	0.000	0.022

362	1.235	7.840	0.9224	--	--	--	--	--	--	--
363	--	--	--	3.301	0.000	5.949	0.0147	2.381	0.000	0.078
366	1.235	7.819	0.9223	--	--	--	--	--	--	--
372	--	--	--	11.754	0.000	3.770	0.0522	--	--	--
382	--	--	--	3.649	0.000	4.552	0.0329	--	--	--
391	0.602	5.676	0.9571	0.000	4.605	4.619	0.0316	--	--	--
398	--	--	--	11.102	1.047	3.909	0.0480	--	--	--
403	--	--	--	1.667	0.000	2.970	0.0848	--	--	--
404	--	--	--	2.810	0.000	4.459	0.0347	--	--	--

Post. prob.: Posterior probability. Prob: Probability. Gray rows indicate negative selection test results.

2.7 Molecular stability of the structural proteins E, M, and N

The program DynaMut2 was used to estimate the molecular stability of the SARS-CoV-2 structural proteins M and N with mutated residues at sites previously identified under positive selection (Table 6). Sites of M and N proteins recognized by the HyPhy tests had their amino acid substitutions evaluated at molecular level using publicly available structures from PDB database (Figure 6). Differently from the spike protein, proteins M and N are less represented in experimentally resolved structures from PDB. Thus, structures: (a) 8CTK - relative to protein M; (b) 7VNU - relative to N-terminal domain of protein N; and (c) 6ZCO - relative to C-terminal domain of protein N were selected to perform these analyzes.

In the analysis of M protein, most alterations promote a stabilizing effect in the protein structure, excepting for S94G ($\Delta\Delta G$ = -0.19 kcal/mol). In the N-terminal domain from N protein, alterations in site 63 were associated with a stabilizing effect,

while mutations in site 151 seems to destabilize the protein structure. Mutations observed in the C-terminal majoritarily suggest a destabilizing effect, excepting for mutation Q289L (ΔΔG = 0.42 kcal/mol). Some mutations such as M:S94G/I and N:Q289H/L showed variable stabilizing/destabilizing patterns for the same site. Mutations M:H125L/Y and N:D63Y not only presented a stabilizing effect but also are associated with a larger predicted stability change, increasing from 1 up to 2.04 kcal/mol in Gibbs Free Energy.

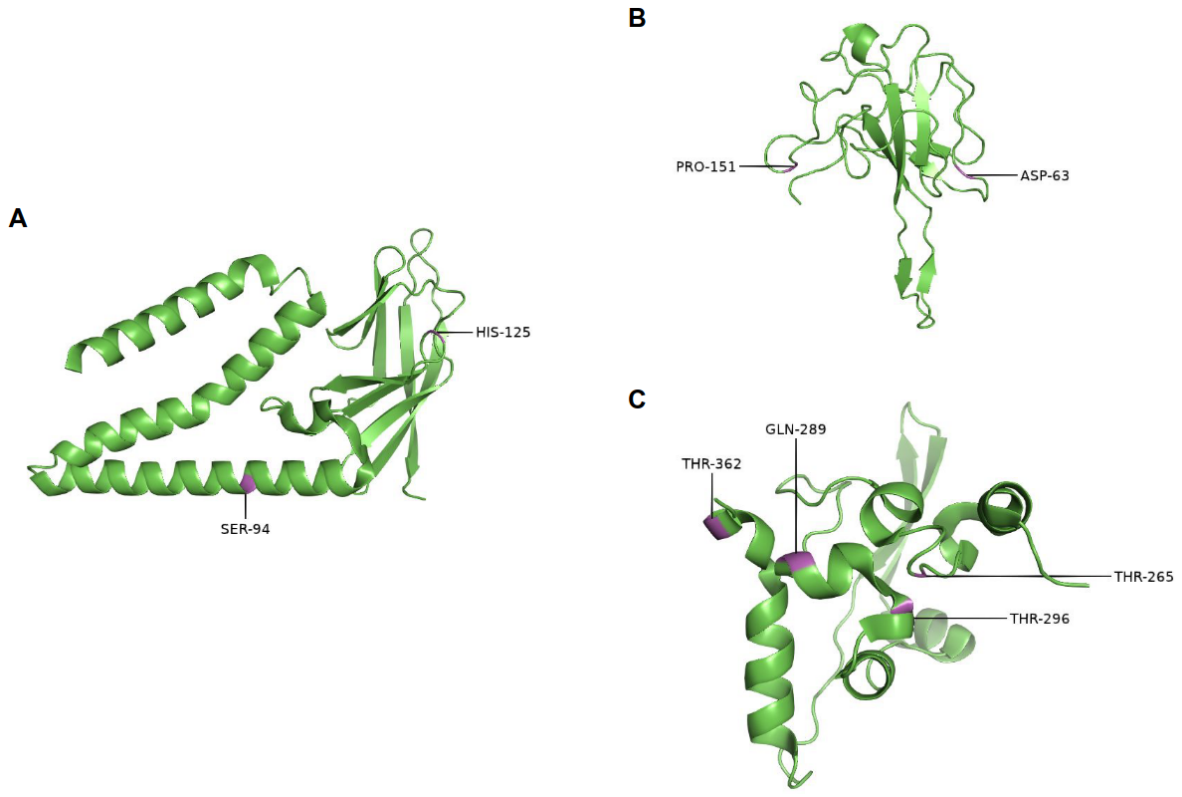

FIGURE 6 Positively selected sites in M and N protein structures. (A) Membrane protein (8CTK, chain A); (B) N-terminal domain from Nucleocapsid protein (7VNU, chain A); (C) C-terminal domain from Nucleocapsid protein (6ZCO, chain A).

TABLE 6 DynaMut2 results for positively selected sites from proteins E, M, and N.

Protein	Site	Mutation	Frequency[1]	Predicted Stability Change ($\Delta\Delta G^{Stability}$)
M	94	S → G	0.45%	-0.19 kcal/mol (Destabilizing)
M	94	S → I	0.09%	0.65 kcal/mol (Stabilizing)
M	125	H → L	0.04%	1.83 kcal/mol (Stabilizing)
M	125	H → Q	0.04%	0.52 kcal/mol (Stabilizing)
M	125	H → Y	0.18%	2.04 kcal/mol (Stabilizing)
N	63	D → G	18.53%	0.08 kcal/mol (Stabilizing)
N	63	D → Y	0.04%	1.0 kcal/mol (Stabilizing)
N	151	P → L	0.22%	-0.43 kcal/mol (Destabilizing)
N	151	P → S	0.09%	-0.17 kcal/mol (Destabilizing)
N	265	T → I	0.04%	-0.32 kcal/mol (Destabilizing)
N	289	Q → L	0.04%	0.42 kcal/mol (Stabilizing)
N	289	Q → H	0.22%	-0.38 kcal/mol (Destabilizing)
N	296	T → I	0.09%	-0.75 kcal/mol (Destabilizing)
N	362	T → I	0.18%	-0.12 kcal/mol (Destabilizing)
N	362	T → K	0.09%	-0.09 kcal/mol (Destabilizing)

[1] Mutation frequency in the multiple sequence alignment (n = 2,240 sequences).

3 DISCUSSION

Rio Grande do Sul (RS) is currently the fourth state most affected by COVID-19 in Brazil (https://covid.saude.gov.br/ accessed on December 13, 2022). The P.1 lineage initiated a new wave of infections in Brazil around November 2020, starting in Manaus (northern Brazil) and spreading across the country (Faria et al., 2021). In RS, P.1 arrived in mid of January 2021, with a high transmission rate until April 2021, characterizing the second COVID-19 wave in the state (Varela et al., 2021; Salvato et al., 2021).

According to the phylogenomic analysis, the genomes from RS state formed monophyletic groups for most of the lineages, with specific clades to lineages and sublineages belonging to Alpha, Gamma, Delta, and Omicron VOCs. This data may suggest some intra-lineage genetic conservation in the SARS-CoV-2 genomes from RS state. However, as seen in AudacityInstant data search, four of our sequenced samples are most closely related to genomes from Chile, Mexico, Canada, and USA. It is not possible to guarantee if it is not an artifact from sequencing related to low quality and covering, sampling or if it is real evidence of possible migration events, since our genomes present older collection dates than their matches. The sequencing reads assembly for these four samples achieved between 98.12 and 99.73% of SARS-CoV-2 reference sequence covering and the occurrence of undefined nucleotides (Ns) ranged from 4.47 up to 17.78%, indicating low coverage sequences.

For those sequences more closely related to Brazilian samples, except for genome RS-44479 - which was collected in June, 2021 and had their closest related sequence dated to February, 2021, in Rio de Janeiro - the remaining ones also have older collection dates than their matches. The sequencing reads assembly for these

four “Brazilian” samples achieved between 98.82 and 99.76% of SARS-CoV-2 reference sequence covering and the occurrence of undefined nucleotides (Ns) ranged from 2.00 up to 13.47%, indicating three low and one high coverage sequence in this set. The remaining four genomes with no identifiable related sequences presented very low sequencing quality and covering, ranging from 84.92 up to 98.26% of SARS-CoV-2 reference sequence covering and 49.55 up to 58.43% of undefined nucleotides.

RNA viruses have a higher mutation rate than DNA viruses and organisms (Duffy, 2018). Selective pressure occurs in a way that the virus can keep its transmission and immune evasion mechanisms updated according to the host characteristics (Zarai et al., 2020). The E protein is the smallest structural protein of SARS-CoV-2 and keeps their structure highly conserved across diverse genres of β-coronaviruses (Yadav et al., 2021). It comprises three main domains, the N-terminal (NT), C-terminal (CT), and transmembrane domain (TMD). Possible modification in TMD could indicate an differential interaction with membrane lipids, as well as the alteration of the capacity of membrane attachment and ER targeting by the E protein (Timmers et al., 2021). Similarly, sites located at the D-L-L-V motif bind to the host protein PALS1 could facilitate infection (Timmers et al., 2021). However, no sites were found under positive selective pressure in E protein. The presence of sites 25 and 63 under negative selective pressure suggest their importance to protein function conservation.

The M protein is very important in the mounting of the virion and the other structural proteins in the coronaviruses (Neuman et al., 2011). In SARS-CoV-2, this protein can be related to antigenic reactions, with the S and N proteins (Lopandić et al., 2021) even reducing the interferon I responses (Sui et al., 2021). Therefore,

modifications in its genomic structure can directly impact the virus survival, which is probably the reason for the low identification of diversifying selection events in that protein.

The N protein structure is composed of three main domains: N-terminal domain (NTD), a linker domain rich on serine and arginine residues (SR-rich linker), and a C-terminal domain (CTD) (Timmers et al., 2021). NTD and CTD comprise major antigenic sites of the N protein in SARS-CoV virus (Surjit & Lal, 2009). This protein has a role in the packing of the viral genetic material besides related to immune escape, blocking interferons and other defense mechanisms of the host (Bai et al., 2021). According to Rahman et al., several alterations in that protein makes it difficult to create vaccines and medications that could use it as a target (Rahman et al., 2020). The co-occurring amino acid mutations R203K and G204R, for example, are known to enhance replication, fitness, and pathogenesis of SARS-CoV-2 (Johnson et al., 2022).

Changes in nucleotides can result in modification of the protein structure, increasing or decreasing their stability (Jaenicke, 1996). The flexibility of a protein is related to its function and conformation (Zhao, 2010). In this way, the supervised machine-learning trained tool DynaMut2 was selected to predict the effect of missense variations from positively selected sites on protein stability. According to the DynaMut2 results for M protein, mutations in site 94 could stabilize or destabilize protein structure according to the alteration of the native serine by an isoleucine or a glycine, respectively. Site 125 achieved a stabilizing effect in all tested mutations. The mutation H125Y is the most frequent on GISAID with 14,355 occurrences in SARS-CoV-2 genomes in the world (accessed on October 31, 2022). Present in all variants of concern, this mutation was found to be prevalently associated to the Delta

and Omicron clades, 35.10% and 24.46% of the occurrences on GISAID, respectively, while the variant H125L is less spread, occurring in ≅0.001% of world genomes, including those from Alpha, Delta, and Omicron groups. Another minor variant, H125Q (≅0.0009%) is also related to Alfa, Delta, and Omicron lineages.

The N-terminal domain of N protein had two sites analyzed with 2 different mutations each. Mutations on site 63 lead to a stabilizing effect and mutations on site 151 seem to destabilize protein structure. The alteration of an aspartic acid to a glycine in D63G mutation is widely found in the world, occurring in 32.74% of world genomes on GISAID, majoritarily in Delta sequences (99.8% of D63G occurrences). D63Y is found in approximately 2,800 genomes, mostly from Omicron and Alpha lineages. P151S substitution (≅1.28% of SARS-CoV-2 genomes in the world) destabilizes N-terminal domain from Nucleocapsid protein by alteration of a non-polar proline by a polar serine. Prevalent in Omicron (95.64% of P151S occurrences on GISAID), P151S is more frequent than P151L (non-polar proline to non-polar leucine), which destabilizes the protein structure, with more than 26,000 occurrences on SARS-CoV-2 genomes in the world, including all VOCs.

In the C-terminal domain of N protein were considered four sites with their respective amino acid alterations and the majority of them seem to destabilize the protein. Only mutation Q289L (≅0.009% of GISAID genomes) demonstrates a stabilizing effect on the C-terminal domain structure. Surprisingly, as observed in N-terminal domain mutations, Q289H (≅0.08% of GISAID genomes in the world) that lead to a destabilizing effect is nine times more frequent than the substitution for a leucine residue and is mostly found in Delta and Omicron genomes.

Similar results were found by Rahman and colleagues (2020) in the analysis of the structural effects of mutation Q289H. For T362I, their results indicate a

stabilizing effect, in contrast with our findings. Finally, more studies are necessary to completely understand how structural changes may lead to advantages of SARS-CoV-2 in the host-pathogen interactions.

4 MATERIALS AND METHODS

4.1 Sample collection and clinical testing

Respiratory secretion were analyzed by Laboratório Central de Saúde Pública do Estado do Rio Grande do Sul (LACEN) (Porto Alegre, Rio Grande do Sul, Brazil) using RT-qPCR AllPlex SARS-CoV-2 assays Seegene Inc. Seoul, Republic of Korea with primers and probes targeting the RNA dependent RNA Polymerase (RdRP) Nucleocapsid (N) and Envelope (E) genes as recommended by the World Health Organization, with remnant samples stored at -20ºC. For the sequencing protocol, positive samples in the first RT-qPCR between April 09, 2021 to June 29, 2021, were selected and submitted to a second RT-qPCR, which was performed by BiomeHub (Florianópolis, Santa Catarina, Brazil), with a charite-berlin protocol. Samples with quantification cycle (Cq) up to 30 for at least one primer were selected for SARS-CoV-2 genome sequencing and assembly by the BiomeHub laboratory. In total, 12 samples who tested positive for SARS-CoV-2 RT-qPCR were included in the study.

4.2 SARS-CoV-2 genome sequencing and assembly

Total RNAs were prepared according to a reference protocol (Eden & Sim, 2020), with cDNA synthesized with SuperScript IV (Invitrogen) and DNA amplified with Platinum Taq High Fidelity (Invitrogen). The library preparation was performed with Nextera Flex (Illumina) and quantification was performed with Picogreen and

Collibri Library Quantification Kit (Invitrogen). The genome sequencing was generated on Illumina MiSeq Platform by 150x150 runs with 500xSARS-CoV-2 coverage (50-100 mil reads/per sample).

For the genome assembly (BiomeHub in-house script), the adapters removal and read trimming for 150 nt read sequences were performed by fastqtools.py. The alignment of the sequenced reads to the reference SARS-CoV-2 genome (GenBank ID: NC_045512.2) was performed by Bowtie v2.4.2 (Langmead & Salzberg, 2012) with additional parameters as end-to-end and very-sensitive. The analyses of the sequencing coverage and depth were generated by samtools v1.11 (Li et al., 2009) with minimum base quality per base (Q) ≥ 30. Finally, the consensus sequence for each SARS-CoV-2 genome was generated by a bcftools pipeline (Li, 2011), including the commands mpileup (parameters: Q ≥ 30; q ≥ 40, depth (d) ≤ 2,500), filter (parameters: DP>50), call and consensus.

4.3 SARS-CoV-2 genomes and data retrieval

In order to compare 12 SARS-CoV-2 genomes from Esteio to other samples from the state, we gathered 2,227 sequences from the GISAID database (Elbe & Buckland-Merrett, 2017) with a collection date between March 1, 2020 and May 27, 2022 (submission up to May 27, 2022). The sequences were selected according to the following filters: (i) Location: South America / Brazil / Rio Grande do Sul; (ii) Clade: all; (iii) Complete genomes; (iv) High Coverage Selected.

The analysis of sequencing efforts, lineage frequencies and genomic characterization in Rio Grande do Sul state was posteriorly performed with GISAID data by retrieval of 4,706 sequences included in the following parameters: (i) Location: South America / Brazil / Rio Grande do Sul, (ii) Collection date between

March 1, 2020 and May 31, 2022, and (iii) Submission date up to September 30, 2022.

4.4 SARS-CoV-2 mutations and lineages

SNPs and insertions/deletions in each sample were identified by the variant calling pipeline (https://github.com/tseemann/snippy), which uses FreeBayes and snpEff to call, annotate and predict variant effects on genes and proteins. The genomes were aligned with MAFFT and the extraction of SNPs and gaps from the sequences in relation to the reference was performed with msastats.py script. The reference sequence comes from the GenBank RefSeq (NC_045512.2), isolated and sequenced from an initial case from Wuhan, China, in 2019. The strains were identified using the dynamic nomenclature implemented in Pangolin (Rambaut, 2020) (https://github.com/cov-lineages/pangolin) and global clades and mutations using Nextstrain from Nextclade (https://clades.nextstrain.org/).

4.5 Phylogenomic analyses

For the global phylogenomics, a search was performed by Audacity *Instant* on the GISAID database (Elbe & Buckland-Merrett, 2017) to find closely related sequences to the sequenced genomes from this study (up to June 23, 2022). The resulting genome set was aligned with the MAFFT v.7 web server (Katoh et al., 2017). The trimming of 5' and 3' UTRs was performed with UGENE (Okonechnikov et al., 2012). The evolutionary model and phylogenomic tree inferences were performed by the IQ-TREE software (Nguyen et al., 2014) with addition of a Shimodaira-Hasegawa-like approximate likelihood ratio test of 1,000 replicates (Guindon et al., 2010) and an approximate Bayes test (Anisimova et al., 2011).

Figtree software (http://tree.bio.ed.ac.uk/software/figtree/) was used to inspect and visualize the phylogenomic tree.

For the local phylogenomic analyses, 2,227 genome sequences from Rio Grande do Sul state, previously downloaded from the GISAID database, were aligned using the MAFFT v.7 web server. The trimming of 5' and 3' UTRs was performed with UGENE, identification of the best evolutionary model and phylogenomic inference performed by IQ-TREE software with the same parameters used for the global phylogenomics previously described, with Figtree software used for the inspection and visualization of phylogenomic tree.

4.6 Phylogenetics and Molecular Evolution of SARS-CoV-2 Structural Proteins

In order to infer the phylogenetic patterns of structural proteins E, M and N, genomic alignment coordinates related to these sequences were exported according to SARS-CoV-2 reference genome (NC_045512.2). Sequences with nucleotide insertions altering the reading frame were excluded from the analysis. For each gene sequence alignment, the evolutionary model and phylogenetic tree were inferred according to the previously described parameters from phylogenomic analysis.

Molecular evolution tests were performed with the HyPhy package (Pond et al., 2004). The methods FUBAR (Murrell et al., 2013), FEL (Kosakovsky Pond & Frost, 2005), and SLAC (Kosakovsky Pond & Frost, 2005) were implemented to evaluate potential sites under adaptive (pervasive) and purifying selection.

4.7 Molecular stability of structural proteins M and N

The estimation of molecular stability of structural proteins was performed by DynaMut2 web server (Rodrigues et al., 2020) using the experimentally resolved

structures by Electron Microscopy: (a) 8CTK (3.52 Å) - relative to protein M; and X-Ray Diffraction: (b) 7VNU (1.95 Å) - relative to N-terminal domain of protein N; and (c) 6ZCO (1.36 Å) - relative to C-terminal domain of protein N, from Protein Data Bank (https://www.rcsb.org/). The selection of tested amino acid mutations was defined according to the sites detected under positive selection by the molecular evolution tests.

DATA AVAILABILITY STATEMENT

Full tables acknowledging the authors and corresponding labs submitting sequencing data used in this study can be found in Supplementary File 2. The genomes sequenced by this study are deposited on GISAID database under identification codes EPI_ISL_16106069 up to EPI_ISL_16106069. Additional information related to the current study as well as the two genome sequences with long Ns stretches (>50%) are available from the corresponding author on reasonable request.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

ETHICS STATEMENT

The research protocol was approved with exemption of written informed consent for viral genome sequencing and bioinformatic analyses by Comitê de Ética em Pesquisa em Seres Humanos of Universidade Federal de Ciências da Saúde de Porto Alegre (CEP - UFCSPA) under process number CAAE 39247920.0.0000.5345.

AUTHOR CONTRIBUTION

Amanda de M. Mayer: Formal analysis; investigation; methodology; writing – original draft; writing - review and editing. **Patrícia A. G. Ferrareze:** Conceptualization, formal analysis; investigation; methodology; writing – original draft; writing - review and editing. **Luiz F. V. de Oliveira:** Methodology; resources; funding acquisition; writing - review and editing. **Tatiana S. Gregianini:** Methodology; resources; writing - review and editing. **Carla L. A. M. N.:** Resources; writing - review and editing. **Gabriel D. Caldana:** Investigation; writing – original draft; writing - review and editing. **Lívia Kmetzsch:** Supervision; writing - review and editing. **Claudia E. Thompson:** Conceptualization, formal analysis; investigation; methodology; resources; supervision; funding acquisition; writing – original draft; writing - review and editing. All authors have read and approved the manuscript.

ACKNOWLEDGEMENTS

We thank the administrators of the GISAID database and research groups across the world for supporting the rapid and transparent sharing of genomic data during the COVID-19 pandemic and the *Governo do Estado do Rio Grande do Sul* and *Ministério da Saúde* for supplies and equipment used in the SARS-CoV-2 diagnosis routine. This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)* - Finance code 001. The genome sequencing was performed by BiomeHub laboratory.

References

Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., & Gascuel, O. (2011). Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Systematic Biology*, *60*(5), 685–699. https://doi.org/10.1093/sysbio/syr041

Bai, Z., Cao, Y., Liu, W., & Li, J. (2021). The sars-cov-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug or vaccine mitigation. *Viruses*, *13*(6). https://doi.org/10.3390/v13061115

Chakraborty, S. (2022). E484K and N501Y SARS-CoV 2 spike mutants Increase ACE2 recognition but reduce affinity for neutralizing antibody. *International Immunopharmacology*, *102*, 108424. https://doi.org/10.1016/j.intimp.2021.108424

Chen, J., Wang, R., Gilby, N. B., & Wei, G.-W. (2022). Omicron variant (B.1.1.529): Infectivity, vaccine breakthrough, and antibody resistance. *Journal of Chemical Information and Modeling*, *62*(2), 412–422. https://doi.org/10.1021/acs.jcim.1c01451

Chen, R. E., Zhang, X., Case, J. B., Winkler, E. S., Liu, Y., VanBlargan, L. A., Liu, J., Errico, J. M., Xie, X., Suryadevara, N., Gilchuk, P., Zost, S. J., Tahan, S., Droit, L., Turner, J. S., Kim, W., Schmitz, A. J., Thapa, M., Wang, D., … Diamond, M. S. (2021). Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nature Medicine*, *27*(4), 717–726. https://doi.org/10.1038/s41591-021-01294-w

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/s1473-3099(20)30120-1

Duffy, S. (2018). Why are RNA virus mutation rates so damn high? *PLoS Biology*, *16*(8). https://doi.org/10.1371/journal.pbio.3000003

Eden, J.-S., & Sim, E. (2020). *SARS-CoV-2 Genome Sequencing Using Long Pooled Amplicons on Illumina Platforms v1*. ZappyLab, Inc. http://dx.doi.org/10.17504/protocols.io.befyjbpw

Elbe, S., & Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, *1*(1), 33–46. https://doi.org/10.1002/gch2.1018

Emam, M., Oweda, M., Antunes, A., & El-Hadidi, M. (2021). Positive selection as a key player for SARS-CoV-2 pathogenicity: Insights into ORF1ab, S and E genes. *Virus Research*, *302*, 198472. https://doi.org/10.1016/j.virusres.2021.198472

Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. da S., Mishra, S., Crispim, M. A. E., Sales, F. C. S., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., de Jesus, J. G., Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., … Sabino, E. C. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*

(New York, N.y.), *372*(6544). https://doi.org/10.1126/science.abh2644

Franceschi, V. B., Caldana, G. D., Perin, C., Horn, A., Peter, C., Cybis, G. B., Ferrareze, P. A. G., Rotta, L. N., Cadegiani, F. A., Zimerman, R. A., & Thompson, C. E. (2021). Predominance of the sars-cov-2 lineage P.1 And its sublineage P.1.2 in patients from the metropolitan region of Porto Alegre, southern Brazil in march 2021. *Pathogens*, *10*(8), 988. https://doi.org/10.3390/pathogens10080988

Ghosh, S., & Chakraborty, S. (2020). Phylogenomics analysis of sars-cov2 genomes reveals distinct selection pressure on different viral strains. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/5746461

Giovanetti, M., Fonseca, V., Wilkinson, E., Tegally, H., San, E. J., Althaus, C. L., Xavier, J., Nanev Slavov, S., Viala, V. L., Ranieri Jerônimo Lima, A., Ribeiro, G., Souza-Neto, J. A., Fukumasu, H., Lehmann Coutinho, L., Venancio da Cunha, R., Freitas, C., Campelo de A e Melo, C. F., Navegantes de Araújo, W., Do Carmo Said, R. F., … de Alcantara, L. C. J. (2022). Replacement of the Gamma by the Delta variant in Brazil: Impact of lineage displacement on the ongoing pandemic. *Virus Evolution*, *8*(1). https://doi.org/10.1093/ve/veac024

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phyml 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Gräf, T., Bello, G., Naveca, F. G., Gomes, M., Cardoso, V. L. O., da Silva, A. F., Dezordi, F. Z., dos Santos, M. C., Santos, K. C. de O., Batista, É. L. R., Magalhães, A. L. Á., Vinhal, F., Miyajima, F., Faoro, H., Khouri, R., Wallau, G. L., Delatorre, E., Siqueira, M. M., Resende, P. C., … Fernandes, S. B. (2022). Phylogenetic-based inference reveals distinct transmission dynamics of SARS-CoV-2 lineages Gamma and P.2 in Brazil. *IScience*, *25*(4), 104156. https://doi.org/10.1016/j.isci.2022.104156

Gularte, J. S., da Silva, M. S., Mosena, A. C. S., Demoliner, M., Hansen, A. W., Filippi, M., Pereira, V. M. de A. G., Heldt, F. H., Weber, M. N., de Almeida, P. R., Hoffmann, A. T., Valim, A. R. de M., Possuelo, L. G., Fleck, J. D., & Spilki, F. R. (2022). Early introduction, dispersal and evolution of Delta SARS-CoV-2 in Southern Brazil, late predominance of AY.99.2 and AY.101 related lineages. *Virus Research*, *311*, 198702. https://doi.org/10.1016/j.virusres.2022.198702

Jaenicke, R. (1996). How do proteins acquire their three-dimensional structure and stability? *Naturwissenschaften*, *83*(12), 544–554. https://doi.org/10.1007/bf01141979

Johnson, B. A., Zhou, Y., Lokugamage, K. G., Vu, M. N., Bopp, N., Crocquet-Valdes, P. A., Kalveram, B., Schindewolf, C., Liu, Y., Scharton, D., Plante, J. A., Xie, X., Aguilar, P., Weaver, S. C., Shi, P.-Y., Walker, D. H., Routh, A. L., Plante, K. S., & Menachery, V. D. (2022). Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. *PLOS Pathogens*, *18*(6), e1010627. https://doi.org/10.1371/journal.ppat.1010627

Junior, R. da S. F., Lamarca, A. P., de Almeida, L. G. P., Cavalcante, L., Machado, D. T., Martins, Y., Brustolini, O., Gerber, A. L., Guimarães, A. P. de C., Gonçalves, R. B., Alves, C., Mariani, D., Cruz, T. F., de Souza, I. V., de Carvalho, E. M., Ribeiro, M. S., Carvalho, S., Silva, F. D. da, Garcia, M. H. de O., … de Vasconcelos, A. T. R. (2021). Turnover of sars-cov-2 lineages shaped the pandemic and enabled the emergence of new variants in the state of Rio de Janeiro, Brazil. *Viruses*, *13*(10). https://doi.org/10.3390/v13102013

Katoh, K., Rozewicki, J., & Yamada, K. D. (2017). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Kosakovsky Pond, S. L., & Frost, S. D. W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, *22*(5), 1208–1222. https://doi.org/10.1093/molbev/msi105

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lopandić, Z., Protić-Rosić, I., Todorović, A., Glamočlija, S., Gnjatović, M., Ćujic, D., & Gavrović-Jankulović, M. (2021). IgM and igg immunoreactivity of sars-cov-2 recombinant M protein. *International Journal of Molecular Sciences*, *22*(9). https://doi.org/10.3390/ijms22094951

Matsuo, T. (2021). Viewing sars-cov-2 nucleocapsid protein in terms of molecular flexibility. *Biology*, *10*(6). https://doi.org/10.3390/biology10060454

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: A fast, unconstrained bayesian approximation for inferring selection. *Molecular Biology and Evolution*, *30*(5), 1196–1205. https://doi.org/10.1093/molbev/mst030

Neuman, B. W., Kiss, G., Kunding, A. H., Bhella, D., Baksh, M. F., Connelly, S., Droese, B., Klaus, J. P., Makino, S., Sawicki, S. G., Siddell, S. G., Stamou, D. G., Wilson, I. A., Kuhn, P., & Buchmeier, M. J. (2011). A structural analysis of M protein in coronavirus assembly and morphology. *Journal of Structural Biology*, *174*(1). https://doi.org/10.1016/j.jsb.2010.11.021

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Okonechnikov, K., Golosova, O., & Fursov, M. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, *28*(8), 1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2004). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. https://doi.org/10.1093/bioinformatics/bti079

Rahman, M. S., Islam, M. R., Alam, A. S. M. R. U., Islam, I., Hoque, M. N., Akter, S., Rahaman, Md. M., Sultana, M., & Hossain, M. A. (2020). Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *Journal of Medical Virology*, *93*(4), 2177–2195. https://doi.org/10.1002/jmv.26626

Rodrigues, C. H. M., Pires, D. E. V., & Ascher, D. B. (2020). DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science*, *30*(1), 60–69. https://doi.org/10.1002/pro.3942

Salvato, R. S., Gregianini, T. S., Campos, A. A. S., Crescente, L. V., Vallandro, M. J., Ranieri, T. M. S., Vizeu, S., Martins, L. G., Da Silva, E. V., Pedroso, E. R., Burille, A., Baethgen, L. F., Schiefelbein, S. H., Machado, T. R. M., Becker, I. M., Ramos, R., Piazza, C. F., Nunes, Z. M. A., & Bastos, C. G. M. B. (2021). Epidemiological investigation reveals local transmission of SARS-CoV-2 lineage P.1 in Southern Brazil. *Revista de Epidemiologia e Controle de Infecção*, *11*(1). https://doi.org/10.17058/reci.v1i1.16335

Satarker, S., & Nampoothiri, M. (2020). Structural proteins in severe acute respiratory syndrome coronavirus-2. *Archives of Medical Research*, *51*(6), 482–491. https://doi.org/10.1016/j.arcmed.2020.05.012

Shiehzadegan, S., Alaghemand, N., Fox, M., & Venketaraman, V. (2021). Analysis of the delta variant B.1.617.2 COVID-19. *Clinics and Practice*, *11*(4). https://doi.org/10.3390/clinpract11040093

Sui, L., Zhao, Y., Wang, W., Wu, P., Wang, Z., Yu, Y., Hou, Z., Tan, G., & Liu, Q. (2021). SARS-CoV-2 membrane protein inhibits type I interferon production through ubiquitin-mediated degradation of TBK1. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.662989

Surjit, M., & Lal, S. K. (2009). The nucleocapsid protein of the SARS coronavirus: Structure, function and therapeutic potential. In *Molecular Biology of the SARS-Coronavirus* (pp. 129–151). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-03683-5_9

Timmers, L. F. S. M., Peixoto, J. V., Ducati, R. G., Bachega, J. F. R., de Mattos Pereira, L., Caceres, R. A., Majolo, F., da Silva, G. L., Anton, D. B., Dellagostin, O. A., Henriques, J. A. P., Xavier, L. L., Goettert, M. I., & Laufer, S. (2021). SARS-CoV-2 mutations in Brazil: From genomics to putative clinical conditions. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-91585-6

Varela, A. P. M., Prichula, J., Mayer, F. Q., Salvato, R. S., Sant'Anna, F. H., Gregianini, T. S., Martins, L. G., Seixas, A., & Veiga, A. B. G. da. (2021). SARS-CoV-2 introduction and lineage dynamics across three epidemic peaks in Southern Brazil: Massive spread of P.1. *Infection, Genetics and Evolution*, *96*. https://doi.org/10.1016/j.meegid.2021.105144

Wang, X., & Powell, C. A. (2021). How to translate the knowledge of COVID‐19 into the prevention of Omicron variants. *Clinical and Translational Medicine*, *11*(12). https://doi.org/10.1002/ctm2.680

Wink, P. L., Ramalho, R., Monteiro, F. L., Volpato, F. C. Z., Willig, J. B., Lovison, O. von A., Zavascki, A. P., Barth, A. L., & Martins, A. F. (2022). Genomic surveillance of sars-cov-2 lineages indicates early circulation of P.1 (Gamma) variant of concern in southern Brazil. *Microbiology Spectrum*, *10*(1). https://doi.org/10.1128/spectrum.01511-21

Yadav, R., Chaudhary, J. K., Jain, N., Chaudhary, P. K., Khanra, S., Dhamija, P., Sharma, A., Kumar, A., & Handu, S. (2021). Role of structural and non-structural proteins and therapeutic targets of sars-cov-2 for COVID-19. *Cells*, *10*(4), 821. https://doi.org/10.3390/cells10040821

Zarai, Y., Zafrir, Z., Siridechadilok, B., Suphatrakul, A., Roopin, M., Julander, J., & Tuller, T. (2020). Evolutionary selection against short nucleotide sequences in viruses and their related hosts. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *27*(2). https://doi.org/10.1093/dnares/dsaa008

Zhao, Q. (2010). Protein flexibility as a biosignal. *Critical ReviewsTM in Eukaryotic Gene Expression*,

20(2), 157–170. https://doi.org/10.1615/critreveukargeneexpr.v20.i2.60

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, *382*(8), 727–733. https://doi.org/10.1056/nejmoa2001017