Deep phylogenetic-based clustering analysis uncovers new and shared mutations in SARS-CoV-2 variants as a result of directional and convergent evolution

Danilo Rosa Nunes[1], Carla Torres Braconi*[1], Louisa F. Ludwig-Begall[2], Clarice Weis Arns[3] and Ricardo Durães-Carvalho*[1,3]

[1]Department of Microbiology, Immunology and Parasitology, Paulista School of Medicine, Federal University of São Paulo, São Paulo-SP, Brazil

[2]Veterinary Virology and Animal Viral Diseases, Department of Infectious and Parasitic Diseases, FARAH Research Centre, Faculty of Veterinary Medicine, University of Liège, Belgium.

[3]Laboratory of Virology, University of Campinas, Campinas-SP, Brazil.

*Corresponding authors: ctbsantos@unifesp.br (CTB) and rdcarval@gmail.com (RD-C).

Abstract

Nearly two decades after the last epidemic caused by a severe acute respiratory syndrome coronavirus (SARS-CoV), newly emerged SARS-CoV-2 quickly spread in 2020 and precipitated an ongoing global public health crisis. Both the continuous accumulation of point mutations, owed to the naturally imposed genomic plasticity of SARS-CoV-2 evolutionary processes, as well as viral spread over time, allow this RNA virus to gain new genetic identities, spawn novel variants and enhance its potential for immune evasion. Here, through an in-depth phylogenetic clustering analysis of upwards of 200,000 whole-genome sequences,

we reveal the presence of not previously reported and hitherto unidentified mutations and recombination breakpoints in Variants of Concern (VOC) and Variants of Interest (VOI) from Brazil, India (Beta, Eta and Kappa) and the USA (Beta, Eta and Lambda). Additionally, we identify sites with shared mutations under directional evolution in the SARS-CoV-2 Spike-encoding protein of VOC and VOI, tracing a heretofore-undescribed correlation with viral spread in South America, India and the USA. Our evidence-based analysis provides well-supported evidence of similar pathways of evolution for such mutations in all SARS-CoV-2 variants and sub-lineages. This raises two pivotal points: the co-circulation of variants and sub-lineages in close evolutionary environments, which sheds light onto their trajectories into convergent and directional evolution (i), and a linear perspective into the prospective vaccine efficacy against different SARS-CoV-2 strains (ii).

Author summary

In this study, through analysis of very robust and comprehensive datasets, we identify a plethora of mutations in the SARS-CoV-2 Spike cell surface protein of several variants of concern and multiple variants of interest. We trace an association of such mutations with viral spread in different countries. We further infer the presence of new SARS-CoV-2 sublineages and show that the vast majority of mutations identified in the SARS-CoV-2 Spike protein are under convergent evolution. If we consider every color of a Rubik´s cube's face to represent a different mutation of a particular variant, evolutionary convergence can be achieved only when all composite pieces of a single face are of the same color and every face has one unique color. Overall, this raises two important points: we provide insight into the presence of SARS-CoV-2 variants and sub-lineages circulating in very close

evolutionary environments and our analyses can serve to facilitate an outlook into the prospective vaccine efficacy against different SARS-CoV-2 strains.

Introduction

In the last two decades, human health has been threatened by the emergence of three important zoonotic and pathogenic betacoronaviruses, namely the severe acute respiratory syndrome coronavirus (SARS-CoV) (Guan et al., 2003), the Middle East respiratory syndrome coronavirus (MERS-CoV) (Zaki et al., 2012) and, most recently, the causative agent of the Coronavirus Disease 2019 (COVID-19) pandemic, SARS-CoV-2 (da Costa, Moreli, and Saivish 2020). Likely originated from bats, pandemic SARS-CoV-2, like other endemic human alpha- (NL63 and 229E) and beta- (OC43 and HKU1) CoVs known for causing upper respiratory tract infections, overcame the interspecies barrier as a result of spillover and/or recombination events, and gained a pervasive ability to rapidly infect and spread around the globe (Corman et al., 2018; Boni et al., 2020; V'kovski et al., 2021).

The COVID-19 pandemic precipitated an intense genomic surveillance via data depositories and sequencing platforms and led to an unprecedented accumulation of public genomic data concerning a human pathogenic virus (Boni et al., 2020**;** Munnink et al., 2021). The sheer amount of available sequencing data has the potential to facilitate higher-precision micro-evolutionary analyses mapping escape and point mutations in presumed positively selected sites and residues putatively associated to an increased virus fitness and pathogenesis and allows inferences concerning the dynamics of SARS-CoV-2 spread (Kosakovsky Pond et al., 2008; Alteri et al., 2021).

Although the analysis of micro-evolutionary mechanisms is of paramount importance and may provide powerful information to promote the prediction of vaccination perspectives and the tracing of SARS-CoV-2 epidemiological chains, there is as yet a lack of data-based investigations examining the presence of eventual shared mutations and their evolutionary characteristics in classified SARS-CoV-2 Variants of Concern (VOC) and Variants of Interest (VOI) (CDC 2021a; Peacock et al., 2021).

Given the importance of monitoring mutations to track the emergence of novel variants, here we investigate the influence of directional selection and the dynamics of SARS-CoV-2 genomic plasticity in VOC and VOI by clustering partition high-scale phylogenetic and directional evolution (DEPS) approaches. Additionally, we show the presence of several mutations common for both VOI/VOC and convergently emerged sub-lineages, and provide a perspective of possible effects on the vaccination efficacy and the ongoing COVID-19 pandemic.

Methods

Sequence data and filtering strategy

High-coverage and complete HCoV-229E and HCoV-NL63 (alpha-CoVs), HCoV-OC43, HCoV-HKU1, MERS-CoV, SARS-CoV and SARS-CoV-2 VOC and VOI (beta-CoVs) genome sequences (≥ 29,000 bp), sampled from humans, were retrieved from the Global Initiative on Sharing Avian Influenza Data-EpiCoV (GISAID-EpiCoV) and GenBank databases at different times: February 12th (MERS-CoV, SARS-CoV and SARS-CoV-2), July 12th (HCoV-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1 and SARS-CoV-2) and August 26th 2021 (SARS-CoV-2), totalling 238,990 sequences. With regards to SARS-CoV-2, we particularly focused

on strains of countries from South America, China, India, and the United States of America (USA). At the time of analysis, India, the USA, and Brazil had reported the largest numbers of cumulative confirmed COVID-19 cases and deaths. This approach was used to compare putative mutual sites and residue changes under directional evolution over time.

Subsequently, sequences were filtered via Sequence Cleaner, a biopython-based program, utilising the following script: sequence_cleaner -q INPUT_DIRECTORY -o OUTPUT_DIRECTORY -ml 29,000 (MINIMUM_LENGTH) -mn 0 (PERCENTAGE_N) --remove_ambiguous. The outcome was a set of unambiguous sequences equal to and greater than 29,000 pb with zero percent of unknown nucleotides. Next, the datasets were aligned by adding coding-sequences related to references for HCoV-229E (NC_002645.1), HCoV-NL63 (NC_005831.2), HCoV-OC43 (NC_006213.1), HCoV-HKU1 (NC_006577.2), MERS-CoV (NC_038294.1), SARS-CoV (NC_004718.3), and SARS-CoV-2 (NC_045512.2), using default settings, with the rapid calculation of full-length multiple sequence alignment of closely-related viral genomes (MAFFT v.7 web-version program; https://mafft.cbrc.jp/alignment/software/closelyrelatedviralgenomes.html) and were edited by the UGENE v.38.1 (Okonechnikov et al., 2012).

Clustering and sub-clustering analysis

A methodological approach to extract large-scale phylogenetic partitions was applied to identify transmission cluster chains on the largest Maximum Likelihood (ML) phylogenetic trees of the SARS-CoV-2 variants on the basis of a depth-first search algorithm which unifies evaluation of node reliability, tree topology and patristic distance (Prosperi et al., 2011). The ML tree was implemented in FastTree

v.2.1.7 by using the standard implementation General Time Reversible (GTR) plus CAT with 20 gamma distribution parameters and a mix of Nearest-Neighbor Interchanges (NNI) and Sub-Tree-Prune-Regraft (SPR) (Price, Dehal, and Arkin 2010). Thereafter, in view of identifying SARS-CoV-2 cluster transmission events, we first selected sequences (one per cluster) from nodes/sub-trees with ≥ 2 distinct individuals and ≥ 90% reliability of statistical support (Shimodaira-Hasegawa test), where initially the patristic distance was adjusted to find a representative number of clusters (n= 100) from each large reconstructed ML tree. In addition to this strategy, a second approach included sub-clustering analysis as an indirect way to infer and investigate the possibility of co-circulating sub-lineages. For this, we selected sequences (two per cluster) with ≥ 95% node reliability of statistical support from a threshold of 0.05, thus corresponding to the 5th percentile when considering the whole-tree patristic distance distribution.

Recombination and directional evolution analyses

Before proceeding to directional evolution analysis, all datasets were submitted to the Genetic Algorithm for Recombination Detection (GARD), a likelihood-based tool to pinpoint recombination breakpoints (Kosakovsky Pond et al., 2006). To double check the outcome of the first of the two strategies described above, an additional test was conducted using the Pairwise Homoplasy Index (PHI; default settings) (Huson and Bryant 2005). Evidence-based analysis through phylogenetic maximum-likelihood was then performed implementing the Datamonkey web-server and the program Hyphy v.2.5 to track directional selection in amino acid sequences (DEPS) (Kosakovsky Pond et al., 2020). The DEPS method identifies both the residue and sites evolving toward it with great accuracy and detects frequency-dependent selection-scenarios as well as selective sweeps

and convergent evolution that can confound most existing tests (Kosakovsky Pond et al., 2008). Further, the DEPS method has shown better performance than (traditional) substitution rate-based analyses (dN/dS) in detecting transient and frequency-dependent selection and directionally evolving sites and residues. For the most part, a Beta-Gamma site-to-site rate variation was used to conduct the analysis. The best-fit protein substitution model was chosen according to the corrected Akaike Information Criterion (cAIC). Only target sites and residues with Empirical Bayes Factors for evidence in favour of a directional selection model equal to or greater than 100 were considered for further exploration. Certain randomly chosen datasets were run multiple times (more than eight) to confirm obtained results.

Statistical analysis

Data pertaining to SARS-CoV and MERS-CoV-related cases and deaths were extracted from the National Health Service (NHS, UK) (https://www.nhs.uk/conditions/sars/) and European Centre for Disease Prevention and Control (ECDC) (https://www.ecdc.europa.eu/en/publications-data/distribution-confirmed-cases-mers-cov-place-infection-and-month-onset-1), respectively. Information concerning SARS-CoV-2 was collected from World Health Organization (WHO) (https://covid19.who.int/). Population demographic data were retrieved from the Our World in Data website (https://ourworldindata.org/grapher/world-population-by-world-regions-post-1820?tab=table&country=Oceania~~North+America~Europe~Africa~Asia).

Statistical analyses were performed using one-way analysis of variance (ANOVA) and nonparametric methods followed by *post hoc* Kruskal-Wallis and Friedman (both with Dunn's Multiple Comparison), and Bartlett's tests (Tukey's, Newman-Keuls and Bonferroni’s multiple comparisons). Additionally, Mann Whitney

and Wilcoxon matched-pairs signed-rank (T test) and Pearson/Spearman (Correlation), all one-tailed methods with 99% confidence interval (CI), were run. *P-values* equal to or less than 0.005 ($p \leq 0.005$, SARS-CoV-2 from South America: DEPS [sites and residues] vs infections) and 0.05 ($p \leq 0.05$, SARS-CoV-2 from Brazil, China, India and the USA: DEPS [residues] vs circulating variants and infections) were considered as statistically significant. Data analyses were carried out using GraphPad Prism v. 5.01 (GraphPad Software, San Diego, California, USA). Figures and data visualization were performed using the ggplot2 v.3.3.5 package in the R (RStudio v.1.4.1717) language environment. Final graphics were edited with the open-source software drawing tool Inkscape v.1.0.2.

Results and discussion

Recombination is known to be a crucial evolutionary process for many RNA viruses (Lai 1992; Lemey, Salemi, and Vandamme 2009; Su et al., 2016); the process is frequently observed in the *Coronaviridae* family where recombination is likely facilitated by discontinuous transcription involving jumps of the replication-transcription complex during minus strand RNA synthesis. However, the consequences of recombination events occurring in the context of the current SARS-CoV-2 evolutionary landscape are still speculative (Li et al., 2020; Singh and Yi 2021; Pollett et al., 2021). Here we address this knowledge gap, revealing the presence of recombination and shared mutations in the SARS-CoV-2 Spike-encoding protein, demonstrating them to be under directional and convergent evolution amongst SARS-CoV-2 VOC/VOI and sub-lineages, and tracing an interconnection with viral spread. First, endemic and epidemic human coronaviruses (HCoVs) were compared to identify similar evolutionary patterns that could help clarify the evolution of SARS-CoV-2. An initial recombination breakpoint analysis showed that four of six HCoVs

analyzed presented such signals (Fig. 1A). Endemic viruses OC43, NL63 and HKU1 also showed a similar pattern of residue accumulation and directional evolution, despite these viruses being subject to differing selective pressures (Forni et al., 2021).

A subsequent comparison of SARS-CoV-2 to the other two pathogenic HCoVs (SARS-CoV and MERS-CoV), highlighted differences in the number of directionally-evolving sites and residues (Fig. 1B). These patterns, putatively reflecting the initial evolutionary paths of the individual viruses, may suggest that SARS-CoV was initially under lower positive evolutionary pressure than MERS-CoV and SARS-CoV-2. In turn, deletions and mutations acquired by SARS-CoV have been shown to have had an impact on adaptation to human-to-human transmission, modifying both the capacity for viral proliferation and profiles of pathogenesis (Muth et al., 2018; Pereira 2020; Pereira 2021).

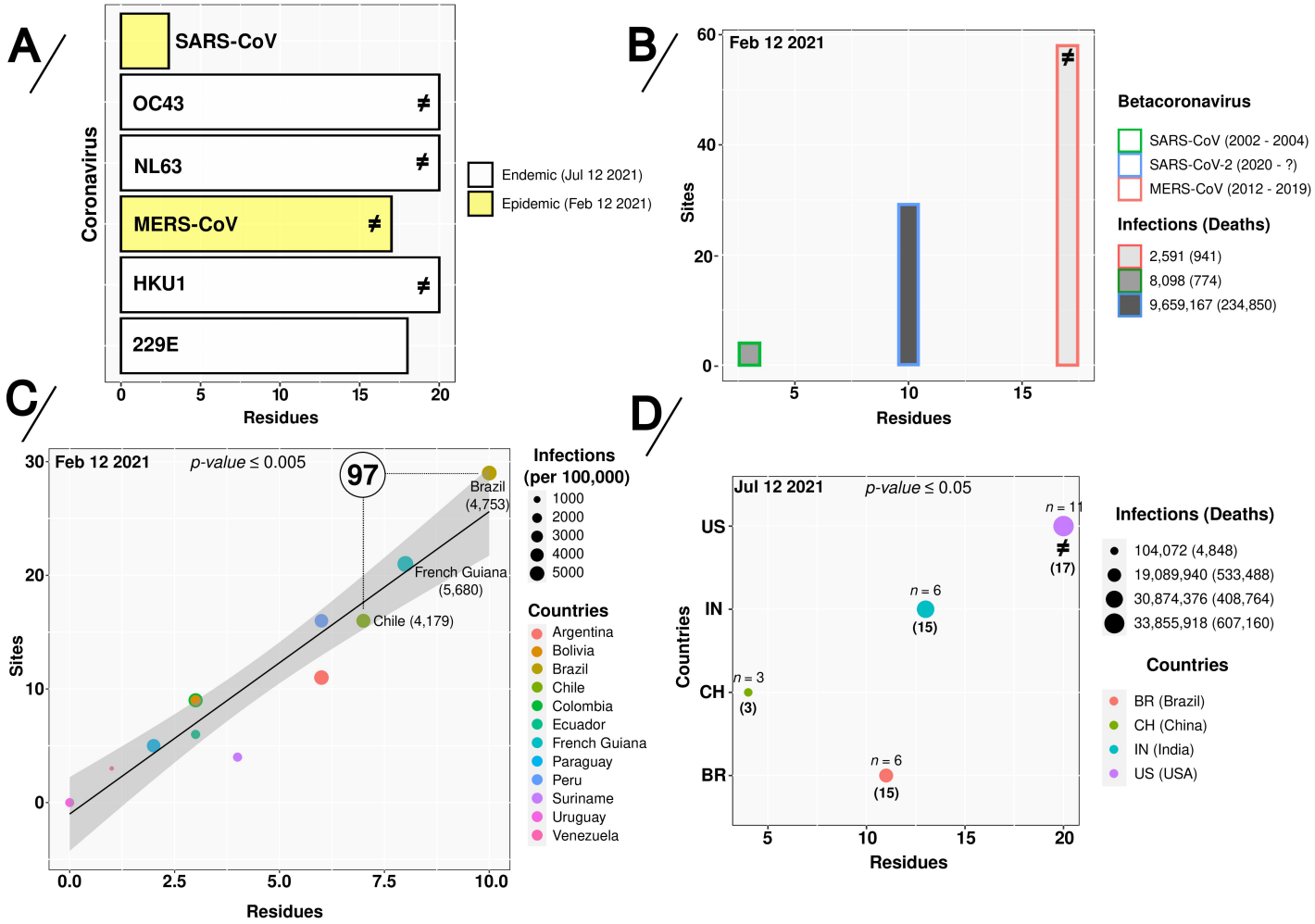

Figure 1. Directionally-evolving sites and residues in alpha- (229E and NL63) and beta- (OC43, HKU1, SARS-CoV, MERS-CoV and SARS-CoV-2) coronavirus sequences. (**A**) Directed-evolving residues in six different endemic and epidemic human coronaviruses (HCoVs), (**B**) Sites and residues vs infections in epidemic and pandemic coronaviruses (CoVs). (**C**) Linear-regression curve and co-relationship on the absolute amount of sites and/or residues under directional-positive selection given SARS-CoV-2 infections per 100,000 people from South America countries and (**D**) the total number of SARS-CoV-2 infections in Brazil, China, India and the USA. In panels A and B, the symbol ≠ represents the presence of recombination breakpoints signals. In panel C, the number inside the circle represents the amount

of clusters found in Brazil and Chile. In panel D, *n* represents the amount of SARS-CoV-2 variants and the numbers in parentheses indicate sites under directional and convergent evolution in the Spike-encoding protein. Colors and symbols used in the panels are defined in the legend to the right of the figure.

The evolution of SARS-CoV-2 was initially marked by genetic drift in a typical process of neutral evolution (Dearlove et al., 2020; MacLean et al., 2021); the virus reached a large number of new and susceptible hosts and, although some mutations appeared along the genome, there was no significant shift (Martin et al., 2021). However, as SARS-CoV-2 spread (Yadav et al., 2020; Hodcroft et al., 2021), fitness changes resulting from mutations in the viral genome as well as the emergence of new variants were increasingly reported (Velazquez-Salinas et al., 2020; Zhang et al., 2020; Plante et al., 2021).

The first epidemic wave of SARS-CoV-2 severely affected most countries in South America as a probable result of multiple viral introductions (Candido 2020); rapid increases of case numbers were especially reported in Brazil, the biggest and most populous country in Latin America (Paiva et al., 2020; Stefanelli et al., 2020). The uncontrolled viral spread created a favorable scenario for the emergence of new variants (Voloch et al., 2021; Faria et al., 2021; Resende et al., 2021; Sabino et al., 2021). To identify the impact of directional-positive selection sites at the rate of infections under these particular conditions, we traced the evolutionary scenario of SARS-CoV-2 in South America (via analysis of a significant and representative amount of genome sequences).

Remarkably, our data showed that an increase of DEPS was correlated with viral spread dynamics, with Brazil exhibiting a lower proportion of COVID-19 cases when compared to French Guiana and the same amount of SARS-CoV-2 clusters inferred in Chile (n= 97) (Fig. 1C), probably due to a higher diversity of circulating viruses. Our results also highlighted a series of mutations; while certain mutations have previously been described, but have hitherto remained unidentified in SARS-CoV-2 VOC and VOI, multiple further mutations are identified for the first time in this study (Table 1).

Table 1: Mutational landscape of SARS-CoV-2 Spike protein VOC and VOI based on the WHO label

Sequences collection date	Inferred substitutions (Spike location) Obs.: new mutations are underlined Convergent evolution: ●	SARS-CoV-2 carrying this mutation (from WHO)	Additional SARS-CoV-2 variants carrying this mutation (from this study)	Empirical Bayes Factors	Time interval (months)
Feb 12th 2021	L18F ● (NTD)	Beta and Gamma	Alpha	$>10^5$	5
	T20N ● (NTD)	Gamma	Alpha	$>10^5$	
	P26S/P26L* ● (NTD)	Gamma	Alpha and Epsilon/Zeta*	2129.2	
	D138H/D138Y* ● (NTD)	Gamma	Alpha and Epsilon/Delta*	$>10^5$	
	R190S ● (NTD)	Gamma	Delta	9869.0	
	K417T ● (RBD)	Gamma	-	1966.6	
	E484K/E484Q ● (RBD)	Beta, Gamma, Eta, Iota, Kappa, Theta and Zeta	-	$>10^5$	
	N501Y ● (RBD)	Alpha, Beta, Gamma, Mu and Theta	Eta, Kappa and Lambda	$>10^5$	
	T1027I ● (CH)	Gamma	-	$>10^5$	
	S13I ● (SP-NTD)	Epsilon	Alpha	166.9	
	R21I/R21T* ● (NTD)	-	Gamma/Epsilon*	168.0	
	R34L/R34P* (NTD)	-	Unsigned/Eta*	262.6	

Jul 12th 2021 Obs.: without Delta variant	S50L (NTD)	-	unsigned	124.0	0
	L54F ● (NTD)	-	Gamma	>10^5	
	W152L*/W152C (NTD)	Epsilon	Gamma*	226.2	
	S255F ● (NTD)	-	Gamma, Delta and Kappa	160.0	
	N501Y ● (RBD)	Alpha, Beta, Gamma, Mu and Theta	Eta, Kappa and Lambda	428.8	
	A570D ● (CT1)	Alpha	Eta, Kappa and Lambda	>10^5	
	P681H (CT2)	Alpha, Mu and Theta	Gamma and Lambda	>10^5	
	A688V ● (S1/S2)	-	Alpha, Gamma and Zeta	113.4	
	T716I ● (S1/S2)	Alpha	Epsilon	2054.1	
	D1118H/D1118Y* (CD1)	Alpha	Lambda/Zeta*	>10^5	
	C1235F ● (CTail)	-	unsigned	317.0	
	S13I ● (SP-NTD)	Epsilon	Alpha	137.1	
	T19R/T19I* ● (NTD)	Delta	Eta*	>10^5	
	R21I/R21T* ● (NTD)	-	Gamma/Epsilon*	138.2	
	R34L (NTD)	-	unsigned	241.4	
	L54F ● (NTD)	-	Gamma	327.6	
	G142D/G142S* ● (NTD)	Delta and Kappa	Zeta*	>10^5	
	W152L (NTD)	-	Gamma	201.2	

	R237M ● (NTD)	-	unsigned	418.7
Jul 12th 2021 Obs.: with Delta variant	L452R/L452Q ● (RBD)	Delta, Epsilon, Iota, Lambda and Kappa	-	$>10^5$
	T478K ● (RBD)	Delta	-	$>10^5$
	E484K (RBD)	Beta, Gamma, Eta, Iota, Kappa, Theta and Zeta	-	$>10^5$
	N501Y ● (RBD)	Alpha, Beta, Gamma, Mu and Theta	Eta, Kappa and Lambda	$>10^5$
	A570D ● (CT1)	Alpha	Eta, Kappa and Lambda	$>10^5$
	D574Y ● (CT1)	-	unsigned	236.3
	P681R/P681H* ● (CT2)	Alpha, Delta, Kappa, Mu and Theta	Gamma* and Lambda*	$>10^5$
	T716I ● (S1/S2)	Alpha	Epsilon	1013.4
	D936Y● (HR1)	-	Gamma and Kappa	236.6
	S982A ● (HR1)	Alpha	-	9852.5
	D1118H ● (CD1)	Alpha	Lambda	$>10^5$
	D1163G ● (HR2)	-	Gamma	237.4

Obs.: R158- and G142- deletions were also found in the Delta and Theta SARS-CoV-2 variants, respectively. NTD, N-terminal domain; RBD, receptor binding domain; CD1, connector domain 1; CH, center helix; CT1, C-terminal domain 1; CT2, C-terminal domain 2; CTail, cytoplasmic tail; HR1, heptad repeat 1; HR2, heptad repeat 2; S1/S2, cleavage site and SP-NTD, Signal peptide-N-terminal domain.

Sources: CDC (https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html) and ECDC (https://www.ecdc.europa.eu/en/covid-19/variants-concern).

Spike mutations such as E484K, N501Y, L452R, S13I and W152C, seem to be fundamentally important in the process of adaptation of SARS-CoV-2 to human hosts, this by enhancing the affinity to the human ACE2 receptor and mediating immune system evasion (McCallum et al., 2021a; Greaney et al., 2021; Harvey et al., 2021). Our analyzes allowed us to follow SARS-CoV-2 spread dynamics over time in Brazil, showing an increasing number of sites under DEPS, primarily in the Spike-encoding protein. Nine sites are highlighted prior to February 2021, followed by fourteen sites until July 2021 (SARS-CoV-2 Delta variant not included). With the introduction of the Delta variant, both the presence of recombination signals as well as an increase of sites under DEPS were detected (Table 1, Table 2 and Fig. 1), allowing for inferences concerning a SARS-CoV-2 reproductive number increase. An increase in virus circulation augments the chance of viral coinfection, which in turn (and as a pre-requisite for recombination), can heighten the risk of emergence of new variants (Haddad et al., 2020; Ritchie et al., 2021; CDC 2021b).

The Delta variant, first identified in late 2020 in India as B.1.617.2 (Kirola 2021), harbors a constellation of non-synonymous mutations in the Spike protein (McCallum et al., 2021b) and has become the leading VOC worldwide. By the end of July 2021, this VOC accounted for 90% of all sequenced samples (Lamarca et al., 2021; PAHO, 2021). Brazil, India and the USA, the countries most severely affected by the pandemic, are now once again threatened by this highly contagious variant. Analysis of the molecular evolution of SARS-CoV-2 taking into account the influence of local demography in these specific scenarios has the potential to generate important insights into the spread and infection dynamics of this pathogen.

Using SARS-CoV-2 sequences from China (the most populated country in the world) as reference, we analyzed all datasets from Brazil, India, and the USA via a

large-scale phylogenetic partitions analysis (Prosperi et al., 2011; Matsuda, Suzuki, and Ogata 2020). Increases in SARS-CoV-2 infections were observed to be proportional to locally circulating variants and were not (in the scenarios analyzed), correlated with any particular demography (Fig. 1D); this indirectly reinforces the importance of measures implemented to avoid viral propagation. Analysis of phylogenetic partition clusters along the length of the circa 30 kb CoV genome evidenced several directionally-evolving sites under convergent evolution (Table 2). Thus, a possible association between the rate of infections and the number of residues as well as sites in the Spike-encoding protein under DEPS can be established (Fig. 1D). Interestingly, this supports a hypothesis of convergent evolution due to repeated and multiple site-specific substitutions in distinct SARS-CoV-2 VOC and VOI (see Table 1 and Table 2).

Table 2: Directional evolution landscape on SARS-CoV-2 variants/lineages circulating in Brazil, China, India and the USA

SARS-CoV-2 from	Directionally selected residues (proteome)	Directionally selected sites (proteome)	Most prevalent selection kind in SARS-CoV-2 Spike protein	Recombination breakpoints?
Brazil (without Delta variant)	11	89	Convergent evolution/Repeated Substitutions (10/15 sites)	No
Brazil (with Delta variant)	16	96	Convergent evolution/Repeated Substitutions (20/22 sites)	Yes
China	5	19	Convergent evolution/Repeated Substitution (3/3 sites)	No
India	13	74	Convergent evolution/Repeated Substitution (15/15 site)	No
USA	20	491	Convergent evolution/Repeated Substitution (12/17 sites)	Yes

Additionally, we also inferred the possible appearance of SARS-CoV-2 sub-lineages and traced the influence of an environment favoring directional evolution acting on SARS-CoV-2 variants. We showed different patterns among sites in the VOC and VOI, with a particular emphasis on the Kappa VOI currently circulating in the USA. We further demonstrated recombination among SARS-CoV-2 VOC and VOI from India (Beta, Eta and Kappa) and the USA (Beta, Eta and Lambda) (Fig. 2A and Table 2).

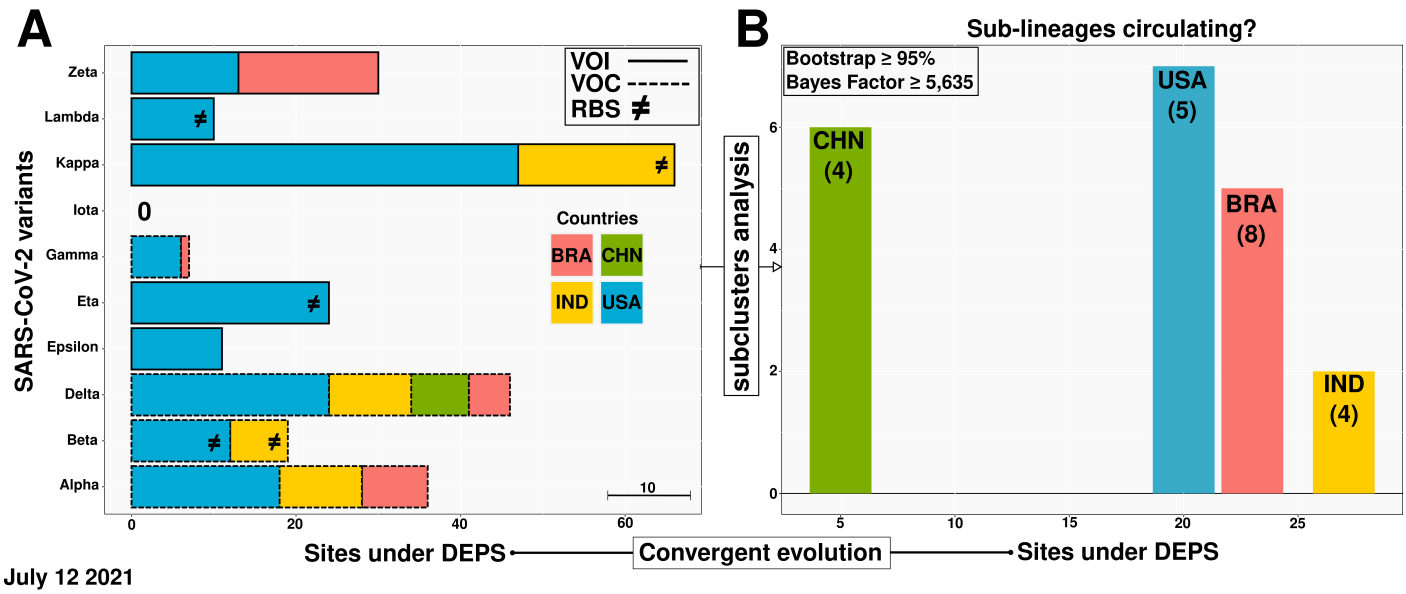

Figure 2. Transmission clustering and sub-clustering analyses on SARS-CoV-2 Variants of Concern (VOC) and Variants of Interest (VOI) sequences. (**A**) Model-based phylogenetic Maximum Likelihood (ML) method to infer transmission clustering and directional evolution in the VOC (dashed line) and VOI (non dashed line), and (**B**) sub-clustering and sub-lineage inferences in strains circulating in Brazil, China, India and the USA. Each color represents a particular country. RBS stands for recombination breakpoints signal (≠) and the scale bar shows the proportion of ten sites under positive selection (A). The numbers in parentheses indicate Spike-encoding protein sites under directional and convergent evolution (B).

As one of the first countries in the world to develop efficient immunizations and implement a vaccination policy (FDA, 2021), the USA vaccinated more than 30% of its population by April 2021. By September 2021, 60% of the booster-immunized population possessed neutralizing antibodies against several viral variants (Ritchie et al., 2021; Pegu et al., 2021). Similar outcomes were observed following widespread vaccination with various SARS-CoV-2 vaccines (different technologies leveraged for vaccine production) in many other regions, including South America and India (Li et al., 2021; Bernal et al., 2021) (Fig. 3). Nonetheless, viral circulation in the face of incomplete immunization has been described as one of the probable causes of the emergence of new variants (Sabino et al., 2021). Accordingly, our own analysis identified SARS-CoV-2 VOC and VOI subclusters (Fig. 2B), thus indicating co-circulation of variants and sub-lineages under convergent evolution. Surprisingly, the same evolutionary pattern was also observed for other endemic and epidemic CoVs studied (see Data availability).

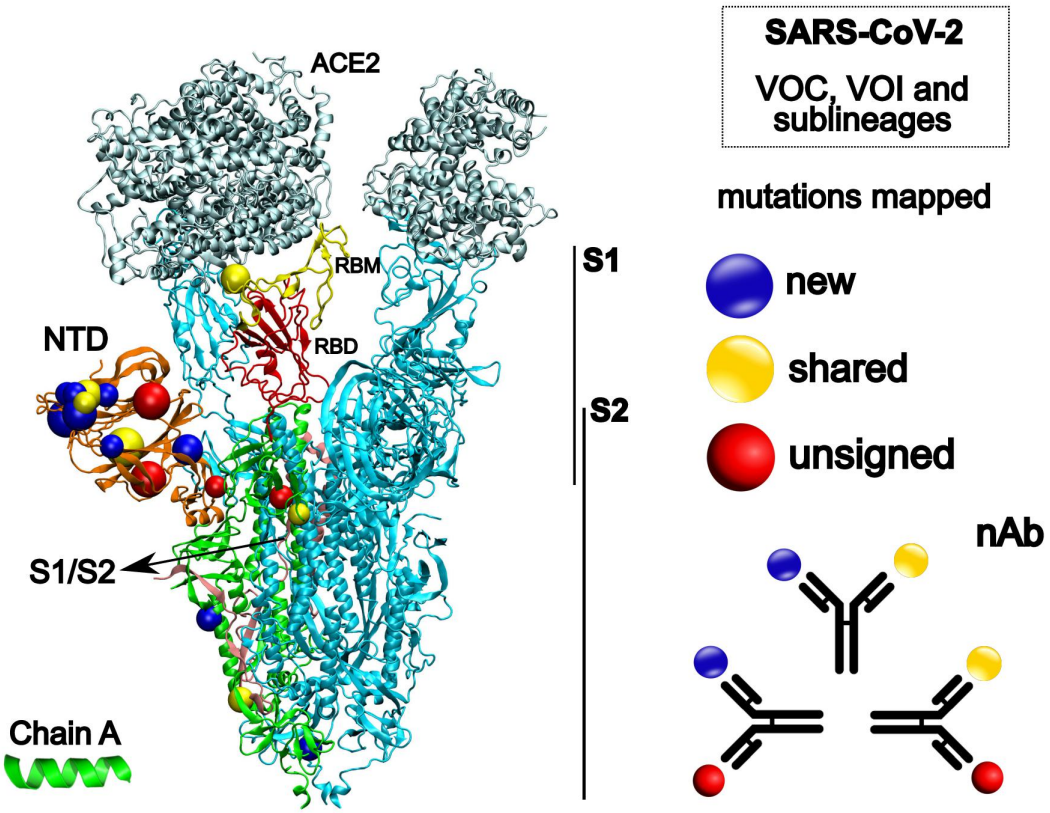

Figure 3. Structural representation of the SARS-CoV-2 Spike glycoprotein (PDB 7A98). On the left side, different colors represent ACE2, angiotensin-converting enzyme 2 (silver); NTD, N-terminal domain (orange); RBM, receptor binding motif (yellow); RBD, receptor binding domain (red) and S1/S2, cleavage site to S2 (pink). The structures in blue represent the chain B and C, respectively. Coloured spheres highlight the mutations mapped in the study. On the right side, our hypotheses about a linear perspective into the prospective vaccine efficacy against different SARS-CoV-2 strains. nAb, neutralizing antibody. Spike protein image was created with the Visual Molecular Dynamics (VMD) v.1.9.3 (Humphrey, Dalke, and Schulten1996).

This study demonstrates the influence of positive directional evolution on SARS-CoV-2 circulating in South America and in those countries most severely affected by the COVID-19 pandemic. Our methodology allowed for the identification of recombination breakpoints and distinct transmission subclusters. We were able to

indirectly infer transmission of a viral epidemiological chain and the generation of new variants. We also further identified and classified several convergently emerged shared mutations in different SARS-CoV-2 VOC and VOI. Lastly, we hypothesize that the co-circulation of SARS-CoV-2 variants and their possible sub-lineages takes place within a very close evolutionary environment, which can be translated to a setting of strong convergent evolution, where the viral effective population size have acquired identical site-specific mutations. Our results can help to anticipate a linear perspective with regards to future vaccine efficacy pandemic.

Data Availability

Some data on which this paper is based are too large to be retained or publicly archived with available resources. Smaller files which comprise information concerning recombination, convergent evolution, phylogenetic-based clustering analysis (ML trees, transmission clusters/subclusters), as well as the filtered and aligned sequences datasets used to map the shared and unsigned mutations in the SARS-CoV-2 VOC and VOI, are publicly available at https://github.com/rosadanilo/SARS-CoV-2_DEPS.

Acknowledgments

We gratefully acknowledge the authors and both the originating and submitting laboratories for the sequence data in GISAID EpiCoV and GenBank on which this work is based. The authors also thank the Rede Corona-Ômica/MCTI/FINEP, the National Laboratory for Scientific Computing (LNCC/MCTI, Brazil) for providing HPC resources of the Santos Dumont supercomputer (ID #45691), and Prof. Luiz Mário Ramos Janini for fruitful discussion.

Funding

This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, grants 2019/01255-9 and 2021/03684-4 (Young Investigator Program) (RD-C), and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, grant 405691/2018-1 (C.T.B). DRN is recipient of an institutional scholarship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, grant 88887.5062234/2020-00.

Conflict of interest: None declared.

References

Alteri C., Cento V., Piralla A. et al. (2021). Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nature Communication*, 12:434.

Bernal J. L., Andrews N., Gower C. et al. (2021). Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *The New England Journal of Medicine*, 385:585-594.

Boni M. F., Lemey P., Jiang X. et al. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology*, 5:1408–17.

Burki T. (2020). COVID-19 in Latin America. *Lancet Infectious Diseases*, 20:547–8.

Candido D. S., Claro I. M., de Jesus J. G. et al. (2020). Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*, 369:1255-60.

Centers for Disease Control and Prevention (CDC) (2021a). SARS-CoV-2 Variant Classifications and Definitions. Available at https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html.

Centers for Disease Control and Prevention (CDC) (2021b). Delta Variant: What We Know About the Science. Available at https://www.cdc.gov/coronavirus/2019-ncov/variants/delta-variant.html.

Corman V. M., Muth D., Niemeyer D. et al. (2018). Hosts and Sources of Endemic Human Coronaviruses. *Advances in Virus Research,* 100:163-88.

da Costa V. G., Moreli M. L., Saivish M. V. (2020). The emergence of SARS, MERS and novel SARS-2 coronaviruses in the 21st century. *Archives of Virology,* 165:1517-26.

Dearlove B., Lewitus E., Bai H. et al. (2020). A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proceedings of the National Academy of Sciences of the United States of America*, 117:23652–62.

Faria N. R., Mellan T. A., Whittaker C. et al. (2021). Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372:815-21.

Forni D., Cagliani R., Arrigoni F. et al. (2021). Adaptation of the endemic coronaviruses HCoV-OC43 and HCoV-229E to the human host. *Virus Evolution*, 7:veab061.

Greaney A. J., Loes A. N., Crawford K. H. D. et al. (2021). Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host & Microbe,* 29:463-76.

Guan Y., Zheng B. J., He Y. Q. et al. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*, 302:276–8.

Haddad D., John S. E., Mohammad A. et al. (2021). SARS-CoV-2: Possible recombination and emergence of potentially more virulent strains. *PLoS One*, 16:e0251368.

Harvey W. T., Carabelli A. M., Jackson B. et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19:409–24.

Hodcroft E. B., Zuber M., Nadeau S. et al. (2021). Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature*, 595:707–12.

Humphrey W., Dalke A., and Schulten K. (1996). VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33-38.

Huson D. H., Bryant D. (2005). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23:254–67.

Kirola L. (2021). Genetic emergence of B.1.617.2 in COVID-19. *New Microbes and New Infections,* 43:100929.

Lai M. M. C. (1992). Genetic Recombination in RNA Viruses. *Current Topics in Microbiology and Immunology,* 176:21-32.

Lamarca, A. P., Almeira, L. G. P., Junior, R. S. F. et al. (2021). Genomic surveillance tracks the first communitary outbreak of Delta (B.1.617.2) variant in Brazil. Available at https://virological.org/t/genomic-surveillance-tracks-the-first-communitary-outbreak-of-delta-b-1-617-2-variant-in-brazil/733.

Lemey P., Salemi M., Vandamme A-M. (2009). The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. *Cambridge University Press*.

Li X., Giorgi E. E., Marichannegowda M. H. et al. (2020). Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Science Advances,* 6:eabb9153.

Li X-N., Huang Y., Wang W. et al. (2021). Effectiveness of inactivated SARS-CoV-2

vaccines against the Delta variant infection in Guangzhou: a test-negative case-control real-world study. *Emerging Microbes & Infection*, 10:1751-1759.

Martin D. P., Weaver S., Tegally H. et al. (2021). The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv*, doi: 10.1101/2021.02.23.21252268.

Matsuda T., Suzuki H., Ogata N. (2020). Phylogenetic analyses of the severe acute respiratory syndrome coronavirus 2 reflected the several routes of introduction to Taiwan, the United States, and Japan. Available at https://arxiv.org/abs/2002.08802.

McCallum M., Bassi J., Marco A. D. et al. (2021a). SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. *Science*, 373:648–54.

McCallum M., Walls A. C., Sprouse K. R. et al. (2021b). Molecular basis of immune evasion by the delta and kappa SARS-CoV-2 variants. *bioRxiv*, doi: 10.1101/2021.08.11.455956.

MacLean O. A., Lytras S., Weaver S. et al. (2021). Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biology*, 19:e3001115.

Munnink B. B. O., Worp N., Nieuwenhuijse D. F. et al. (2021). The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nature Medicine*, 27:1518–24.

Muth D., Corman V. M., Roth H. et al. (2018). Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Scientific Reports*, 8:15177.

Okonechnikov K., Golosova O., Fursov M. et al. (2012). Unipro UGENE: a unified

bioinformatics toolkit. *Bioinformatics*, 28:1166–7.

PAHO (2021). Epidemiological Update: Increase of the Delta variant and its potential impact in the Region of the Americas - 28 September 2021. Available at https://www.paho.org/en/documents/epidemiological-update-increase-delta-variant-and-its-potential-impact-region-americas-8.

Paiva M. H. S., Guedes D. R. D., Docena C. et al. (2020). Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. *Viruses,* 12:1414.

Peacock T. P., Penrice-Randal R., Hiscox J. A. et al. (2021). SARS-CoV-2 one year on: evidence for ongoing viral adaptation. *Journal of General Virology*, 102:001584.

Pegu A., O'Connell S. E., Schmidt S. D. et al. (2021). Durability of mRNA-1273 vaccine-induced antibodies against SARS-CoV-2 variants. *Science,* 373:1372-77.

Pereira F. (2020) 'Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. *Infection Genetics and Evolution*, 85:104525.

Pereira F. (2021). SARS-CoV-2 variants lacking a functional ORF8 may reduce accuracy of serological testing. *Journal of Immunological Methods*, 488:112906.

Plante J. A., Liu Y., Liu J. et al. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, 592:116–121.

Pollett S., Conte M. A., Sanborn M. et al. (2021). A comparative recombination analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic. *Scientific Reports,*11:17365.

Pond S. L. K., Poon A. F. Y., Brown A. J. L. et al. (2008). A maximum likelihood method for detecting directional evolution in protein sequences and its

application to influenza A virus. *Molecular Biology and Evolution*, 25:1809–24.

Pond S. L. K., Poon A. F. Y., Velazquez R. et al. (2020). HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution*, 37:295-99.

Pond S. L. K., Posada D., Gravenor M. B. et al. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics*, 22:3096–8.

Price M. N., Dehal P. S., Arkin A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, 5:e9490.

Prosperi M. C. F., Ciccozzi M., Fanti I. et al. (2011). A novel methodology for large-scale phylogeny partition. *Nature Communication*, 2:321.

Resende P. C., Gräf T., Paixão A. C. D. et al. (2021). 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. *Viruses*, 13:724.

Ritchie H., Mathieu E., Rodés-Guirao L. et al. (2021). Coronavirus Pandemic (COVID-19). *Our World in Data.* Available at https://ourworldindata.org/coronavirus.

Sabino E. C., Buss L. F., Carvalho M. P. S. et al. (2021). Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *Lancet*, 397:452–5.

Singh D., Yi S. V. (2021). On the origin and evolution of SARS-CoV-2. *Experimental & Molecular Medicine*, 53:537–47.

Stefanelli P., Faggioni G., Lo Presti A. et al. (2020). Whole genome and phylogenetic analysis of two SARS-CoV-2 strains isolated in Italy in January and February 2020: additional clues on multiple introductions and further circulation in Europe. *Eurosurveillance*, 25:2000305.

Su S., Wong G., Shi W. et al. (2016). Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends in Microbiology*, 24:490–502.

U.S. Food & Drug Administration (FDA) (2020). Moderna COVID-19 Vaccine. Available at https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/moderna-covid-19-vaccine.

Velazquez-Salinas L., Zarate S., Eberl S. et al. (2020). Positive Selection of ORF1ab, ORF3a, and ORF8 Genes Drives the Early Evolutionary Trends of SARS-CoV-2 During the 2020 COVID-19 Pandemic. *Frontiers in Microbiology*, 11:550674.

V'kovski P., Kratzel A., Steiner S. et al. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology,* 19:155-170.

Voloch C. M., da Silva F. R., de Almeida L. G. P. et al. (2021). Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *Journal of Virology*, 95:e00119-21.

Worobey M., Pekar J., Larsen B. B. et al. (2020). The emergence of SARS-CoV-2 in Europe and North America. *Science*, 370:564–70.

Yadav P. D., Potdar V. A., Choudhary M. L. et al. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian Journal of Medical Research,* 151:200-209.

Zaki A. M., van Boheemen S., Bestebroer T. M et al. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *The New England Journal of Medicine*, 367:1814–20.

Zhang L., Jackson C. B., Mou H. et al. (2020). SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nature Communication*, 11:6013.