The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the Spike protein

Paola Cristina Resende[1]*, Felipe G Naveca*[2], Roberto D. Lins[3], Filipe Zimmer Dezordi[4,5], Matheus V. F. Ferraz[3,6], Emerson G. Moreira[3,6], Danilo F. Coêlho[3,6], Fernando Couto Motta[1], Anna Carolina Dias Paixão[1], Luciana Appolinario[1], Renata Serrano Lopes[1], Ana Carolina da Fonseca Mendonça[1], Alice Sampaio Barreto da Rocha[1], Valdinete Nascimento[2], Victor Souza[2], George Silva[2], Fernanda Nascimento[2], Lidio Gonçalves Lima Neto[7], Irina Riediger[8], Maria do Carmo Debur[8], Anderson Brandao Leite[9], Tirza Mattos[10], Cristiano Fernandes da Costa[11], Felicidade Mota Pereira[12], Ricardo Khouri[13], André Felipe Leal Bernardes[14], Edson Delatorre[15], Tiago Gräf[16], Marilda Mendonça Siqueira[1], Gonzalo Bello**[17], and Gabriel L Wallau**[4,5] on behalf of Fiocruz COVID-19 Genomic Surveillance Network.

1. Laboratory of Respiratory Viruses and Measles (LVRS), Instituto Oswaldo Cruz, FIOCRUZ-Rio de Janeiro, Brazil.
2. Laboratório de Ecologia de Doenças Transmissíveis na Amazônia (EDTA), Instituto Leônidas e Maria Deane, FIOCRUZ-Amazonas, Brazil.
3. Department of Virology, Instituto Aggeu Magalhães, FIOCRUZ-Pernambuco, Brazil.
4. Departamento de Entomologia, Instituto Aggeu Magalhães, FIOCRUZ-Pernambuco, Brazil.
5. Núcleo de Bioinformática (NBI), Instituto Aggeu Magalhães FIOCRUZ-Pernambuco, Brazil.
6. Department of Fundamental Chemistry, Federal University of Pernambuco, Recife, Brazil
7. Laboratório Central de Saúde Pública do Estado do Maranhão (LACEN-MA), Brazil.
8. Laboratório Central de Saúde Pública do Estado do Paraná (LACEN-PR), Brazil.
9. Laboratório Central de Saúde Pública do Estado do Alagoas (LACEN-AL), Brazil.
10. Laboratório Central de Saúde Pública do Amazonas (LACEN-AM), Brazil.
11. Fundação de Vigilância em Saúde do Amazonas, Brazil.
12. Laboratório Central de Saúde Pública do Estado da Bahia (LACEN-BA), Brazil.
13. Laboratório de Enfermidades Infecciosas Transmitidas por Vetores, Instituto Gonçalo Moniz, FIOCRUZ-Bahia, Brazil.
14. Laboratório Central de Saúde Pública do Estado de Minas Gerais (LACEN-MG).
15. Departamento de Biologia. Centro de Ciências Exatas, Naturais e da Saúde, Universidade Federal do Espírito Santo, Alegre, Brazil.
16. Plataforma de Vigilância Molecular, Instituto Gonçalo Moniz, FIOCRUZ-Bahia, Brazil.
17. Laboratório de AIDS e Imunologia Molecular, Instituto Oswaldo Cruz, FIOCRUZ-Rio de Janeiro, Brazil.

*, ** These authors contributed equally to this work.

Abstract

Mutations at both the receptor-binding domain (RBD) and the amino (N)-terminal domain (NTD) of the SARS-CoV-2 Spike (S) glycoprotein can alter its antigenicity and promote immune escape. We identified that SARS-CoV-2 lineages circulating in Brazil with mutations of concern in the RBD independently acquired convergent deletions and insertions in the NTD of the S protein, which altered the NTD antigenic-supersite and other predicted epitopes at this region. These findings support that the ongoing widespread transmission of SARS-CoV-2 in Brazil is generating new viral lineages that might be more resistant to neutralization than parental variants of concern.

Background

Recurrent deletions in the amino (N)-terminal domain (NTD) of the spike (S) glycoprotein of SARS-CoV-2 has been identified during long-term infection of immunocompromised patients [1–4] as well as during extended human-to-human transmission [3]. Most of those deletions (90%) maintain the reading frame and cover four recurrent deletion regions (RDRs) within the NTD at positions 60-75 (RDR1), 139-146 (RDR2), 210-212 (RDR3), and 242-248 (RDR4) of the S protein [3]. The RDRs that occupy defined antibody epitopes within the NTD and RDR variants might alter antigenicity [3]. Interestingly, the RDRs overlap with four Indel Regions (IR) at the NTD (IR-2 to IR-5) that are prone to gain or lose short nucleotide sequences during sarbecoviruses evolution both in animals and humans [5,6].

Since late 2020, several more transmissible variants of concern (VOCs) and also variants of interest (VOI) with convergent mutations at the receptor-binding domain (RBD) of the S protein (particularly E484K and N501Y) arose independently in humans [7,8]. Some VOCs also displayed NTD deletions like lineages B.1.1.7 (RDR2 Δ144), B.1.351 (RDR4 Δ242-244), and P.3 (RDR2 Δ141-143) that were initially detected in the United Kingdom, South Africa, and the Philippines, respectively [3]. The VOCs B.1.1.7 and B.1.351 are resistant to neutralization by several anti-NTD monoclonal antibodies (mAbs) and NTD deletions at RDR2 and RDR4 are crucial for such phenotype [9–14]. Thus, NTD mutations and deletions represent an important mechanism of immune evasion and accelerate SARS-CoV-2 adaptive evolution in humans.

Several SARS-CoV-2 variants with mutations in the RBD have been described in Brazil, including the VOC P.1 [15] and the VOIs P.2 [16] and N.9 [17]. The VOC P.1 also displayed NTD mutations (L18F) that abrogate binding of some anti-NTD mAbs [14], but none of those variants displayed indels in the NTD. Importantly, although VOC P.1 showed reduced binding to RBD-directed antibodies, it is more susceptible to anti-NTD mAbs than other VOCs [9–14]. In this study, we monitored and characterized the emergence of RDR variants within VOC and VOI circulating in Brazil that were genotyped by the Fiocruz COVID-19 Genomic Surveillance Network between November 2020 and February 2021.

Results

Our genomic survey identified 11 SARS-CoV-2 sequences from five different Brazilian states (Amazonas, Bahia, Maranhao, Parana, and Rondonia) that harbor a

variable combination of mutations in the RBD (K417T, E484K, N501Y) and indels in the NTD of the S protein (**Table 1**). One VOI P.2 sequence and one VOC P.1 sequence displayed a convergent amino acid deletion Δ144 in the RDR2, while two VOC P.1 sequences displayed a four amino acid deletion Δ141-144 in the RDR2. On the other hand, one VOC P.1 sequence harbors a two amino acid deletion Δ189-190; two B.1.1.33(E484K) sequences carried deletions Δ141-144, Δ211 and Δ256-258, and four B.1.1.28 sequence displayed a four amino acid insertion ins214ANRN. We also identified B.1.1.28 ins214ANRN variants sharing six out of 10 P.1 lineage-defining mutations in the Spike protein (L18F, P26S, D138Y, K417T, E484K, N501Y) as well as P.1 lineage-defining mutations in the NSP3 (K977Q), NS3 (S253P) and N (P80R) proteins, thus defined as P.1-like variants. Inspection of sequences available at EpiCoV database in the GISAID (https://www.gisaid.org/) at March 1st, 2021, revealed one B.1.1.28 from the Amazonas state and three P.1 sequences from the Bahia state with deletion Δ144 **(Table 1).** All three P.1 Δ144 sequences from Bahia were recovered from individuals reporting a travel history to the Amazonas state [18].

Table 1. SARS-CoV-2 Brazilian variants with indels at NTD of the Spike protein.

Sample(s)	Lineage	NTD Indel	RBD	GISAID ID
AM-FIOCRUZ-20842572LS/2020*	B.1.1.28	Δ144	-	EPI_ISL_1068132
MG-FIOCRUZ-8180/2021	P.2	Δ144	E484K	EPI_ISL_1219137
BA53/2021*	P.1	Δ144	K417T	EPI_ISL_1067729
BA54/2021*			E484K	EPI_ISL_1067733
BA55/2021*			N501Y	EPI_ISL_1067734
BA-FIOCRUZ-7029/2021*				EPI_ISL_1219136
AL-FIOCRUZ-4795/2021*	P.1	Δ141-144	K417T	EPI_ISL_1219134
PR-FIOCRUZ-5273/2021**			E484K	EPI_ISL_1219133
			N501Y	
AL-FIOCRUZ-4786/2021*	P.1	Δ189-190	K417T	EPI_ISL_1219135
			E484K	
			N501Y	
MA-FIOCRUZ-6871/2021	B.1.1.33	Δ141-144	V445A	EPI_ISL_1181371
MA-FIOCRUZ-6874/2021	(E484K)	Δ211	E484K	EPI_ISL_1181370
		Δ256-258		
AM-FIOCRUZ-20897269OP*	B.1.1.28	ins214ANRN	K417T	EPI_ISL_1068256
AM-FIOCRUZ-20897281WS*	(P.1-like)		E484K	EPI_ISL_1219132
AM-FIOCRUZ-21840593CL*			N501Y	EPI_ISL_1261122
PR-FIOCRUZ-5241/2021				EPI_ISL_1261123

*Patient from Amazonas state or traveller returning from Amazonas state. ** Patient from Rondonia.

The Maximum Likelihood (ML) phylogenetic analyses showed that P.1 variants Δ141-144 were intermixed among non-deleted sequences (**Fig. 1A**). The four P.1 Δ144 sequences detected in Bahia state; however, branched in a subclade (aLRT = 77%) together with the Δ189-190 variant and the other two lineages P.1 sequences that share the synonymous mutation A18945G (**Fig. 1A**). The four P.1-like ins214ANRN and the two B.1.1.33(E484K) Δ141-144/211/256-258 variants also clustered in highly supported (aLRT = 100%) monophyletic clades (**Fig. 1A and B**). These findings suggest that P.1 Δ141-144 variants resulted from independent convergent NTD deletions events, while P.1 Δ144, P.1-like ins214ANRN and B.1.1.33(E484K) Δ141-144/211/256-258 variants might represent newly emergent VOIs or VOCs. It is interesting to note that most P.1 sequences with NTD deletions were detected in individuals from or with travel history to the Amazonas state.

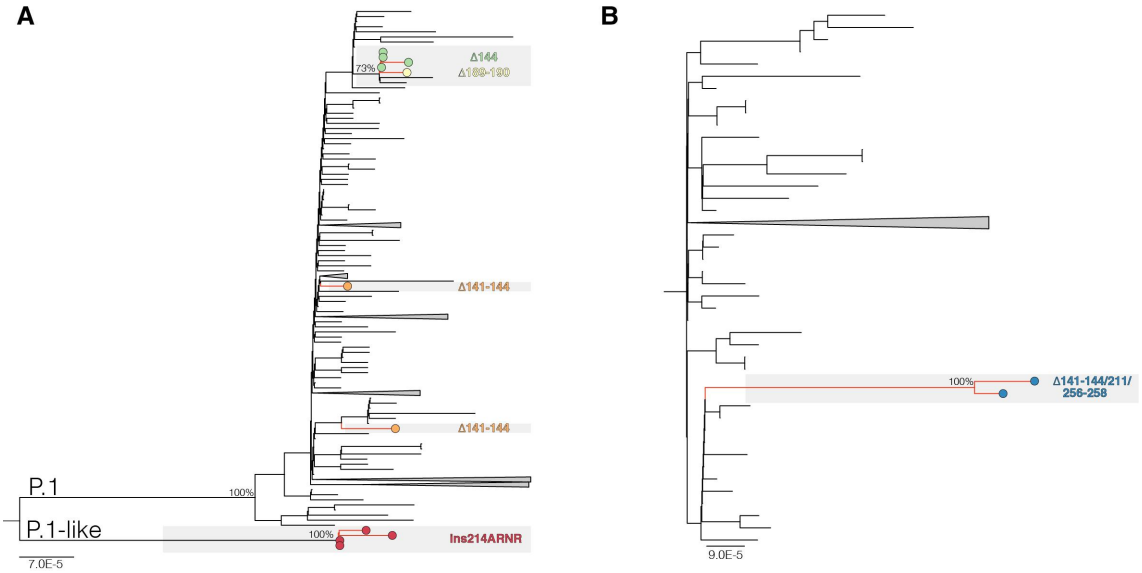

Figure 1. ML phylogenetic tree of whole-genome lineage P.1/P.1-like (**A**) and B.1.1.33 (**B**) Brazilian sequences showing the recurrent emergence of deletions at the NTD of the S protein. Tip circles representing the SARS-CoV-2 sequences with NTD indels are colored as indicated. The trees were rooted at midpoint and branch lengths are drawn to scale with the left bar indicating nucleotide substitutions per site. For visual clarity, some clades are collapsed into triangles.

While SARS-CoV-2 variants harboring NTD deletions at RDR2 and RDR4 have emerged in many different lineages globally, the ins214 in the S protein is a more rare event. Our search of SARS-CoV-2 sequences available at EpiCoV database in the

GISAID (https://www.gisaid.org/) at March 1st retrieved only 146 SARS-CoV-2 sequences of lineages A.2.4 (n = 52), B (n = 3), B.1 (n = 7), B.1.1.7 (n = 1), B.1.177 (n = 1), B.1.2 (n = 1), B.1.214 (n = 80) and B.1.429 (n = 1) that displayed an insert motif of three to four amino acids (AKKN, KLGB, AQER, AAG, KFH, KRI, and TDR) in position 214 (**Appendix Table 1**). Most ins214 motifs were unique, except ins214TDR, which seems to have arisen independently in B.1 and B.1.214. With the only exception of one lineage B sequence sampled in March 2020, all SARS-CoV-2 ins214 variants were only detected since November 2020, and its frequency increased in 2021 mainly due to the recent dissemination of lineage A.2.4 ins214AAG in Central and North America and lineage B.1.214 ins214TDR in Europe.

Next, we aligned the S protein of representative sequences of SARS-CoV-2 lineages with NTD indels and SARS-CoV-2-related coronavirus (SC2r-CoV) lineages from bats and pangolins [19]. Inspection of the alignment confirms that most NTD indels detected in the SARS-CoV-2 lineages occur within IR previously defined in sarbecovirus (**Fig. 2**). The **Δ**141-144 occurs in the IR-3 located in the central part of the NTD, where some bats SC2r-CoV also have deletions. The ins214 occurs in the IR-4 where an insertion of four amino acids was detected in three bat SC2r-CoV isolated in China (RmYN02, ins214GATP), Thailand (RacCS203, ins214GATP), and Japan (Rc-o319, ins214GATS). Despite amino acid motifs at ins214 were very different across SARS-CoV-2 and SC2r-CoV lineages, the insertion size was conserved (3-4 amino acids). The **Δ**256-258 occurs near the IR-5, where some bat and pangolin SC2r-CoV lineages also displayed deletions. Thus, the NTD regions that are prone to gain indels during viral transmission among animals are the same as those detected during transmissions in humans.

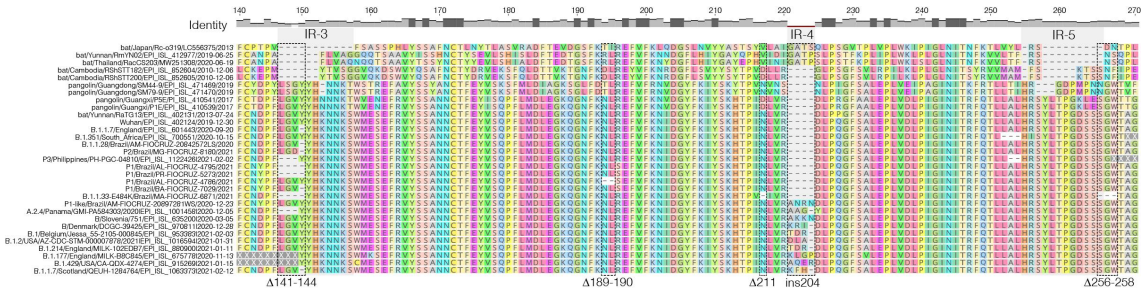

Figure 2. Amino acid alignment of positions 140-270 of the S protein of representative sequences of SARS-CoV-2 lineages harboring indels in the NTD and SARS-CoV-2-related coronavirus (SC2r-CoV) from bats and pangolins. IRs positions (gray shaded areas) are approximations due to the high genetic variability in these alignment positions. Dotted

rectangles highlight the indels identified in this study. The identity level estimated for each position of the alignment is displayed at the top.

Epitope mapping showed that neutralizing antibodies are primarily directed against the RBD and NTD of the S protein [9,20–23]. Some of the RBD mutations (K417T and E484K) detected in the VOCs and VOIs circulating in Brazil have been associated with increased resistance to neutralization by mAbs, or polyclonal sera from convalescent and vaccinated subjects [24–27]. The RDR2 **Δ**144 and RDR4 **Δ**242-244 deletions observed in VOCs B.1.1.7 and B.1.35, respectively, are located in the N3 (residues 141 to 156) and N5 (residues 246 to 260) loops that composes the NTD antigenic-supersite [28,29] and confers resistance to neutralization by anti-NTD mAbs [3,9,10,30]. Moreover, *in vitro* co-incubation of SARS-CoV-2 with highly neutralizing plasma form COVID-19 convalescent patient, has revealed an incremental resistance to neutralization followed by the stepwise acquisition of indels at N3/N5 loops [31]. Notably, SARS-CoV-2 challenge in hamsters previously treated with anti-NTD mAbs revealed selection of two escape mutants harboring NTD deletions **Δ**143-144 and **Δ**141-144 [14]. Thus, NTD indels might represent a mechanism of ongoing adaptive evolution of VOC and VOI circulating in Brazil to escape from dominant neutralizing antibodies directed against the NTD antigenic-supersite.

To test this hypothesis, we performed a modeling analysis of the binding interface between wildtype/indels NTD variants and the NTD-directed neutralizing antibody (NAb) 2-51 derived from a convalescent donor [20]. The NAb 2-51 makes several contacts with the wildtype NTD antigenic-supersite (EPI_ISL_402124), primarily through the heavy-chain (**Fig. 3**). The loops N3 and N5 play a pivotal role in the binding process with a predominance of hydrophobic contacts and dispersion interactions in N5 and saline interactions in N3. Our result shows that deletions at RDR 2 (**Δ**144, **Δ**141-143) and RDR4 (**Δ**242-244) impact the loops' size and conformation, disrupting the native contacts and reducing the interacting hydrophobic surface accessible area, mainly due to the loss of the hydrophobic pocket (**Figure S1**). Indels around the N3/N5 loops resulted in a significant loss of interactions (both electrostatic and hydrophobic) (**Table 2**) that will dramatically impact the binding free energy, and therefore the binding affinity, between those NTD deletion variants and the NAb 2-51. Although NTD indels **Δ**189-190 and ins214ANRN did not affect the NTD antigenic-supersite, they occur at putative epitope regions covering residues

168/173-188 and 209-216 (**Appendix Table 2**) and leads to conformational changes in exterior loops **(Figure S1 G-H)** which might affect Ab binding outside the antigenic-supersite. These findings suggest that NTD deletions Δ144, Δ141-143, and Δ242-244 here detected probably abrogate the binding of NAb directed against the NTD antigenic-supersite and confirm that deletions at RDRs 2/4 are an essential mechanism for SARS-CoV-2 immune evasion [3,14].

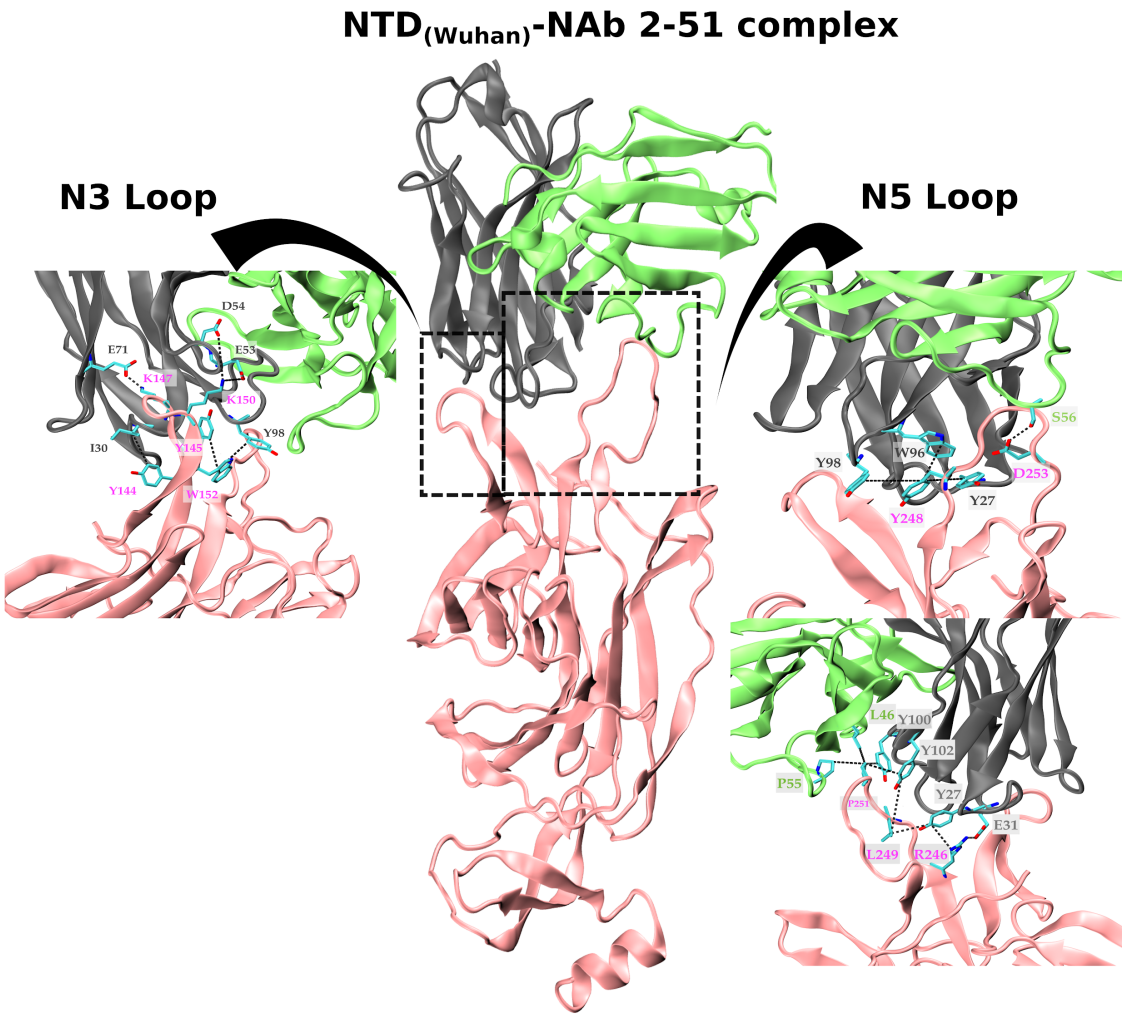

Figure 3. List of native interactions showed onto the 3D structure of the S protein NTD targeted by a natural mAb. Cartoon representation of the structure of the NTD protein complexed to the NAb 2-51. The NTD is colored in pink; the heavy and light chains of the NAb 2-51 are colored in gray and green, respectively. The insets show a close-up of the binding interface of the loops N3 and N5 interacting with the variable chains of the NAb 2-51. The N5 loop representation is also rotated 180° around its z-axis. Residues making contact in the interface are depicted in the licorice representation, with carbon

atoms in cyan, nitrogen atoms in blue and oxygen atoms in red. The dotted lines indicate the interacting residues-pair.

Table 2. Impact of indels on the binding between SARS-CoV-2 NTDs and NAb 2-51, expressed as loss of putative interactions.

Variant	ΔH-bond	ΔSalt-bridge	Δpi-stacking	ΔHydrophobic SASA [Å^2]	Native Contacts Lost (NTD - Ab)
P.1 Δ189-190	0	0	0	0	-
P.1-like ins214ANRN	0	0	0	0	-
B.1.1.28 Δ144	-2	-3	-1	-1	K147-E71 K150-E53 K150-D54 Y145-Y98
P.2 Δ144	-2	-3	-1	-104	K147-E71 K150-E53 K150-D54 Y145-Y98
P.1 Δ144	-2	-3	-1	-111	K147-E71 K150-E53 K150-D54 Y145-Y98
P.1 Δ141-144	-2	-3	-1	-313	K147-E71 K150-E53 K150-D54 Y145-Y98
B.1.1.33 Δ141-144 Δ256-258	-3	-3	-1	-439	Y147-E71 K150-E53 K150-D54 Y145-Y98 D253-S56 P251-P55 P251-L46 P251-Y100

Recent genomic findings are showing a sudden landscape change in SARS-CoV-2 evolution since October 2020, coinciding with the independent emergence of VOCs carrying multiple convergent amino acid replacements at the RBD of the S protein [32]. One hypothesis is that such a major selection pressure shift on the virus genome is driven by the increasing human population immunity worldwide acquired from natural SARS-CoV-2 infection. Our findings suggests that SARS-CoV-2 is continuously adapting in Brazil and that RDRs 2/4 variants here detected might have

evolved to escape from NAb against NTD supersite and could be even more resistant to neutralization than the parental P.1, P.2, and B.1.1.33(E484K) viruses. The sequential evolution steps observed in Brazil recapitulates the pattern observed in South Africa where the VOC B.1.351 first acquired key RBD mutations (E484K and N501Y) and some weeks later the NTD deletion Δ242-244 [7]. These findings highlight the urgent need to address the SARS-CoV-2 vaccines' efficacy towards those emergent SARS-CoV-2 variants and the risk of ongoing uncontrolled community transmission of SARS-CoV-2 in Brazil for the generation of more transmissible variants.Furthermore, the recurrent emergence of NTD ins214 variants in different SARS-CoV-2 lineages circulating in the Americas and Europe since November 2020 deserves further attention.

Methods

SARS-CoV-2 and SARS-CoV-2-related coronavirus (SC2r-CoV) sequences

Our genomic survey of SARS-CoV-2 positive samples sequenced by the Fiocruz COVID-19 Genomic Surveillance Network between 12th March 2020 and 28th February 2021 identified 11 sequences with mutations of concern in the RBD and indels in the NTD (**Appendix 1**). The SARS-CoV-2 genomes were recovered using Illumina sequencing protocols as previously described [33,34]. The FASTQ reads obtained were imported into the CLC Genomics Workbench version 20.0.4 (Qiagen A/S, Denmark), trimmed, and mapped against the reference sequence EPI_ISL_402124 available in EpiCoV database in the GISAID (https://www.gisaid.org/). The alignment was refined using the InDels and Structural Variants module. Additionally, the same reads were imported in a different pipeline [35] based on Bowtie2 and bcftools [36] mapping and consensus generation allowing us to further confirm the indels for all samples sequenced in this study. BAM files were used as input to generate sequencing coverage plots onto indels using the Karyoploter R package [37]. Sequences were combined with SARS-CoV-2 and SC2r-CoV from bats and pangolins available in the EpiCoV database in GISAID by 1st March 2021 (Appendix Table 1). This study was approved by the FIOCRUZ-IOC (68118417.6.0000.5248 and CAAE 32333120.4.0000.5190) and the Amazonas State University Ethics Committee (CAAE: 25430719.6.0000.5016) and the Brazilian Ministry of the Environment (MMA) A1767C3.

Maximum Likelihood Phylogenetic Analyses

SARS-COV-2 sequences here obtained were aligned with high quality (<1% of N) and complete (>29 kb) lineages B.1.1.28, P.1, P2 and B.1.1.33 sequences that were available

in EpiCoV database in the GISAID (https://www.gisaid.org/) at March 1st, 2021 and subjected to maximum-likelihood (ML) phylogenetic analysis using IQ-TREE v2.1.2 [38]. The S amino acid sequences from selected SARS-CoV-2 and SC2r-CoV lineages available in the EpiCoV database were also aligned using Clustal W [39] adjusted by visual inspection.

Structural Modeling

The resolved crystallographic structure of SARS-CoV-2 NTD protein bound to the neutralizing antibody 2-51 was retrieved from the Protein Databank (PDB) under the accession code 7L2C [28]. Missing residues of the chain A, corresponding to the NTD coordinates, were modeled using the user template mode of the Swiss-Model webserver (https://swissmodel.expasy.org/) [40] and was used as starting structure for the NTD wildtype. This structure was then used as a template to model the NTD variants using the Swiss-Model. The modeled structures of the NTDs variants were superimposed onto the coordinates of the PDB ID 7L2C to visualize the differences between the NTD-antibody binding interfaces. Image rendering was carried out using Visual Molecular Dynamics (VMD) software [41]. The NTD-antibodies complexes were geometry optimized using a maximum of 5,000 steps or until it reached a convergence value of 0.001 REU (Rosetta energy units) using the limited-memory BroydenFletcher-Goldfarb-Shanno algorithm, complying with the Armijo-Goldstein condition, as implemented in the Rosetta suite of software 3.12 [42]. Geometry optimization was accomplished through the atomistic Rosetta energy function 2015 (REF15), while preserving backbone torsion angles. Protein-protein interface analyses were performed using the Protein Interactions Calculator (PIC) webserver (http://pic.mbu.iisc.ernet.in/)[43], the 'Protein interfaces, surfaces and assemblies' service (PISA) at the European Bioinformatics Institute (https://www.ebi.ac.uk/pdbe/pisa/pistart.html) [44] and the InterfaceAnalyzer protocol of the Rosetta package interfaced with the RosettaScripts scripting language [45]. For the interfaceAnalyzer, the maximum SASA that is allowed for an atom to be defined as buried is 0.01 $Å^2$, with a SASA probe radius of 1.2 Å.

Epitope prediction

Epitopes in the NTD region were predicted by the ElliPro Antibody Epitope Prediction server [46]. NTD are shown as predicted linear epitopes when using PDB accession codes

6VXX [47] and 6VSB [48], (structural coordinates corresponding to the entire S protein), along with a minimum score of 0.9, *i.e.*, a highly strict criterion.

Acknowledgements

The authors wish to thank all the health care workers and scientists who have worked hard to deal with this pandemic threat, the GISAID team, and all the EpiCoV database's submitters, GISAID acknowledgment table containing sequences used in this study are attached to this post (**Appendix Table 3**). We also appreciate the support of the Fiocruz COVID-19 Genomic Surveillance Network (http://www.genomahcov.fiocruz.br/) members, the Respiratory Viruses Genomic Surveillance Network of the General Laboratory Coordination (CGLab), Brazilian Ministry of Health (MoH), Brazilian Central Laboratory States (LACENs) and the Amazonas surveillance teams for the partnership in the viral surveillance in Brazil. Financial support was provided by FAPEAM (PCTI-EmergeSaude/AM call 005/2020 and Rede Genômica de Vigilancia em Saúde - REGESAM); Ministério da Ciência, Tecnologia, Inovações e Comunicações/Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq/Ministério da Saúde - MS/FNDCT/SCTIE/Decit (grants 402457/2020-9 and 403276/2020-9); Inova Fiocruz/Fundação Oswaldo Cruz (Grants VPPCB-007-FIO-18-2-30 and VPPCB-005-FIO-20-2-87) and INCT-FCx (465259/2014-6). Computer allocation was partly granted by the Brazilian National Scientific Computing Center (LNCC). FGN, GLW, RDL and GB are supported by the CNPq through their productivity research fellowships (306146/2017-7, 303902/2019-1, 425997/2018-9 and 302317/2017-1 respectively). G.B. is also funded by the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ (Grant number E-26/202.896/2018).

References

1 Avanzato VA, Matson MJ, Seifert SN *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* 2020; **183**: 1901-1912.e9.

2 Choi B, Choudhary MC, Regan J *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med* 2020; **383**: 2291–2293.

3 McCarthy KR, Rennick LJ, Nambulli S *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* 2021; **6950**: 6–6.
4 Kemp SA, Collier DA, Datir RP *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* 2021; : 1–10.
5 Spike protein mutations in novel SARS-CoV-2 'variants of concern' commonly occur in or near indels. Virological. 2021.https://virological.org/t/spike-protein-mutations-in-novel-sars-cov-2-variants-of-concern-commonly-occur-in-or-near-indels/605 (accessed 14 Mar2021).
6 Spike protein sequences of Cambodian, Thai and Japanese bat sarbecoviruses provide insights into the natural evolution of the Receptor Binding Domain and S1/S2 cleavage site. Virological. 2021.https://virological.org/t/spike-protein-sequences-of-cambodian-thai-and-japanese-bat-sarbecoviruses-provide-insights-into-the-natural-evolution-of-the-receptor-binding-domain-and-s1-s2-cleavage-site/622 (accessed 14 Mar2021).
7 Tegally H, Wilkinson E, Giovanetti M *et al.* Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature* 2021; : 1–8.
8 Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. Virological. 2020.https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed 14 Mar2021).
9 Wang R, Zhang Q, Ge J *et al.* Spike mutations in SARS-CoV-2 variants confer resistance to antibody neutralization. *bioRxiv* 2021; : 2021.03.09.434497.
10 Wang P, Nair MS, Liu L *et al.* Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. *bioRxiv* 2021; : 2021.01.25.428137.
11 Collier DA, De Marco A, Ferreira IATM *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* 2021; : 1–8.
12 Gobeil S, Janowska K, McDowell S *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure, conformation and antigenicity. *bioRxiv* 2021; : 2021.03.11.435037.
13 Wang P, Wang M, Yu J *et al.* Increased Resistance of SARS-CoV-2 Variant P.1 to Antibody Neutralization. *bioRxiv* 2021; : 2021.03.01.433466.
14 McCallum M, Marco AD, Lempp FA *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* 2021; **0**. doi:10.1016/j.cell.2021.03.028.
15 Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. Virological. 2021.https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (accessed 14 Mar2021).
16 Voloch CM, da Silva Francisco R, de Almeida LGP *et al.* Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J Virol* 2021. doi:10.1128/JVI.00119-21.
17 Resende PC, Gräf T, Paixão ACD *et al.* A potential SARS-CoV-2 variant of interest (VOI) harboring mutation E484K in the Spike protein was identified within lineage B.1.1.33 circulating in Brazil. *bioRxiv* 2021; : 2021.03.12.434969.
18 Tosta S, Giovanetti M, Nardy VB *et al.* Early genomic detection of SARS-CoV-2 P.1 variant in Northeast Brazil. *medRxiv* 2021; : 2021.02.25.21252490.
19 Wacharapluesadee S, Tan CW, Maneeorn P *et al.* Evidence for SARS-CoV-2 related

coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun* 2021; **12**: 972.

20 Liu L, Wang P, Nair MS *et al.* Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* 2020; **584**: 450–456.

21 Voss WN, Hou YJ, Johnson NV *et al.* Prevalent, protective, and convergent IgG recognition of SARS-CoV-2 non-RBD spike epitopes in COVID-19 convalescent plasma. *bioRxiv* 2020; : 2020.12.20.423708.

22 Piccoli L, Park Y-J, Tortorici MA *et al.* Mapping Neutralizing and Immunodominant Sites on the SARS-CoV-2 Spike Receptor-Binding Domain by Structure-Guided High-Resolution Serology. *Cell* 2020; **183**: 1024-1042.e21.

23 Barnes CO, West AP, Huey-Tubman KE *et al.* Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* 2020; **182**: 828-842.e16.

24 Greaney AJ, Loes AN, Crawford KHD *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* 2021; **29**: 463-476.e6.

25 Hoffmann M, Arora P, Groß R *et al.* SARS-CoV-2 variants B.1.351 and B.1.1.248: Escape from therapeutic antibodies and antibodies induced by infection and vaccination. *bioRxiv* 2021; : 2021.02.11.430787.

26 Baum A, Fulton BO, Wloga E *et al.* Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science* 2020; **369**: 1014–1018.

27 Nelson G, Buzko O, Spilman P, Niazi K, Rabizadeh S, Soon-Shiong P. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *bioRxiv* 2021; : 2021.01.13.426558.

28 Cerutti G, Guo Y, Zhou T *et al.* Potent SARS-CoV-2 Neutralizing Antibodies Directed Against Spike N-Terminal Domain Target a Single Supersite. *bioRxiv* 2021; : 2021.01.10.426120.

29 Chi X, Yan R, Zhang J *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* 2020; **369**: 650–655.

30 Wibmer CK, Ayres F, Hermanus T *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat Med* 2021; : 1–4.

31 Andreano E, Piccini G, Licastro D *et al.* SARS-CoV-2 escape in vitro from a highly neutralizing COVID-19 convalescent plasma. *bioRxiv* 2020; : 2020.12.28.424451.

32 Martin DP, Weaver S, Tegally H *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv* 2021; : 2021.02.23.21252268.

33 Nascimento VA do, Corado A de LG, Nascimento FO do *et al.* Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. *Mem Inst Oswaldo Cruz* 2020; **115**. doi:10.1590/0074-02760200310.

34 Resende PC, Motta FC, Roy S *et al.* SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. *bioRxiv* 2020; : 2020.04.30.069039.

35 Paiva MHS, Guedes DRD, Docena C *et al.* Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. *Viruses* 2020; **12**: 1414.

36 Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.

Bioinforma Oxf Engl 2011; **27**: 2987–2993.

37 Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 2017; **33**: 3088–3090.

38 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 2015; **32**: 268–274.

39 Larkin M a, Blackshields G, Brown NP *et al.* Clustal W and Clustal X version 2.0. *Bioinforma Oxf Engl* 2007; **23**: 2947–8.

40 Waterhouse A, Bertoni M, Bienert S *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018; **46**: W296–W303.

41 Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J Mol Graph* 1996; **14**: 33–38.

42 Leaver-Fay A, Tyka M, Lewis SM *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 2011; **487**: 545–574.

43 Tina KG, Bhadra R, Srinivasan N. PIC: Protein Interactions Calculator. *Nucleic Acids Res* 2007; **35**: W473-476.

44 Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* 2007; **372**: 774–797.

45 Fleishman SJ, Leaver-Fay A, Corn JE *et al.* RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLOS ONE* 2011; **6**: e20161.

46 Ponomarenko J, Bui H-H, Li W *et al.* ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* 2008; **9**: 514.

47 Wrapp D, Wang N, Corbett KS *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020; **367**: 1260–1263.

48 Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 2020; **181**: 281-292.e6.