SARS-CoV-2 Worldwide Replication Drives Rapid Rise and Selection of Mutations across the Viral Genome: A Time-Course Study Potential Challenge for Vaccines and Therapies

Stefanie Weber¹[§], Christina M. Ramirez²[§], Barbara Weiser³, Harold Burger³, and Walter Doerfler^{1,4*}

Institute for Clinical and Molecular Virology, Friedrich-Alexander University (FAU) Erlangen-Nürnberg, 91054 Erlangen, Germany¹, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095-1772, USA², Department of Medicine, University of California, Davis, Sacramento, CA 95817, USA³, and Institute of Genetics, University of Cologne, 50674 Cologne⁴, Germany

Running title: Mapping the Mutants of SARS-CoV-2 on the Viral Genome

[§] These authors have contributed equally to this analysis

*Address for correspondence Walter Doerfler Institute for Clinical and Molecular Virology Friedrich-Alexander University (FAU) Erlangen-Nürnberg Schlossgarten 4 D-91054 Erlangen, Germany Tel.: +49-171-205-1587 E-mail: walter.doerfler@t-online.de

Significance and New Aspects of Study – Synopsis

- We examine the time course of emerging mutations in the SARS-CoV-2 genome that have rapidly been selected in the world's populations through the beginning of 2021. A study of the prevalence of viral mutations in the GISAID database in ten different countries United Kingdom, South Africa, Brazil, US, India, Russia, France, Spain, Germany, and China revealed widespread mutations along the genome.
- We previously identified about 10 hotspot mutations in the SARS-CoV-2 genome that became prevalent in many of the countries studied¹. Since the beginning of February, many new mutations arose in the ten countries (and worldwide). The preponderance of variants and mutations correlated with the increased spread of Covid-19.
- There was a temporal progression from about 10 predominant mutants shared by several countries up to the end of May 2020, followed by a consistent and rapid increase in the number of new mutations between June and December along with the emergence of variants of concern, first reported in December 2020.
- We examine the relative frequencies of mutations, along with variants of interest, in 10 countries up until January 20, 2021. Investigations on the pathogenic properties of individual SARS-CoV-2 mutations will be urgently needed to understand the kaleidoscopic patterns of worldwide Covid-19 outbreaks and symptoms. Monitoring the frequency and speed of mutant selection have direct relevance to diagnostic testing, vaccines and therapeutics.
- As an explanation for efficient viral mutagenesis, we hypothesize that the viral spike protein as documented facilitates viral entry via the cell's ACE receptor². This in turn interacts with the APOBEC polypeptide, an m-RNA editing function. The actually observed frequent C to U (T) transitions and other base exchanges are thus effected. Hence, as one of the earliest steps upon viral entry, active mutagenesis commences, since SARS-CoV-2 exploits one of the cell's defenses against viral infections.

Abstract

Scientists and the public were alarmed at the first large viral variant of SARS-CoV2 reported in December 2020. We have followed the time course of emerging viral mutants and variants during the SARS-CoV-2 pandemic in ten countries on four continents. We examined complete SARS-CoV-2 nucleotide sequences in GISAID, (Global Initiative of Sharing All Influenza Data) with sampling dates extending until January 20, 2021. These sequences originated from ten different countries: United Kingdom, South Africa, Brazil, USA, India, Russia, France, Spain, Germany, and China. Among the novel mutations, some previously reported mutations waned and some of them increased in prevalence over time. VUI2012/01 (B.1.1.7) and 501Y.V2 (B.1.351), the so-called UK and South Africa variants, respectively, and two variants from Brazil, 484K.V2, now called P.1 and P.2, increased in prevalence. Despite lockdowns, worldwide active replication in

genetically and socio-economically diverse populations facilitated selection of new mutations. The data on mutant and variant SARS-CoV-2 strains provided here comprise a global resource for easy access to the myriad mutations and variants detected to date globally. Rapidly evolving new variant and mutant strains might give rise to escape variants, capable of limiting the efficacy of vaccines, therapies, and diagnostic tests.

Keywords: SARS-CoV-2 mutation; UK variant B.1.1.7, VOC202012/01, (501Y.V2); South Africa variant B.1.351 variant; Spike mutations, mutational hotspots, time course of mutant emergence; Brazil variant, P.1, P.2; APOBEC-mediated editing

Introduction

Between December 2019 and January 28, 2021, the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) pandemic has expanded world-wide to 219 countries and territories; about 101.9 million people have been infected, and about 2.2 million (2.16 %) have lost their lives according to Johns Hopkins³.

In our laboratory, we have set out to follow the rapid rise of new mutations in the SARS-CoV-2 genome as Covid-19 cases soared worldwide. We identified mutation hotspots in different populations. Initially, we analyzed SARS-CoV-2 sequences that had been deposited in data bases between January and May/June of 2020. At least 10 prevalent sites of sequence mutations were observed and up to 80% of nucleotides at the mutated site were changed¹ Several of these mutations led to non-synonymous amino acid changes in different open reading frames across the viral genome. These alterations in functional viral proteins were selected during active worldwide replication of SARS-CoV-2. We have now extended the time frame of mutant analyses to January 20, 2021, and found increased prevalence of mutations along the genome worldwide. We specifically examined mutations from the US, India, Brazil, Russia, the UK, France, Spain, Germany, South Africa and China that were deposited in the GISAID (Global Initiative of Sharing All Influenza Data) database⁴.

As of January 28, 2021, infection rates worldwide were extremely high, surpassing the levels seen at the peak in April 2020³. The uncontrolled spread may have led to a proliferation of mutants and variants, which we define as viruses with a specific set of mutations. The so-called UK variant, also known as B.1.1.7 or alternatively VOC202012/01, was first identified in England in September 2020, and reported on December 8th as a rapidly spreading variant of concern that had 23 total mutations, 14 non-synonymous including 8 mutations in Spike, 6 synonymous, 3 deletions (**Table 1**) ⁵. Some of the mutations involve the gene for the Spike protein, which mediates binding, fusion, and entry of the virus into the host cell. One of these deletions, H69/V70 del (Δ H69/ Δ V70), has been reported to emerge during convalescent plasma treatment^{6,7}. Another Spike mutation, N501Y, is of concern as it is suggested to interact with ACE2 and could reduce the effectiveness of neutralizing antibodies⁸. This variant has been

associated with higher transmissibility^{9,10} and at least one confirmed case of reinfection¹¹ leading to lockdowns and travel bans in efforts to contain its spread. On December 23, 2020, the time of the lockdown, the variant was already found in Australia, Denmark and Italy. As of January 29, 2021. this variant is now reported in 54 countries according to GISAID (https://www.gisaid.org/hcov19-variants).

On December 18, 2020¹², another variant of concern, unrelated to the UK variant but also having the N501Y mutation, was announced in South Africa, and was dubbed 501Y.V2 or B.1.351¹³. This variant is characterized by 8 mutations in Spike including K417N, E484K and N501Y^{13,14} (Table 1). As of January 29, 2021 this variant has been reported in 24 countries and 5 continents.

Also rising independently, are 2 Brazil variants that are now called P.1 and P.2. P.1 has 17 unique amino acid changes, 3 deletions, 4 synonymous mutations and one 4 nucleotide insertion¹⁵ (Table 1). P.1 shares the 501Y and a deletion in ORF1ab with both the UK and the South Africa Variant. It is interesting to note that the N501Y mutation was not widely spread in Brazil before this variant was described while the E484K is more prevalent, although Brazil is not sequencing large numbers of samples. The E484K and the N501Y mutations are of particular concern in that they have been suggested to reduce neutralization by antibodies and increase the affinity for ACE2. P.1 and B.1.351 share both mutations N501Y and E484K (Table 1). P.1 has been associated with a case of documented reinfection¹⁶ and one case has been reported in the United States. P.2, unrelated to P.1, is characterized by the E484K mutation and has been implicated in two cases of reinfection^{17,18}.

These variants have caused concerns regarding efficacy of the vaccines. Recently Wu et al. described the efficacy of mRNA-1273 vaccine against many spike mutations tested both separately and in combination¹⁹. They show that sera from both vaccinated non-human primates and vaccinated humans are effective against the UK variant and various other spike mutations. They also found neutralization, albeit at lower levels, against the full South Africa variant B.1.135. It has been shown that the Pfizer BNT162b2 vaccine is effective against the N501Y mutant alone²⁰ as well as the UK variant B.1.117²¹, There have also been preliminary data from two other vaccine manufacturers showing efficacy against the South African variant. To illustrate the rise of mutations and variants over time, we list the number of variants and mutations deposited in GISAID worldwide across time (Figure 1). Table 2 lists the number of variant sequences deposited in GISIAD by country.

The rapid appearance of the variants across the world illustrates the importance of sequencing viral pathogens and tracking mutations. There is emerging evidence that these variants may alter transmissibility and have the potential to reduce the efficacy of existing COVID-19 vaccines. Sequencing SARS-CoV-2 is both a scientific and clinical imperative²². Because nucleic acid sequencing of SARS-CoV-2 samples is not part of routine clinical practice at this time, it is

necessary to institute programs to monitor sequence variation as a matter of course in order to detect mutations in the viral genome.

A consequence of the lack of routine viral sequencing is that it may contribute to selection bias. Sequences deposited to GISAID may not be representative of viral prevalence as different countries contribute different numbers of sequences. It is also possible that selection bias may be inherent, as different countries deposit sequences at different rates. Further, it was found that the Spike Δ H69/ Δ V70 causes the so-called S-dropout, rendering the nucleic acid test (NAT) negative for Spike (S) and positive for nucleocapsid (N). As this is one of the mutations in B.1.1.7, it has been used as a screening tool for this variant²³. While useful for screening, this deletion might create selection bias because patients who were positive for SARS-CoV-2 with an S dropout may have their samples preferentially sequenced as the prevalence for the new variant is being assessed.

Rapid increases in the number and types of new SARS-CoV-2 mutations in the world population within a time span of weeks to months are a remarkable biologic event. The uncontrolled rapid replication of SARS-CoV-2 in an immunologically naïve world population during one year constituted a wake-up call of the need to sequence and track the evolution of novel pathogens as these mutations and variants have raised concerns regarding increased transmissibility, immune escape and the efficacy of vaccines and the validity of diagnostic tests.

Methods

We analyzed complete SARS-CoV-2 genome sequences with known dates of sampling that were downloaded from GISAID: (i) Only complete sequences were included. (ii) For a chosen time period, all complete sequences with a sampling date from each country were included. Sequences were binned according to sampling date. iii) Sequences by country were filtered by country using the GISAID interface²⁴. Nucleotide sequences from the UK, South Africa, Brazil, the US, India, Russia, France, Spain, Germany and China were compared to the reference genome of the SARS-CoV-2 isolate from Wuhan-Hu-1, NCBI Reference Sequence: NC_045512.2. The programs Vector NTI Advance[™] 11 (Invitrogen[™]), Tool Align X, or Snapgene (GSL Biotech), by using the algorithm MUSCLE (Multiple Sequence Comparison by Log-Expectation), for the alignment of sequences. Amino acid sequences were also analyzed with the program Snapgene. DNA sequence analyses of reverse-transcripts of an RNA genome will have to be considered with the possibility that errors may have been introduced at several steps. e.g., by preferred reading mistakes of the reverse transcriptase due to specific sequence or structural properties of SARS-CoV-2 RNA. We have tried to overcome this obvious complication by analyzing a large number of genomes. Percentages were calculated by dividing the number of sequences with the mutation that were sampled at that time and available in the database by the total number of complete sequences with a known sampling date. In addition to the determination of mutants for defined time spans in ten countries, the total number of individual mutations was also determined in all sequences deposited to GISAID up until January 20, 2021 by using GESS (Global Evaluation of

SARS-COV-2/hCOV-19 Sequences²¹ as well as CoV-Glue²⁵ and PANGOLIN (Phylogenetic Assignment of Named Global Outbreak LINeages) https://github.com/hCoV -2019/pangolin)²⁶.

In the present study, somewhat arbitrarily, we set a 2% mark of mutations at a given nucleotide in the viral sequence as the cutoff for hotspot status and mutations recording in Tables 3 to 12. The SARS-CoV-2 RNA sequences investigated for mutant status had been deposited at time intervals of 2020 as follows:

Brazil: 02/25 to 08/15/2020; China-I: 12/23/2019 to 03/18/2020; China-II: 03/20 to 07/22/2020; France: April to 09/12/2020; Germany-I: February to 03/23/2020; Germany-II: February to 06/17/2020; Germany-III: 06/24 to 08/28/2020; Germany-IV 09/10 to 10/13; India: 01/27 to 05/27/2020 and 06/03 to 07/04/2020; Russia: 03/24 to 06/07/2020; South Africa: 09/01 to 12/07/2020; Spain: 06/01 to 09/20/2020; UK: 01/29 - 12/04/2020; US-I: 02/29 to 04/26/2020; US-II: 06/12 to 07/07/2020; US-III: 07/09 to 07/22/2020; US-IV 08/01 to 12/01. Some of the data had been reported previously in Table 1 of Weber et al. 2020^1 , but were included here again for comparison. These data were designated with an asterisk.

Results

Time course of emerging mutations in ten different countries

We examined mutations in 383,570 complete sequences with known sampling dates in GISAID up until January 20, 2021. **Figure 1** shows the worldwide distribution of Spike mutations as well as other variants of interest over time from April 2020 to January 20, 2021. **Table 1** lists the signature mutations for the variants. **Table 2** shows the total number of each variant of interest (B.1.1.7 (the UK Variant), 501Y.V2 (the South African Variant) and 484K.V2 (B.1.1 lineage with S: E484K/D614G, V1176F N: A199S/R203K/G204R) deposited in GISAID by each country as of January 20, 2021.

Selection of novel mutations in humans was rapid and frequent in 2020. Among the novel mutations discovered in the current study, some were seen only in one country and others occurred in several different countries. We will present the identified mutations arising in the SARS-CoV-2 RNA country by country for the designated time periods (**Tables 3 to 12**). The data covering time course analyses of the appearance of mutations and their nature in most of the ten different countries are presented in Tables 3A to 12A, The corresponding B Tables summarize the total number of mutations in individual sequence position at a cut off of 2% preponderance for the time period 01/19/2020 to 01/20/2021, i.e. of the entire first Covid-19 year.

Mutation analyses in ten different countries

The following paragraphs document the mutational repertoire of SARS-CoV-2 in different regions of the world. The results are somewhat biased in that countries differed considerably in the number of sequences that had become available for inspection in the GISAID database

(www.gisaid.org)²⁴. We have emphasized the time course of appearance of novel mutations in SARS-CoV-2 isolates that had a history of vigorous replication in some of the most severely affected populations on the globe, such as UK, South Africa, Brazil, the US, India, Russia, France, Spain, Germany and China. The most recent update [January 30, 2021] of Covid-19 cases and fatalities in the ten countries, whose isolates were analyzed for mutations, are presented in Table 13.

(i) United Kingdom

For mutations arising in the UK, we have not followed the time course of emerging mutations during earlier periods of the pandemic. In a total of >71.000 viral isolates of SARS-CoV-2 genomes from around the world, that were deposited between 01/19/2020 and 01/20/2021, 4 of the prevalent mutations found worldwide, at positions 241, 3,037, 14,408 and 23,403, had reached almost 100% representation (**Table 3**). In a total of 70 sequence positions >2% deviations in comparison to the Wuhan reference were noted, > 50% were C to U (T) transitions (see also Tables 3 to 11B). Twelve novel mutations reached prevalence values between 15% and 49%, 7 of them around 49%. Several of these mutations were also found in other countries (Tables 4 to 12). High prevalence of new mutations correlated with active replication in countries of high Covid-19 incidence.

On December 8, 2020 Rambaut et al⁵ described a novel variant of SARS-CoV-2 that was circulating in England starting in October and increased in prevalence suggesting a possible increase in transmissibility^{9,10,22}. An analysis of its genome revealed 14 non-synonymous mutations and 3 deletions that comprised a few nucleotides. In the spike glycoprotein 6 of these mutations and 2 deletions were located, one of them N501Y due to an A23063T replacement. This particular variant is now considered a variant of concern VOC202012/01²². Current reports have described increased infectivity of this variant, whereas its pathogenicity is currently being assessed^{10,27}.

Recent reports suggest that Pfizer BNT162b2 vaccine is effective against the UK variant as well as the N501Y mutant alone^{20,21}. Wu et al show preliminary effectiveness for the Moderna vaccine (mRNA-1273)¹⁹ against the variant. Press reports from Novavax²⁸ are also suggestive of the effectiveness of NVX- CoV2373 against the UK variant. Table 2 lists mutations found in the GISAID database up until February 1, 2021 and also includes data on VOC202012/01 as well as B.1.351 and 484K.V2.

(ii) South Africa

We analyzed 95 SARS-CoV-2 sequences from viral isolates in South Africa that were deposited in the GISAID databank [**Table 4A**]; 28 mutations overall were found in those sequences. Four of the 7 prevalent mutations, known from isolates all over the world, had reached 100% representation in the SARS-CoV-2 sequences, except those at positions 1,059 (~10%), 25,563

(~10%), and 28,881 (~63%). There were 7 new mutations unique to the South African variant, four of which caused non-synonymous amino acid exchanges. Twelve of the novel mutations were shared with other countries, eight of these mutations led to amino acid exchanges, many of them to non-synonymous replacements. Twenty-five percent of the mutations affected the spike glycoprotein, a finding that should alert us to the capacity of the virus to respond to potential vaccines directed against the viral spikes. There was one each mutation that involved the viral endoRNAse and the RNA-dependent RNA polymerase.

For the entire year 2020 (January 19, 2020 to January 20, 2021), the four prevalent mutations at positions 241, 3,037, 14,408, and 23,403 were again [Table 4B] represented close to 100%, the mutation at 28,881/2/3 in the nucleocapsid phosphoprotein gene at about 70% [**Table 4B**]. There were 8 new mutations at >10% prevalence. In a total of 63 positions in the viral genome deviations from the Wuhan reference sequence were noted above the 2% cutoff.

Recently, the N501Y variant was detected in South Africa which also had two additional point mutations, K417 and E484K. Data about its possible increased infectivity and transmissibility were preliminary²⁹. Also in December 2020, another variant called 501Y.V2, B.1.351 also known South African variant is characterized by 8 lineage defining mutation with 3 in the receptor binding domains: K417N, E484K and N501Y. This variant also appeared to spread quickly in South Africa giving rise to travel bans from South Africa. It has been suggested that this variant is able to escape neutralization by donor plasma³⁰. Increased transmissibility has also been suggested²⁹. Furthermore, there is early evidence that the efficacy of multiple existing vaccines against the B.1.351 variant may be diminished^{19,28,31}. It will be important to continue to perform sequence analysis of viral strains and to correlate the evolution of mutants and variants with viral transmission and vaccine efficacy.

(iii) Brazil

In the nine SARS-CoV-2 mutations identified in a subset of about 100 published sequences available from Brazil in one time frame between 02/25 and 08/15 [**Table 5A**], five belonged to the worldwide prevalent hotspots at nucleotide numbers 241, 3,037, 14,408, 23,403, and 28,881. Two mutations at positions 12,053 and 25,088 were unique to the sequences from Brazil, and were noted in between 15.7 and 34.4 % of the analyzed sequences, respectively. Two of the novel shared mutations were also identified in sequences from France and Russia (27,299 and 29,148) at frequencies of about 40 %. The mutation at nucleotide position 28,881 was found in 71.6 % of the viral sequences studied. This mutation occurred in viral sequences from all countries investigated, except in those from China.

Of note, among the nine different new mutations observed in the SARS-CoV-2 isolates from Brazil, two were not observed in isolates from any of the eight other countries investigated. Possibly, they had recently emerged in the Brazilian population in which the virus had been replicating very actively, and the mutations had been selected under conditions of pandemic viral

abundance. The frequent $C \rightarrow T$ mutations amounted to 44.4% frequency in this selection. Note that the time analysis cut off occurred before the reported emergence of variant strains P.1 and P.2 were identified. We include the related 484K.V2 variant in Table 5B along with the number of individual mutations for all complete sequences with known sampling dates deposited to GISAID by January 20, 2021.

Impact on Coding Capacity: The two Brazil-unique mutations at positions 12,053 (viral replicase) and 25,088 (viral spike protein) led to leu to phe and val to phe synonymous replacements, respectively. The two novel shared mutations at positions 27,299 (ORF6 protein) and 29,148 (nucleocapsid phosphoprotein) both caused ile to thr replacements of a non-conservative nature.

Table 5B shows 27 individual mutations for the >1100 complete sequences with known sampling dates deposited to GISAID by January 20, 2021. The predominant mutations at positions 241, 3,037, 14,408, 23,403 showed frequencies at 99%. The mutation in the nucleocapsid phosphoprotein at position 28,881/2/3 presented with 93%, the highest frequency for this mutation among all 10 countries studied. As shown in Table 5A, in the time course study the nucleocapsid mutation reached a similarly high of 89%. C to U transitions in these samples reached only 29%. As of January 20, 5 cases of occurrence of the B.1.1.7 variant from the UK were reported.

(iv) USA

Table 6A lists mutations from a random subset of sequences selected in the US at 4 different time points. Some of the long-term prevalent mutations presented in the table under US-I and US-II were already included in a previous analysis as indicated by an asterisk¹. They were listed here again to facilitate comparisons to the wider spectrum of new mutations that arose in the US (US-III, US-IV) and in different countries in the course of a few weeks. In addition to the worldwide occurring prevalent mutations, at nucleotide (nt) numbers 241, 1,059, 3,037, 8,782, 14,408, 23,403, 25,563, 28,144, and 28,881, there were a total of 13 unique, i.e. not previously described mutations in our analyses of which nine were found exclusively in the US-III sample cohort at frequencies between 4 % and 29.3 % (Table 6A, unique). Except for three of these mutations, many attained their highest frequency of occurrence at the time point US-III. Two of the novel unique mutations in sequence positions 17,858 and 18,060 had disappeared in the US-III samples. Seventeen of the novel mutations were shared by other regions in the world, seven appeared in most or all ten countries investigated. We listed 13 mutations that had disappeared in the July samples of US-III, possibly they had proved not to be penetrating enough or were not sampled due to selection bias. As apparent in the table, five of the 15 new mutations among the US-II sequences deposited between June 12 and July 07 occurred at low frequencies (< 10%) exclusively in this collection of sequences, others, also at low frequencies, were also present in isolates from other countries as indicated. There were a number of novel shared mutations which were also represented in other countries- BR Brazil, CN China, FR France, DE Germany; IN India, RU Russia, ES Spain, ZA South Africa. The more recently selected SARS-CoV-2

mutations under US-III stemmed from the time period between July 09 and July 22, 2020. The comparison of June and July US-III sequences and their mutations to their counterparts from a month earlier (US-II) revealed the complex vitality of new mutants arising in a SARS-CoV-2 population that had been replicating during a most critical phase of the US pandemic during the summer of 2020. During the four months' period 08/01 to 12/01 (US-IV), another 117 SARS-CoV-2 sequences were added to Table 6A. Several of the predominant mutations reached 100% representation. Eight novel mutations, some unique, others shared, were listed at nucleotide positions 8,083, 10,139, 18,424, 21,304, 25,907, 28,472, 28,869, and 28,887; most of them reached >20% representation. At many nucleotide positions in the viral genome, the frequencies of the long-term predominant mutations increased over the entire time period between the last days of February to the end of July. This study has thus allowed us to witness the spread of mutations in the US population and at the same time the constant emergence of novel mutations and their increase in frequency with time.

Impact on Coding Capacities:

There is the idea that all mutations exist at a low level, but are detected when they are selected and proliferate. Of the 39 SARS-CoV-2 RNA sites mutated, 13 mutations, i.e. 42%, remained without effect on the encoded protein. In contrast, 18, i.e. 58%, exhibited changes in the genomes coding capacity [noted in red in Table 6A] which affected most of the virus-encoded proteins.. Most amino acid exchanges were non-synonymous and were likely responsible for functionally important alterations as judged from the type of amino acid replacements, e.g. pro to ser (nucleotide position 4,226) in nsp3; leu to phe (7,837), also in nsp3; tyr to cys (17,858) in the viral helicase; asp to gly (23,403) in the spike glycoprotein; arg-gly to lys-arg (28,881) in the nucleocapsid phosphoprotein and others. Among the additional eight mutations in the US-IV period, four led to non-synonymous amino acid exchanges in functionally important proteins as the 2'-O ribose-methyltransferase, the 5'-3' exonuclease, and the nucleocapsid phosphoprotein.

The asp to gly exchange due to the mutation in position 23,403 that affected the viral spike glycoprotein, was described earlier¹⁸. The mutant grows to higher titers in cell cultures, reaches higher viral loads in the upper respiratory tract but does not lead to increased disease severity¹⁸. The mutation has been reported to increase susceptibility to neutralization¹⁹. At this point, the functional consequences of most of the identified mutations for viral replication and/or pathogenicity need to be assessed. The SARS-Co-V-2 variant discovered in the UK in December 2020 will be discussed in part (*iii*) of the Conclusion section.

Analysis of mutation frequencies during short periods of time as compared to those observed over the entire year 2020: In addition, a total of 52,934 SARS-CoV-2 sequences from the US in GISAID was analyzed for the presence of mutations as compared to the original Wuhan sequence (**Table 6B**) over the entire year 2020. A total of 42 sequence positions showed >2% deviations from the reference sequence; 21 (50%) were C to U (T) transitions. Similarly high C to U preferences in sequence exchanges were observed in isolates from some of the other 9 countries that were analyzed. In the Conclusion section of this work, a presumptive editing function

(APOBEC) is discussed to account for the prevalence of C to U transitions in all these viral genomes. SARS-CoV-2 represents itself as a highly adaptable virus that optimally utilizes its and the host cell's capacities to generated mutations and has them efficiently selected under a wide range of conditions in human populations.

As of January 20, 2021, a total of 81 isolates of the UK variant B.1.1.7 was reported in the US which is probably a gross underestimate, by January 22 this variant had reached 12 US states. Worldwide, the occurrence of SARS-CoV-2 mutations and variants is changing daily as expected at the height of this pandemic.

(v) India

During the periods of sequence analyses between January 27, 2020 to May 27, 2020 (IN-I) and June 03, 2020 to July 04, 2020 (IN-II) the prevalent hotspot mutations at sequence positions 241, 3,037, 14,408, 23,403, and 25,563 had reached values of representation approaching 100 %, except at position 25,563 which amounted to 52% of sequences [Table 7A]. New mutations emerged during these time periods. A set of nine novel mutations, unique to the Indian population, were observed, i.e. 39.1% out of a total of 23 mutations in all sub-samples from India. These unique mutations were located in genome positions which were completely different from the newly arising SARS-CoV-2 mutations in the US or in any other population investigated in our study [Table 7A]. A total of seven of these novel mutations originated or increased in frequency in the late IN-II time period, whereas two of the mutations could no longer be detected during that same period. An additional nine newly arising mutations were shared with those in countries as indicated, some of which reached a frequency of up to 50%. Among all mutations from the Indian samples, $C \rightarrow T$ transitions held the majority of 15/23, i.e. 65.2%. We note that 18 out of 23 (78.3%) mutations in the SARS-CoV-2 isolates from our sub-samples from India were novel. About 7/9 of the India-unique mutations appeared *de novo* or increased in frequency within a time period of a few weeks of very active replication of the virus in the Indian population. New mutations are not only perpetually arising during the present stage of a nearly uncontrolled Covid-19 pandemic, but are also capable of becoming selected in Indian population. Table 7B lists 46 individual mutations for >3270 complete sequences with known sampling dates deposited to GISAID by January 20, 2021. The prevalent mutations at positions 241, 3,037, 14,408, and 23,403 (Tables 3 - 12) were represented at about 86%, at position 28,881 at 44%. In total 46 positions showed mutations at frequency levels >2%, ten of them >10%. The frequency of C to U transitions among all mutations in the samples from India was 50%.

Impact on Coding Capacity: The change in coding capacity of the long-term prevalent mutations in positions 241, 3,037, 14,408, and 23,403 was described for the US samples. Among the nine India-unique mutations, the following four led to functionally significant amino acid exchanges: Position 2,292 (nsp2) gln – pro; 18,568 (3'–5'-exonuclease) leu – phe; 19,154 (3'–5'-exonuclease) thr – ile; and 28,311 (nucleocapsid phosphoprotein) ser – leu. Among the nine additional mutations, which were shared by one or several countries, only the following four led to amino acid exchanges: 6,312 (nsp3) thr – lys; 11,083 (nsp6) leu/tyr – phe; 21,724 (spike

protein) leu-phe – phe-phe; 28,854 (nucleocapsid phosphoprotein) ser – leu (Table 7A). Again, many of the new SARS-CoV-2 mutations were responsible for functionally important non-synonymous amino acid exchanges in the corresponding protein.

(vi) **Russia**

Among the RU-I subsample of 226 SARS-CoV-2 RNA sequences analyzed between 03/24 and 06/07/2020 in the isolates from Russia, there were ten mutations of which six belonged to the previously described long-term prevalent mutations at positions 241, 3,037, 14,408, 23,403, 25,563, and 28,881 [**Table 8A**]. The latter mutation in position 28,881 at a frequency of representation of 76.1% stood out in that it was not a point mutation but involved a three nucleotide exchange creating a highly basic domain in the 3' terminal region of the SARS-CoV-2 nucleocapsid phosphoprotein as reported earlier¹. The 28,881 mutation in the Russian sequences had reached one of the highest frequency at 88%. The four new mutations were located at sequence positions 3,140 (CC \rightarrow TC, with a pro to asn-leu exchange in the amino acid sequence of nsp3), 20,268 (AG \rightarrow GG, without change in amino acid composition in the endo RNase), 26,750 (CA \rightarrow TA, without effect on the membrane glycoprotein), and at 27,415 (GC \rightarrow TC, and an ala to ser change in the ORF6 protein).

Table 8B presents similar results of analyses on about 1,330 sequences collected during one year between 01/19/2020 - 01/20/2021. Again the prevalent mutations had reached close to 100% frequency, the nucleocapsid phosphoprotein about 90%. New mutations were not apparent. C to U transitions stood at 38%.

As of January 20, the detection of any of the new SARS-CoV-2 variants was not reported in Russia.

(vii) France

Mutation frequencies were determined between 04 and 09/12, 2020 (116 SARS-CoV-2 sequences). In the sequences a total of 27 mutations were documented. Among them, seven of the previously described long-term prevalent mutations were identified at frequencies as follows: Nucleotide position 241 (100%), 1,059 (13.8%), 3,037 (99.1%), 14,408 (98.3%), 23,403 (100%), 25,563 (49.1%), 28,881 (14.7%). There were 20 new mutations at frequencies between 10 and 20% that were not described previously¹. C - U transitions reached 37%. [**Table 9A**]. Of interest, none of the new mutations was unique to France in the 116 sequences displayed in Table 9A. Instead, a large percentage of the mutations were shared with Germany and Spain, both neighboring countries. Most novel mutations occurred between 10% and 20% [Table 8A] at frequencies between 10 and 20%. C to U transitions were found in 46% of sequences. Among the novel mutations, 20 occurred at >10, many of them >20% frequencies.

Table 9B lists mutational frequency in sequences deposited up until January 20, 2021, including data on variants of interest and of concern. There are scant data on the occurrence of variants [Table 2], the UK variant B.1.1.7 was counted 20 times, the South African one 5 times. As

complete sequence analyses on Covid-19 isolates are progressing rapidly, new data on the emergence of new variants can be expected.

Impact on Coding Capacity: Among these 20 not-previously described novel mutations, eight did not affect the coding capacity of the relevant viral proteins. Most of the 12 coding-relevant mutations led to amino acid exchanges that were non-synonymous: nsp2, 3, 4, RNA-dependent RNA polymerase, the helicase, the endoRNAse, the spike glycoprotein, and the nucleocapsid phosphoprotein [Table 9A, B].

(viii) Spain

In the Spanish isolates from the period between 06/01 and 09/20/2020, we analyzed 135 sequences and observed 20 mutations [**Table 10A**]. Of these, four, the long-term prevalent ones, had been described earlier in positions 241, 3,037, 14,408, and 28,881. Except for the latter one at 10.4% frequency, the three former came close to 100% occurrence. Of the 16 new mutations, six occurred in Spanish isolates exclusively (termed unique), namely in positions 5,572 (GT \rightarrow TT, frequency 8.1%, changing the amino acid sequence met to ile in nsp3), 5,784 (CT \rightarrow TT, frequency 9.6%, thr to ile in nsp3), 25,062 (GT \rightarrow TT, frequency 13.3%, amino acid change gly to val in the spike glycoprotein), 27,982 (CA \rightarrow TA, frequency 9.6%, changing the sequence from pro to leu in the ORF8 protein), 28,657 (CG \rightarrow TG, at frequency of 14.1%, without affecting the amino acid composition in this position in the nucleocapsid phosphoprotein from ala to val). The remaining 10 novel shared mutants were also found in isolates from other countries and were located in positions as shown in previous tables. With the exception of a point mutation at position 25,049 in the spike glycoprotein and an ensuing amino acid exchange from asp to tyr, none of the other nine mutations in the shared category led to an amino acid exchange.

We also note that in the Spanish collection of SARS-CoV-2 mutations, there were four in the spike glycoprotein, all different from the well-known position 23,403. Two of these new spike mutations led to non-synonymous amino acid exchanges in the spike glycoprotein: In position 25,049 asp to tyr, and in 25,062 gly to val [Table 10A]. Such mutations might become relevant when evaluating the efficacy of a solely spike-directed SARS-CoV-2 vaccine. As a note of caution, one should not rule out functional consequences of nominally silent mutations for SARS-CoV-2 competence, since they might affect the secondary structure of the viral RNA with sequelae in replication and relevant interactions of the viral genome with viral and/or cellular proteins.

It is interesting to note that although the latest Spanish collection of SARS-CoV-2 mutations contains four mutations in the spike glycoprotein, in earlier time points, the D614G mutation at position 23,403, the site of a prevalent mutation^{1,32} was not present [Table 10A]. In **Table 10B**, describing mutant frequencies between 01/19/2020 and 01/20/2021, the 23,403 mutant was present at about 80%, whereas in France and England prevalence was >96%. Moreover, for the 01/2020 to 01/2021 period, mutations in 38 sequences lay above the 2% cut off. The predominant

mutations reached values around 80% representation. C to T transitions were at 42%. Among the novel mutations 17 showed prevalence of >10%, 8 of them of >20%.

(ix) Germany

During the course of the pandemic, we tabulated the occurrence of SARS-CoV-2 mutants which arose between February to 03/23 (DE-I)¹, February to 06/17 (DE-II), 06/24 to 08/28 (DE-III), the latter isolates with only 17 sequences available for analyses, and 09/10 to 10/13 (DE-IV) with 70 sequences. Apart from the prevalent mutations, there were relatively few mutations exceeding 10% representation in the time frame of DE-II. Among the total of 33 mutations in the SARS-CoV-2 RNA sequence [Table 11A], seven belonged to the previously described collection of long-term prevalent sequences - at positions 241, 1,059, 3,037, 14,408, 23,403, 25,563, 28,881 with coding frame alterations as outlined in previous Tables. In the DE-III sample, four of these long-term prevalent mutations had reached 100% representation, two had disappeared, and the mutation at 28,881 had remained at about 53%. Six mutations could be detected exclusively in the DE-III samples from Germany, in positions 3,602 (CA \rightarrow TA), 6,941 (CT \rightarrow TT), 21,855 $(CT \rightarrow TT)$, 25,505 (AA \rightarrow GA), 25,906 (GG \rightarrow CG), 28,869 (CA \rightarrow TA), all of them at 29% of representation. There were mutations in six positions which had been observed also in isolates from other countries, as indicated, and all of them showed modest frequencies. It is interesting to note that 52% of the mutations detected in sequences from France were shared with Germany, but only 16% of the mutations identified from Germany were shared with those from France [Table 9A]. During the time interval of about a month, 09/10 to 10/13 (DE-IV), that immediately preceded a marked rise in Covid-19 cases in Germany, 23 new mutations were identified 6 of which reached a prevalence of >20% and 7 of >10% in the SARS-CoV-2 sequences studied. During the same period, 4 of the prevalent mutations were represented in 100% of sequences, one, at 28,881 of 54%.

Table 11B lists the total number of mutations and variants up until January 20, 2021 from GISAID complete sequences with 52 entries at >2% incidence. The prevalent mutations reach about 86% occurrence. Only at three sites, mutations were found at >10%. C to U transitions were recorded in 46% of the studied sites.

Impact on Coding Capacity: With the exception of the point mutation at 6,941 which was synonymous, the five other mutations were non-synonymous: 3,602 his to tyr (nsp3); 21,855 ser to phe (nsp3); 25,505 glu to arg (ORF3a protein); 25,906 gly to arg (ORF3a protein); 28,869 pro to leu (nucleocapsid phosphoprotein).

(x) China

In late December of 2020, the first cases of Covid-19 emerged in Wuhan, Hubei Province in China, reportedly among workers and customers of the Huanan Seafood Market. The Chinese authorities eventually reacted with a very strict shutdown in Hubei Province, the epicenter of Covid-19, to limit the spread of the new disease. At present, most new cases of Covid-19 are

reportedly being registered in Shanghai and a few additional places. The analyses of SARS-CoV-2 mutants up to March 18, 2020 (CN-I) revealed point mutations in only two genome positions, 8,782 (CC \rightarrow TC, without amino acid exchanges) and 28,144 (TA \rightarrow CA causing a leu to ser exchange in ORF8 protein), both at frequencies of 29.3% [Table 12A]. An extension of our mutant research among a relatively limited number of published sequences to the period from 03/20 to 06/22, 2020 (CN-II) revealed mutations in five of the long-term prevalently affected sequence positions: 241 (CG \rightarrow TG at a frequency of 69.7% without coding changes), 3,037 (CT \rightarrow TT, at a frequency of 69.7%, without coding changes), 14,408 (CT \rightarrow TT at a frequency of 57.6% and a codon change pro to leu in the gene for the RNA-dependent RNA polymerase), 23.403 (AT \rightarrow GT at a frequency of 66.7% and an asp to gly exchange in the spike glycoprotein), and at 28,881 (GGG \rightarrow AAC at a frequency of 33.3% and the codon exchange arg-gly to lys-arg, reported previously). Remarkably, the novel shared point mutations in positions 8,782 and 28,144 had disappeared at the later time point [Table 21]. These latter mutations may have been introduced to China by visitors or business travelers, and then died out because they did not confer a strong evolutionary advantage or due to not enough sequencing. The total counts of mutations up until January 20th are presented in Table 12B.

Conclusions and Problems

(i) SARS-CoV-2 genetics will require in-depth analyses

It has been the intent of this project to follow the genetic evolution of SARS-CoV-2 after the virus transgressed a host barrier and during the ensuing major pandemic in the human population. The virus has shown great replicative and mutagenic potential and appeared in the large human population of 7.8 billion that lacked previous encounters with SARS-CoV-2. In this context, the primary question was not to understand viral mutagenesis in general in its biochemical or genetic details, but to identify mutants that have potential to become prevalent with possible fitness advantages. Which mutants and variants would have the capability to persist and multiply in the course of rapid spread of SARS-CoV-2 within the human population? It will be a continuing long-term challenge to pursue the outcome and time course of a competition in that 29,903 nucleotides in the viral genome were pitted against about 3 billion in the human genome. The SARS-CoV-2 has a repertoire of mutable sites in a stretch of 29,903 nucleotides that cannot only be varied by introducing point mutations but be extended by an almost inexhaustible combination of multiple mutations in the same genome, by deletions and insertions. Before the viral dominance in the human population began, SARS-CoV-2 had already made a major leap, its transition from an animal to the novel human host, an undocumented step in its own right in which mutagenesis and selection must have played a major role. Thus, the impact of ethnic and socio-economic differences in the human population will have to be considered as important factors. In a summary of all mutation analyses we have compared the number and types of mutations to the extent of the Covid-19 pandemic in ten different countries that currently report high numbers of cases and fatalities [Table 13].

Of course, this summary offers only a broad temporal correlation of mutant data and extent of the pandemic in individual countries. High current incidence of Covid-19 is paralleled by high numbers of new mutations and variants, although this relationship was not observed in Brazil or Russia. In anticipation, it will be a further challenge to evaluate the real-world success of the numerous Covid-19 vaccination programs.

(ii) Replication and selection

Rapid worldwide replication of SARS-CoV-2 in heterogeneous populations has been paralleled by the rise of novel mutations. In this report, we have studied mutations in SARS-CoV-2 RNA sequences isolated in the UK, South Africa, Brazil, the US, India, Russia, France, Spain, Germany and China that have become available in the GISAID database during a one-year period between January 19, 2020 and January 20, 2021. We have examined the rise of novel mutations both using sequence subsets segregated by date and also overall in a large cross-section. It seems that towards the end of the year, more mutations in combination were found and propagated rapidly despite lockdowns and other efforts to contain the spread, perhaps owing to potential increased transmissibility. The current data are compatible with the interpretation that rapid regional expansion and efficient viral replication in human populations of very different genetic and socio-economic backgrounds further the selection of new mutations in the viral RNA genome. Differences in defense mechanisms operative in various populations infected by SARS-CoV-2 and/or the various therapeutic measures employed in fighting the infection might also have influenced the selection of new mutants. It is uncertain whether there was regions-specific selection of specific mutations or whether other factors might have furthered differences in unique versus shared novel mutations.

Figure 1 and Table 1 show the number of novel variants in each country as of January 20, 2021. The speed by which the virus traveled even during lockdowns emphasizes the difficulty in suppressing transmission of highly contagious respiratory viruses. The new variants have not been associated with increased pathogenesis although more research needs to be done. The preliminary finding of increased transmissibility of the B.1.1.7 and B.135 variant hinder efforts to contain the virus^{5,9,10,13,22,27,29,30}. The vaccines are expected to work against the novel variants, although with some at reduced efficacy^{19,20,28,31}, but caution is urged to watch viral evolution.

(iii) Rise of novel mutations and variants with new properties – A hypothesis

After initially demonstrating the prevalence of about 10 mutants in at least 10 different countries, SARS-COV-2 evolved to display new point mutations worldwide that were selected among affected populations in a time period of weeks [Tables 3 A, B to 12 A, B]. In **Table 13**, column 5, the number of new point mutations in some of the countries analyzed ranged between 16 and 38. As a consequence of highly efficient sequencing programs in the UK [UK Consort], previously not recognized variants started to appear in in late 2020 and are currently spreading worldwide

(Figure 1,Table 2). The impact of these variants on potential increases of the already existing pathogenicity cannot be UK, South Africa, Brazil, the US, India, Russia, France, Spain, Germany and China assessed at present.

We posit the following hypothesis: SARS-CoV-2 uses the ACE2 (angiotensin-converting enzyme 2) receptor for its entry into human cells². It remains to be determined how the interaction of ACE2 receptor with the spike protein of SARS-CoV-2 affects the location and activity of APOBEC (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide). This class of m-RNA editing functions causes deamination of cytosine to $uracil^{33}$. The high frequency of C to T (U) transitions among SARS-CoV-2 mutants-- among them, 40.7% (France), 59% (US) to 69.9% (India) – as examples - were C to U (T) transitions (see also Tables 4B to 12B) [see Table 13] – up to >88% in the UK samples - might be linked to an m-RNA-editing mechanism^{1,34,35}. Moreover, the high incidence of C to T (U) transitions renders research on the occurrence of methyl-cytosine bases in SARS-CoV-2 RNA a project of considerable importance. In this context, the introduction of 14 point mutations and 3 small deletions in the genome of the B.1.1.7 suggests m-RNA editing as a plausible model. The rapid generation of SARS-CoV-2 point mutations and the sudden rise of ubiquitous and efficiently selected SARS-CoV-2 variants also supports the m-RNA editing mechanism as an attractive hypothesis. The editing function has been interpreted as a cellular defense against intruding viral genomes, and SARS-CoV-2 exploits exactly this mechanism to further its mutagenic potential.

(iv) Will the constant selection of new mutants impinge upon the success of therapeutic or vaccination strategies?

There are multiple sources of vaccines against Covid-19 available now or at various stages of including those from Pfizer/BioNTec; AstraZeneca/Oxford University; development, Moderna/US National Institutes of Health; Johnson and Johnson Novavax; Curevac/Bayer and firms in Russia, China, India, and many more. It is impossible to assess the vaccines' overall long-term efficacy against SARS-CoV-2 infections at this time. Vaccines available now have demonstrated a high level of clinical efficacy. There are also, however, preliminary data suggesting that evolution of viral variants may have diminished efficacy of several vaccines against one of the new SARS-CoV-2 variants^{19,28,31,36}. The emergence of novel mutants of SARS-CoV-2 in short temporal succession and their difficult-to-assess impact on pathogenicity in vivo further complicate predictions about future vaccine efficacy at this time. For example, some of the early laboratory assessments of efficacy versus the new variants have focused on neutralization by sera from immunized individuals; the clinical efficacy of vaccines, however, is likely to benefit from cell-mediated immunity as well. Furthermore, sophisticated and specific plans are already in place to alter the Covid-19 vaccines to compensate for possible escape mutants. SARS-CoV-2 is a new and evolving pathogen. Effective vaccines have been developed within one year of the identification of the pathogen, a remarkably short time. Ingenuity and basic research are likely to offer solutions to help control the spread of SARS-CoV-2 and future emerging viruses.

Note added during submission

Domingo and Perales [J. Virol. 2021, doi: 10.1128/JVI.02437-20.] point out that the time for Covid-19 vaccination should be at low incidence of disease, in order to avoid selection of SARS-CoV-2 escape mutants.

(v) "Conceptual bridge between natural language and viral evolution"

With this conceptual bridge, Hie et al. studied viral escape mechanisms to circumvent cellular defenses by adapting machine learning algorithms that had been developed to analyze human languages³⁷. The viral mutation seeks to escape by looking different to the immune system akin to word changes keeping the grammar of a sentence while altering its meaning. With that seemingly remote approach, the authors hope to predict viral structures that will be able to escape immunological defenses. This idea has been applied to influenza hemagglutinin, HIV-1 envelope glycoprotein and SARS-CoV-2 spike glycoprotein. Will an intellectual *spiel* from linguistics can actually help solve a complex biological problem remains to be seen and a subject for future work.

Acknowledgments

We acknowledge and are very grateful to the GISAID Initiative and for the hard work and openscience of the individual research labs and public health agencies that have made their genome data accessible on GISAID, on which this research is based.

This research was supported by the Dr. Robert Pfleger Stiftung in Bamberg, Germany [5.12.2018]. W.D. is indebted to the Institute for Clinical and Molecular Virology of FAU in Erlangen, Germany for their continued hospitality extended to the Epigenetics Group.

Competing interests

The authors declare no competing interests.

Author contributions

S.W. carried out all work involving sequence selection and formal analyses, was involved in the conceptualization of the project and in the analysis and interpretation of data. C.R. performed the analysis on the large sequence database and variants of interest/concern using GISAIS, GESS, CoV-Glue and other computational tools, statistical analyses, interpretation of the data, and writing of the manuscript. B.W. and H.B. contributed to the analysis and interpretation of the data. B.W. contributed to writing the manuscript. W.D. initiated the project, was involved in the conceptualization of the project, in the analysis and interpretation of data and wrote the manuscript with C.R.'s and B.W.'s contributions.

References:

1. Weber S, Ramirez C, Doerfler W. Signal hotspot mutations in SARS-CoV-2 genomes evolve as the virus spreads and actively replicates in different parts of the world. Virus Research 2020;289.

2. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;182:812-27.

3. Dong E, Du H, Garner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis 2020;20:533-4.

4. Elbe S, Buckland-Merritt G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges 2017;1:33-46.

5. Rambaut A, Loman N, Pybus O. Preliminary genomic characterisation of an emergentSars-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virologicalorg 2020.

6. Kemp S, Collier D, Datir R, al. e. Neutralizing antibodies in Spike mediated SARS-CoV-2 adaption. MedRxiv 2020.

7. Kemp S, Datir R, Collier D. Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion Δ H69/ Δ V70 MedRxiv 2020.

8. Yi C, Sun X, Ye J, al e. Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. Cell Mol Immunol 2020;17:621-30.

9. NERVTAG. New and Emerging Respiratory Virus Threats Advisory Group. NERVTAG meeting on SARS-CoV-2 variant under investigation VUI-202012/01. 2020.

10. Volz E, Mishra S, Chand M, et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. MedRxiv 2021;2020.12.30.20249034.

11. Harrington D, Kele B, Pereira S, Coutyo-Parada X, Riddell A, al e. Confirmed Reinfection with SARS-CoV-2 Variant VOC-202012/01. Clinical Infectious Diseases 2021;ciab014.

12. WHO. SARS CoV-2 Variants. In: News DO, ed.2020.

13. Tegally H, Wilkinson E, Giovanetti M, et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. MedRxiv 2020.

14. Pond S, Wilkison E, Weaver S, James S, H T, al e. A preliminary selection analysis of the South African V501.V2 SAR-CoV-2 clade. virologicalorg 2020.

15. Faria N, Claro I, Candido D, Franco L, al e. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. virologicalorg 2021.

16. Naveca F, da Costa C, Nascimento V, Souza V, Corado A, al e. SARS-CoV-2 reinfection by the new Variant of Concern (VOC) P.1 in Amazonas, Brazil. Virologicalorg 2021.

17. Nonaka C, Franco M, Graf T, Mendes A, de Aguiar R, al e. Genomic evidence of a SARS-CoV-2 reinfection case with E484K spike mutation in Brazil. preprintsorg 2021.

18. Resende P, Bezerra J, de Vasconcelos R, Arantes I, Appolinario L, al e. Spike E484K mutation in the first SARS-CoV-2 reinfection case confirmed in Brazil, 2020. virologicalorg 2021.

19. Wu K, Werner A, Molivs J, al e. mRNA-1273 vaccine induces neutralizing antibodies against spike mutations from global SARS-CoV-2 variants. . bioRxiv 2021.

20. Xie X, Zou J, Fnter-Garfias, al e. Neutralization of N501Y mutant SARS-CoV-2 by BNT162b2 vaccine-elicted sera. bioRxiv 2021.

21. Collier DA, Meng B, Ferreira I, Datir R. Impact of SARS-CoV-2 B.1.1.7 Spike variant on neutralisation potency of sera from individuals vaccinated with Pfizer vaccine BNT162b2. . MedRxiv 2021.

22. Consortium CUC-GU. COG-UK report on SARS-CoV-2 spike mutations of interest in the UK2021 01/15/2021.

23. Washington N, White S, Schiabor Barrett K, al e. S gene dropout patterns in SARS-CoV-2 tests suggest repread of the H69/V70del mutation in the US. medRxiv 2020.

24. Shu Y, J M. GISAID: Global initiative on sharing all influenza data - from vision to reality. EuroSurveillance 2017;22.

25. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. Preprint 2020;2020060225.

26. Rambaut A, Holmes E, O'Toole A, al e. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403-7.

27. Santos J, GA P. The high infectivity of SARS-CoV-2 B.1.1.7 is associated with increased interaction force between Spike-ACE2 caused by the viral N501Y mutation. biorxiv 2021.

28. Novavax. Novavax COVID-19 Vaccine Demonstrates 89.3% Efficacy in UK Phase 3 Trial. 2021.

29. Cheng M, Krieger J, Kaynak B, Arditi M, Bahar I. Impact of South African 501.V2 variant on SARS-CoV-2 spike infectivity and neutralization: A structure-based computational assessment. bioRxiv 2021.

30. Wibner C, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, al e. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. bioRxiv 2021.

31. Johnson J. Johnson & Johnson Announces Single-Shot Janssen COVID-19 Vaccine Candidate Met Primary Endpoints in Interim Analysis of its Phase 3 ENSEMBLE Trial. 2021.

32. Korber B, Fisher W, Gnankaran S, Yoon H, Theiler J, al e. Tracking changes in SARS-CoV-2 spike: evidence that D614G

increases infectivity of the COVID-19 virus

. Cell Mol Immunol 2020;182:812-7.

33. Anant S, Davidson N. Molecular mechanisms of apolipoprotein B mRNA editing. Curr Opin Lipidol 2001;12:159-65.

34. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for hostdependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv 2020;17.

35. Simmonds P. Rampant C \rightarrow U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: Causes and consequences for their short- and long-term evolutionary trajectories. mSphere 2020.

36. Xie X, Liu Y, J L, Zhang X, Zou J, al e. Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K, and N501Y variants by BNT162b2 vaccine-elicited sera. biorxiv 2021.

37. B H, Zhong E, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2021;371.

Legends to Figures and Tables

Figure 1:

Global Mutations/Variants over time from April 1, 2020 to January 20, 2021 for all countries. Complete sequences with known dates of accession were downloaded from GISAID on January 20, 2021. Months were assessed from the submission data and subsetted on those with reported sampling dates.

Table 1:

Mutations associated with variants B.1.1.7 (UK), B.1.135 (South Africa), P.1 (Brazil) and P.2 (Brazil).

Table 2

Variants of SARS-CoV-2 by country as of January 20, 2021. Currently, new variants are being detected and characterized in rapid succession. This Table could be outdated by the time of publication. For updating of data consult GISAID²⁴.

Table 3 – United Kingdom

Details of the mutant analyses of 7,144 SARS-CoV-2 isolates for deviations from the Wuhan reference sequence. These sequences were deposited in the GISAID initiative between 01/19/2020 and 01/20/2021. For design of Tables see legend to Table 4.

Table 4 – South Africa

The Table presents characteristics of SARS-CoV-2 mutants from South African isolates. For Table design, see legend to Table 6.

Table 5 - Brazil

The Table presents characteristics of SARS-CoV-2 mutants from isolates collected in the Brazilian population. For Table design, see legend to Table 6.

Table 6 - USA

The general design of this Table is similar to Tables 3 to 5, and 7 to 12, with minor modifications. **Part A:** From the overall analyses of the entire SARS-CoV-2 RNA sequence from 112 (US-I), 97 (US-II), 99 (US-III), and 117 (US-IV) randomly chosen isolates, the mutated nucleotides (nt) – as compared to the original Wuhan sequence – were tabulated. The actual time periods of mutant selections for the US-I to US-IV samples were indicated. Mutations previously designated as "signal hotspots" [Weber et al, 2020, i.e. 241 - 1,059 - 1,440 - 2,891 - 3,037 - 8,782 - 14,408 - 23,403 - 25,563 - 28,144 - 28,881] were now designated "prevalent". The * in the US-I and US-II columns designates previous publication in [Weber et al. 2020^1]. The actual nucleotide changes were indicated in the third column, the most frequent being C \rightarrow T (here 59 %), as reported previously [Simmonds et al. 2020^{35} ; Weber et al. 2020^1]. Locations of mutations were tabulated in columns 2 and 3, respectively. In columns 4 to 7, the actual frequencies of mutations at the four time intervals (US-1 to US-IV) were listed. The following designations for individual countries were chosen: BR for Brazil, CN for China, DE for Germany, FR for France, IN for India, RU for Russia, ES for Spain, ZA for South Africa, UK for United Kingdom, and US for USA.

The GGG \rightarrow AAC is a non-point mutation in nucleotide position 28,881 that generated a highly basic amino acid sequence in the SARS-CoV-2 nucleocapsid phosphor-protein. We have speculated that this mutation might have originated from a recombination event between different viral RNA molecules [Weber et al, 2020¹]

Part B: A total of 5,710 SARS-CoV-2 from the GISAID source was analyzed. Deviations from the Wuhan reference sequence of >2% incidence were found at 42 sites in the sequence. Further details were described in the text.

Table 7 – India, Table 8 – Russia, Table 9 – France, Table 10 – Spain, Table 11 – Germany, Table 12 - China

The general design of these Tables follows the outline described in detail in the legend to Table 6 (USA). The number of sequences investigated for SARS-CoV-2 mutations was detailed in Tables for individual countries.

Table 13 (Survey)

The rise of new SARS-CoV-2 mutations in many countries was juxtaposed to the high Covid-19 incidence values around the world. World incidence of Covid-19, as of January 30, 2021 in 219 Countries was Covid-19 cases – 102,87 million, fatalities – 2.22 million (columns 10 and 11). Column 5 lists the total of novel mutations for each country, percentage values related this sum to the total number of mutations. Source for worldwide spread of Covid-19 - <u>https://www.worldometers.info/coronavirus/</u>

The UK data in this Table do not contain results from the analysis of the SARS-CoV-2 variant B.1.1.7 which were shown in Table 1, as of January 20, 2021.



| | B. 1. 1.7 | B.1.135 | P.1 | P.2 |
|----------------|--|--|--------------------------------|----------------------------|
| Gene | Mutation | Mutation | Mutation | Mutation |
| | T1001I | | | |
| ORF1ab | A1708D | | | |
| medRxiv prepri | I2230T nt doi: https://doi.org/10.1101/ | 2021.02.04.21251111; this version posted | February 6, 2021, The copyrigh | t holder for this preprint |
| nsp5 | It is mad | e available under a CC-BY-NC-ND 4.0 Int | ernational license. | L205V |
| nsp6 | | | | |
| | H69/V70 del | L18F | | |
| | Y144 del | D80A | | |
| | N501Y | D215G | | |
| Snike | A570D | R246I | | |
| брікс | P681H | K417N | K417N | |
| | T716I | E484K | E484K | E484K |
| | S982A | N501Y | N501Y | |
| | D118H | A701Y | | V1176F |
| | Q27stop | | | |
| Orf8 | R52I | | | |
| | Y73C | | | |
| | D3L | | | A119S |
| Nucleocansid | S235F | | | R203K |
| Tuctocapsiu | | | | G204R |
| | | | | M234I |

Table 1 Mutations associated with variants B.1.1.7, B.1.135, P.1 and P.2

| | B.1.1.7 (UK | | B.1.1.28 | |
|---|---|---|--|--------------------|
| Country | variant) | B.1.351 (South Africa Variant) | (E484K.V2) | |
| ArgentinadRxiv preprint doi: (which was not co | https://doi.org/1 ⊉ .1101/2021.02. rtified by peer review) is the a | 04.21251111; this version posted February 6, 20 uthor/funder, who has granted medRxiv a licens | 21. The copyri 5 ht holder for this e to display the preprint in perpe | preprint tuity. |
| Australia | 52 | He under a CC-BY-NC-ND 4.0 International licen 4 | se. 0 | |
| Austria | 41 | 0 | 0 | |
| Bangladesh | 3 | 0 | 0 | |
| Belgium | 194 | 12 | 0 | |
| Botswana | 0 | 7 | 0 | |
| Brazil | 15 | 0 | 175 | |
| Canada | 2 | 0 | 13 | |
| China | 2 | 0 | 0 | |
| Czech Republic | 12 | 0 | 0 | |
| Denmark | 531 | 3 | 6 | |
| Faroe Islands | 0 | 0 | 1 | |
| Dominican Republic | 1 | 0 | 0 | |
| Ecuador | 1 | 0 | 0 | |
| Finland | 40 | 2 | 0 | |
| France | 221 | 16 | 0 | |
| Germany | 618 | 82 | 4 | |
| Gibraltar | 1 | 0 | 0 | |
| Greece | 5 | 0 | 0 | |
| Hungary | 5 | 0 | 0 | |
| Iceland | 20 | 0 | 0 | |
| India | 23 | 0 | 0 | |
| Iran | 1 | 0 | 0 | |
| Ireland | 159 | 7 | 4 | |
| Israel | 99 | 0 | 0 | |
| Italy | 133 | 0 | 5 | |
| Jamaica | 4 | 0 | 0 | |
| Japan | 36 | 6 | 9 | |
| Jordan | 38 | 0 | 0 | |
| Kenya | 0 | 2 | 0 | |
| Luxembourg | 32 | 2 | 1 | |
| Malaysia | 1 | 0 | 0 | |
| Mayotte | 18 | 0 | 0 | |

| Mexico | 1 | 0 | 0 | |
|-------------------------|---|---|--|-------------------|
| Mayotte | 0 | 21 | 0 | |
| Mozambiquech was not ce | nttps://doi.org/10.1101/2021.02. artified by peel review) is the a It is made availab | 04.21251111, this version posted February 6, 20 uthor/funder, who has granted medRxiv a licens le under a CC-BY-NC-ND 4.0 International licen | 21. The copyright holder for this perpetion of the perpetion of the preprint in perpetion of the perpetition of the pe | preprint uity. |
| Netherlands | 320 | 22 | 2 | |
| New Zealand | 19 | 6 | 0 | |
| Nigeria | 4 | 0 | 0 | |
| Norway | 30 | 1 | 3 | |
| Oman | 1 | 0 | 0 | |
| Pakistan | 5 | 0 | 0 | |
| Panama | 0 | 1 | 0 | |
| Peru | 1 | 0 | 0 | |
| Poland | 7 | 0 | 0 | |
| Portugal | 88 | 1 | 0 | |
| Romania | 9 | 0 | 0 | |
| Singapore | 38 | 0 | 1 | |
| Slovakia | 37 | 0 | 0 | |
| Slovenia | 2 | 0 | 0 | |
| South Africa | 0 | 490 | 0 | |
| South Korea | 13 | 1 | 1 | |
| Spain | 177 | 1 | 0 | |
| St. Lucia | 3 | 0 | 0 | |
| Sweden | 15 | 1 | 0 | |
| Switzerland | 183 | 11 | 1 | |
| Thailand | 2 | 0 | 0 | |
| Trinidad and Tobago | 1 | 0 | 0 | |
| Turkey | 50 | 0 | 0 | |
| United Arab Emirates | 21 | 5 | 0 | |
| United Kingdom | 36,643 | 74 | 15 | |
| USA | 333 | 1 | 24 | |
| Vietnam | 1 | 0 | 0 | |

 Table 2 – Variants by country as of February 1, 2021(https://www.gisaid.org/hcov19-variants)

| Tuble 5 Chi | teu Kiliguolli | | 01/13/2020 | - 01/20/2021 |
|---------------------------------------|--|------------------------------------|----------------|--------------------------------------|
| Position | Location | Mutation | Total Count | Percentage |
| preprint@toi: http h was2040tcerti | bs://doi.org/10.1101/2021.02.04.21251111; this version po fied by peer review) is the diffkor/funder, who has grante It is made available under a CC-BX-NC-ND 4. | sted February 6 d medRxiv a lic | 2024.7 The co | pyright holder f the 29 @7 int in |
| 241nt 445nt | $ORE1ab polyprotein \rightarrow leader protein$ | T→C | 34 505 | 96.81 48.3 |
| 1.163nt | ord fuo polyprotein / reader protein | $A \rightarrow T$ | 2,544 | 3.56 |
| 1,210nt | | G→T | 1,440 | 2.02 |
| 1,513nt | nsp2 | C→T | 1,528 | 2.14 |
| 1,947nt | | T→C | 1,576 | 2.21 |
| 1,987nt | | A→G | 3,018 | 4.22 |
| 3,037nt | | $C \rightarrow T$ | 69,231 | 96.91 |
| 3,256nt | | $T \rightarrow C$ | 2,523 | 3.53 |
| 4,002nt | | $C \rightarrow I$ | 1,519 | 2.13 |
| 4,343III 6 286nt | nsp3 | $C \rightarrow T$ | 34,650 | 2.12 |
| 6.807nt | | $C \rightarrow T$ | 2.220 | 3.11 |
| 7,528nt | | $C \rightarrow T$ | 1,524 | 2.13 |
| 7,926nt | | C→T | 2,818 | 3.94 |
| 8,683nt | | C→T | 2,189 | 3.06 |
| 9,745nt | nsp4 | $C \rightarrow T$ | 3,640 | 5.1 |
| 9,802nt | | G→T | 1,449 | 2.03 |
| 10,097nt | 3C-like proteinase | G→A | 2,954 | 4.13 |
| 10,870nt | | $G \rightarrow T$ | 3,186 | 4.46 |
| 11,083nt | | $G \rightarrow I$ | 5,/34 | 8.03 |
| 11,590iit 11,533nt | nsp6 | $C \rightarrow I$ | 2,280 | 2.74 |
| 11,555ht | | $A \rightarrow G$ | 2,368 | 3.31 |
| 12,067nt | nsp7 | $G \rightarrow T$ | 1,709 | 2.39 |
| 13,536nt | | $C \rightarrow T$ | 1,502 | 2.1 |
| 14,202nt | | G→T | 2,522 | 3.53 |
| 14,408nt | RNA-dependent RNA polymerase | $C \rightarrow T$ | 69,237 | 96.92 |
| 14,805nt | | C→T | 1,860 | 2.6 |
| 15,406nt | | $G \rightarrow T$ | 2,077 | 2.91 |
| 18,8//nt | 3'-to-5' exonuclease | $C \rightarrow I$ | 3,827 | 5.36 |
| 19,342m 19,718nt | | $G \rightarrow I$ | 2,382 | 3.01 |
| 20.268nt | endoRNAse | $A \rightarrow G$ | 1,999 | 2.8 |
| 21,255nt | 2'-O-ribose methyltransferase | G→C | 34,494 | 48.28 |
| 21,575nt | * | $C \rightarrow T$ | 1,502 | 2.1 |
| 21,614nt | | $C \rightarrow T$ | 17,561 | 24.58 |
| 21,637nt | | C→T | 2,697 | 3.78 |
| 22,227nt | | $C \rightarrow T$ | 34,855 | 48.79 |
| 22,346nt | | $G \rightarrow T$ | 2,244 | 3.14 |
| 22,37/nt | Spike glycoprotein | $C \rightarrow I$ | 1,518 | 2.12 |
| 22,388llt 22 444nt | | $C \rightarrow T$ | 2,340 | 2.92 |
| 22.992nt | | $G \rightarrow A$ | 1.636 | 2.29 |
| 23,403nt | | A→G | 69,262 | 96.95 |
| 23,731nt | | $C \rightarrow T$ | 2,940 | 4.12 |
| 24,334nt | | $C \rightarrow T$ | 10,442 | 14.62 |
| 25,563nt | | $G \rightarrow T$ | 5,774 | 8.08 |
| 25,614nt | ORF3a | $C \rightarrow T$ | 2,737 | 3.83 |
| 26,060nt | | $C \rightarrow I$ | 2,032 | 3.08 |
| 20,144IIt 26.424nt | Envelope protein | $T \rightarrow C$ | 1,740 | 2.43 |
| 26.735nt | Envelope protein | $C \rightarrow T$ | 3.760 | 5.26 |
| 26,801nt | Membrane glycoprotein | C→G | 34,459 | 48.24 |
| 27,769nt | ORF7b | $C \rightarrow T$ | 2,706 | 3.79 |
| 27,944nt | ORES | $C \rightarrow T$ | 25,177 | 35.24 |
| 28,169nt | UNIO | A→G | 2,693 | 3.77 |
| 28,854nt | | $C \rightarrow T$ | 3,683 | 5.16 |
| 28,881nt | | $G \rightarrow A$ | 23,975 | 33.56 |
| 28,882nt | | $G \rightarrow A$ | 23,947 | 33.52 |
| 20,00011 28 932nt | Nucleocapsid phosphoprotein | G→C C→T | 23,940 | 33.32 48.34 |
| 29.227nt | | $G \rightarrow T$ | 2.566 | 3.59 |
| 29,366nt | | $C \rightarrow T$ | 1,743 | 2.44 |
| 29,466nt | | $C \rightarrow T$ | 2,578 | 3.61 |
| 29,555nt | at upstream downstream region of ORF10 ORF9 | C→T | 1,466 | 2.05 |
| 29,645nt | ORF10 | $G \rightarrow T$ | 34,684 | 48.55 |
| 20.771mt | 31178 | $A \rightarrow G$ | 2 475 | 3 46 |

| Fable 4A - | South Africa | | 09/01 - 12/07/2020 | |
|------------|--|---|---|---|
| Position | Location | Mutation | _ Count | Incidence . |
| / preprint | doi: https://doi.org/10. ot certified by peer r | 1101/2021.02.04.21251111; this version posted Februa | ny 6, 2021. Th The license to div | e copyright holder for splay the mengint in pe |
| | 5 UTR It is | made available under an effective NC-ND 4.0 Internation | al license | play the <u>pr</u> opy in the pe |
| 241nt | 0 CIII | $CG \rightarrow TG$ | 95/95 | prevalent |
| | | $\frac{1}{1}$ | | * |
| 1,059nt | 2 | $\frac{ACC}{ACC}$ (Threonine) $\rightarrow ATC$ (Isoleucine) | 10/95 | prevalent |
| 2,164nt | nsp2 | GA → CA GAGAAG (Glutamic Acid Lysine) → GACAAG (Aspartic Acid Lysine) | 11/95 | IN |
| 3,037nt | | $CT \rightarrow TT$ noneffective | 95/95 | prevalent |
| 5,230nt | nsp3 | GT → TT AA <mark>GT</mark> GG (Lysine Tryptophan) → AA TT GG (Asparagine Tryptophan) | 12/95 | DE |
| 6,762nt | | $CT \rightarrow TT$ ACT (Threonine) \rightarrow ATT (Isoleucine) | 13/95 | unique |
| 10,323nt | 3C-like proteinase | $AG \rightarrow GG$ AAG (Lysine) $\rightarrow AGG$ (Arginine) | 11/95 | unique |
| 11,230nt | nsp6 | $\begin{array}{c} GC \rightarrow TC \\ ATGCCT (Methionine Proline) \rightarrow ATTCCT (Isoleucine \\ Proline) \end{array}$ | 11/95 | unique |
| 12,503nt | nsp8 | $TA \rightarrow CA$ TA T (Tyrosine) \rightarrow CA T (Histidine) | 26/95 | unique |
| 14,408nt | RNA-dependent RNA polymerase | $CT \to TT$ $CCT (Proline) \to CTT (Leucine)$ | 95/95 | prevalent |
| 20,268nt | endoRNAse | $AG \rightarrow GG$ | 21/95 | FR,ES,RU |
| 21,801nt | | $AT \rightarrow CT$ $CAT (Acception Acid) \rightarrow CCT (Algoring)$ | 10/95 | unique; S501Y.V2 |
| 22,675nt | | $CG \rightarrow TG$ | 10/95 | unique |
| 22,813nt | | $\begin{array}{c} \text{Hole the transformation}\\ \text{GA} \rightarrow \text{TA}\\ \text{Hole the transformation} \end{array}$ | 10/95 | DE |
| 23,012nt | Spike glycoprotein | $GA \rightarrow AA$ | 12/95 | IN |
| 23.403nt | | $\frac{GAA}{Glutamic Acid} \rightarrow AAA (Lysine)$ $AT \rightarrow GT$ | 95/95 | prevalent |
| 23.66/nt | | $\frac{\text{GAT (Aspartic Acid)} \rightarrow \text{GGT (Glycine)}}{\text{CA} \rightarrow \text{TA}}$ | 14/95 | FS IN |
| 25,00411 | | GCA (Alanine) $\rightarrow GTA$ (Valine) | 14/95 | 2.5,11 |
| 25,563nt | | $GA \rightarrow TA$ CAGAGC (Glutamine Serine) \rightarrow CATAGC (Histidine Serine) | 10/95 | prevalent |
| 25,770nt | ORF3a protein | $GC \rightarrow TC$ AGGCTT (Arginine Leucine) \rightarrow AGTCTT (Serine Leucine) | 20/95 | RU |
| 25,904nt | | $CA \rightarrow TA$ TCA (Serine) \rightarrow TTA (Leucine) | 10/95 | BR,DE |
| 26,456nt | Envelope protein | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT (Leucine)$ | 10/95 | unique |
| 28,253nt | ORF8 protein | $CA \rightarrow TA$ | 14/95 | BR,DE,ES,FR,US |
| 28,854nt | | $CA \rightarrow TA$ $TCA (Second) \rightarrow TTA (Levelag)$ | 23/95 | CN,DE,ES,FR,IN,RU |
| | Nucleocapsid | $GGG \rightarrow AAC$ | | |
| 28,881nt | phosphoprotein | $\begin{array}{c} A \textbf{GGGGGA} \text{ (Arginine Glycine)} \rightarrow A \textbf{AACGA} \text{ (Lysine} \\ A \textbf{rginine)} \end{array}$ | 61/95 | prevalent |
| 28,887nt | | $CT \rightarrow TT$ $ACT \text{ (Threonine)} \rightarrow ATT \text{ (Isoleucine)}$ | 11/95 | BR,CN,FR,IN,RU |
| 29 721nt | 3/IITR | $CC \rightarrow TC$ | 26/95 | unique |

| Table 4B - South Africa | | outh Africa | | 01/19/2020 - 01/20/2021 | | |
|-------------------------|-----------------------------------|--|-------------------------------------|----------------------------------|----------------------|-----------------------------------|
| | Position | Location | Mutation | Total Count | Percentage | |
| medRxiv p (which | reprih7 doi: htt was24 outcert | ps://doi.org/10.1101/2021.92,94,21251111; this version pc fied by peer review) is the author/funder, who has grante | sted February 6 d medRxiv a lice | 2021.87he co ense to7702splay | pyrid the preprint i | for this preprin n perpetuity. |
| | 355nt | ORTHREPOBLYCAUROUND UNDER PROTEINT - INC - IND 4. | u metraaiphani | cense 59 | 3.32 | |
| | 1,059nt | | $C \rightarrow T$ | 149 | 8.38 | |
| | 2,094nt | nsp2 | $C \rightarrow T$ | 38 | 2.14 | |
| | 2,164nt | а. - | $G \rightarrow C$ | 84 | 4.72 | |
| | 2,692nt | | $A \rightarrow I$ | 41 | 2.3 | |
| | 3,03/nt | | $C \rightarrow I$ | 1,746 | 98.15 | |
| | 4,002nt | | $C \rightarrow I$ | 165 | 9.27 | |
| | 4,093nt | nen2 | $C \rightarrow I$ | 48 | 2.1 | |
| | 5,230m | iisp5 | | 147 | 2 50 | |
| | 6.762nt | | $C \rightarrow T$ | 178 | 10.01 | |
| | 7.064nt | | $A \rightarrow G$ | 178 | 6.97 | |
| | 8 660nt | | $C \rightarrow T$ | 69 | 3.88 | |
| | 8.964nt | nsn4 | $C \rightarrow T$ | 69 | 3.88 | |
| | 9.498nt | | $T \rightarrow C$ | 36 | 2.02 | |
| | 10.097nt | | $G \rightarrow A$ | 163 | 9.16 | |
| | 10,323nt | 3C-like proteinase | A→G | 169 | 9.5 | |
| | 11.083nt | | G→T | 60 | 3.37 | |
| | 11,230nt | nsp6 | G→T | 75 | 4.22 | |
| | 11,447nt | L | G→A | 129 | 7.25 | |
| | 12,503nt | nsp8 | T→C | 389 | 21.87 | |
| | 13,536nt | | $C \rightarrow T$ | 170 | 9.56 | |
| | 14,408nt | RNA-dependent RNA polymerase | $C \rightarrow T$ | 1,773 | 99.66 | |
| | 14,925nt | | $C \rightarrow T$ | 71 | 3.99 | |
| | 16,376nt | | $C \rightarrow T$ | 54 | 3.04 | |
| | 16,490nt | Haliansa | $C \rightarrow T$ | 39 | 2.19 | |
| | 16,853nt | nencase | $G \rightarrow T$ | 47 | 2.64 | |
| | 16,946nt | | $C \rightarrow T$ | 43 | 2.42 | |
| | 18,747nt | 3'-to-5' exonuclease | $C \rightarrow T$ | 115 | 6.46 | |
| | 20,234nt | endoRNAse | $C \rightarrow T$ | 42 | 2.36 | |
| | 20,268nt | Chuoki Wise | A→G | 209 | 11.75 | |
| | 21,801nt | | A→C | 142 | 7.98 | |
| | 22,206nt | | A→G | 71 | 3.99 | |
| | 22,287nt | | $T \rightarrow A$ | 86 | 4.83 | |
| | 22,299nt | | $G \rightarrow T$ | 69 | 3.88 | |
| | 22,6/5nt | | $C \rightarrow T$ | 290 | 16.3 | |
| | 22,813nt | Spike glycoprotein | $G \rightarrow T$ | 139 | 7.81 | |
| | 23,012nt | | $G \rightarrow A$ | 146 | 8.21 | |
| | 23,003nt | | $A \rightarrow I$ | 140 | /.8/ | |
| | 23,405lit | | $A \rightarrow G$ | 52 | 99.01 | |
| | 23,023III 23,664nt | | $C \rightarrow T$ | 55 154 | 2.70 8.66 | |
| | 23,731nt | | $C \rightarrow T$ | 161 | 9.05 | |
| | 25.455nt | | $G \rightarrow T$ | 65 | 3.65 | |
| | 25.521nt | | C→T | 66 | 3.71 | |
| | 25,563nt | ORF3a | G→T | 148 | 8.32 | |
| | 25,770nt | | $G \rightarrow T$ | 285 | 16.02 | |
| | 25,904nt | | C→T | 143 | 8.04 | |
| | 26,456nt | Envelope protein | C→T | 140 | 7.87 | |
| | 26,586nt | Membrane glycoprotein | $C \rightarrow T$ | 62 | 3.49 | |
| | 27,384nt | ORF6 | T→C | 120 | 6.75 | |
| | 27,504nt | ORF7a | T→C | 50 | 2.81 | |
| | 28,077nt | OPEO | $G \rightarrow T$ | 74 | 4.16 | |
| | 28,253nt | υκγο | $C \rightarrow T$ | 178 | 10.01 | |
| | 28,854nt | | $C \rightarrow T$ | 173 | 9.72 | |
| | 28,881nt | | $G \rightarrow \overline{A}$ | 1,238 | 69.59 | |
| | 28,882nt | Nucleocapsid phosphoprotein | G→A | 1,238 | 69.59 | |
| | 28,883nt | ruereoeupsid phosphoptotem | G→C | 1,238 | 69.59 | |
| | 28,887nt | | $C \rightarrow T$ | 152 | 8.54 | |
| | 29,425nt | | $G \rightarrow T$ | 117 | 6.58 | |
| | 29,721nt | 3'UTR | $C \rightarrow T$ | 388 | 21.81 | |

| | Table 5A | - Brazil | | 02/25 - 08/15/2020 | | |
|------|------------------------|---|--|-----------------------|-----------------------------|--------------|
| medR | civ Pposepoin t | doi: https://deittorg/10.1101 | 2021.02.04.21251111; this wassion posted February 6, 202 | 1. Too popyrig | gh theide rdor i | his preprint |
| (wł | 241nt | ot certified by peer review 5 UTR It is mad |) is the author/funder, winchas granted medRxiv a license e available under a CC-BY-NC-ND 4.0 International license monether the second | to display the | preprint in pe prevalent | rpetuity. |
| | 3,037nt | nsp3 | $CT \rightarrow TT$ noneffective | 97/102 | prevalent | |
| | 12,053nt | nsp7 | $CT \rightarrow TT$ CT T (Leucine) \rightarrow TT T (Phenylalanine) | 16/102 | unique | |
| | 14,408nt | RNA-dependent RNA polymerase | $CT \rightarrow TT$ CCT (Proline) $\rightarrow CTT$ (Leucine) | 96/102 | prevalent | |
| | 23,403nt | Spika alvooprotoin | $AT \rightarrow GT$ GAT (Aspartic Acid) $\rightarrow GGT$ (Glycine) | 97/102 | prevalent | |
| | 25,088nt | Spike grycoprotein | $GT \rightarrow TT$ GTT (Valine) \rightarrow TTT (Phenylalanine) | 25/102 | unique | |
| | 27,299nt | ORF6 protein | $TA \rightarrow CA$ ATA (Isoleucine) $\rightarrow ACA$ (Threonine) | 41/102 | FR | |
| | 28,881nt | Nucleocapsid | $GGG \rightarrow AAC$ AGGGGA (Arginine Glycine) $\rightarrow AAACGA$ (Lysine Arginine) | 73/102 | prevalent | |
| | 29,148nt | phosphoprotein | $TC \rightarrow CC$ $ATC \text{ (Isoleucine)} \rightarrow ACC \text{ (Threonine)}$ | 41/100 | FR,RU | |

| Table 5B - Br | Table 5B - Brazil | | | - 01/20/2021 |
|-------------------|---|--------------------|-----------------------------|----------------------------|
| Position | Location | Mutation | Total Count | Percentage |
| reprint-506i: htt | ps://doi.org/10.1101/2021.02.04.21251111; this version po | sted February 6 | . 2021 ⁴³ The co | pyright ⁸ Polde |
| was not cert | fied by peer review) is the author/funder, who has grante | d međRxŵ a lice | ense t@3display | the preprint i |
| 100nt | It is made available under a CC-BY-NC-ND 4. |) International li | cense 97 | 8.77 |
| 241nt | | $C \rightarrow T$ | 1,087 | 98.28 |
| 3,037nt | | C→T | 1,093 | 98.82 |
| 3,766nt | nsp3 | T→C | 49 | 4.43 |
| 6,319nt | | A→G | 32 | 2.89 |
| 10,667nt | 3C-like proteinase | T→G | 98 | 8.86 |
| 11,083nt | | $G \rightarrow T$ | 29 | 2.62 |
| 11,824nt | lispo | $C \rightarrow T$ | 98 | 8.86 |
| 12,053nt | nsp7 | C→T | 318 | 28.75 |
| 12,964nt | nsp9 | A→G | 89 | 8.05 |
| 14,408nt | RNA-dependent RNA polymerase | $C \rightarrow T$ | 1,091 | 98.64 |
| 23,012nt | | G→A | 98 | 8.86 |
| 23,403nt | Spike glycoprotein | A→G | 1,093 | 98.82 |
| 25,088nt | | $G \rightarrow T$ | 463 | 41.86 |
| 26,149nt | ORF3a | T→C | 31 | 2.8 |
| 27,299nt | ORF6 | T→C | 459 | 41.5 |
| 28,253nt | ORF8 | $C \rightarrow T$ | 110 | 9.95 |
| 28,628nt | | $G \rightarrow T$ | 99 | 8.95 |
| 28,881nt | | G→A | 1,031 | 93.22 |
| 28,882nt | NT 1 1 1 1 / 1 | G→A | 1,031 | 93.22 |
| 28,883nt | nucleocapsia phosphoprotein | G→C | 1,031 | 93.22 |
| 28,975nt | | $G \rightarrow T$ | 101 | 9.13 |
| 29,148nt | | T→C | 466 | 42.13 |
| 29,754nt | | $C \rightarrow T$ | 95 | 8.59 |
| 29.861nt | 3 UIK | $G \rightarrow T$ | 33 | 2.98 |

| Table 6A - US | | | US-I 02/29 - 04/26/2020* | US-II 06/12 - 07/07/2020* | US-III 07/09 - 07/22/2020 | US-IV 08/01 - 12/01/2020 | |
|---------------|----------------------------------|--|-----------------------------|------------------------------|------------------------------|-----------------------------|----------------|
| Position | Location | Mutation | Count | Count | Count | Count | Incidence |
| 241nt | 5 UTR | $CG \rightarrow TG$ noneffective | 76/111 | 74/96 | 99/99 | 116/117 | prevalent |
| 1,059nt | | $CC \rightarrow TC$ ACC (Threenine) $\rightarrow ATC$ (Isoleucine) | 42/112 | 45/97 | 30/99 | 56/117 | prevalent |
| 1,917nt | nsp2 | $CT \rightarrow TT$ ACT (Threenine) $\rightarrow ATT$ (Isoleucine) | 0/112 | 11/97 | 0/99 | 0/117 | CN |
| 2,416nt | | $CA \rightarrow TA$ noneffective | 9/112 | 4/97 | 1/99 | 3/117 | CN,ES,FR,RU,ZA |
| 3,037nt | | $CT \rightarrow TT$ noneffective | 75/112 | 72/97 | 99/99 | 117/117 | prevalent |
| 3,871nt | | $GA \rightarrow TA$ AAGATC (Lysine Isoleucine) $\rightarrow AATATC$ (Asparagine Isoleucine) | 0/112 | 0/97 | 29/99 | 4/117 | FR,ZA |
| 3,931nt | | $TG \rightarrow CG$ noneffective | 0/112 | 0/97 | 29/99 | 4/117 | unique |
| 4,226nt | nsp3 | $CC \to TC$ $CCA (Proline) \to TCA (Serine)$ | 0/112 | 0/97 | 28/99 | 0/117 | unique |
| 5,672nt | | $CC \rightarrow TC$ $CCT (Proline) \rightarrow TCT (Serine)$ | 0/112 | 0/97 | 28/99 | 0/117 | unique |
| 7,837nt | | $AG \rightarrow CG$ TTAGAC (Leucine Aspartic Acid) \rightarrow TTCGAC (Phenylalanine Aspartic Acid) | 0/112 | 0/97 | 28/99 | 0/117 | CN |
| 8,083nt | | $GG \rightarrow AG$ ATGGAA (Methionine Glutamic Acid) \rightarrow ATAGAA (Isoleucine Glutamic Acid) | 0/112 | 0/97 | 0/99 | 18/117 | unique |
| 8,782nt | nsp4 | $CC \rightarrow TC$ noneffective | 15/112 | 15/97 | 0/99 | 0/117 | CN,DE,ES,IN |
| 10,139nt | 3C-like proteinase | $CT \rightarrow TT$ $CTT (Leucine) \rightarrow TTT (Phenylalanine)$ | 0/112 | 0/97 | 0/99 | 29/117 | unique |
| 12,025nt | nsp7 | $CA \rightarrow TA$ noneffective | 0/112 | 0/97 | 11/99 | 2/117 | unique |
| 14,408nt | RNA-dependent RNA polymerase | $CT \to TT$ $CCT (Proline) \to CTT (Leucine)$ | 78/112 | 71/97 | 99/99 | 117/117 | prevalent |
| 17,747nt | helicase | $CT \to TT$ $CCT (Proline) \to CTT (Leucine)$ | 8/112 | 12/97 | 0/99 | 0/117 | FR |
| 17,858nt | heneuse | $AT \to GT$ TAT (Tyrosine) \to TGT (Cysteine) | 8/112 | 12/97 | 0/99 | 0/117 | ZA |
| 18,060nt | | $CT \rightarrow TT$ noneffective | 9/112 | 11/97 | 0/99 | 0/117 | ZA |
| 18,424nt | 3'- to – 5'exonuclease | $AA \rightarrow GA$ AAT (Asparagine) $\rightarrow GAT$ (Aspartic Acid) | 0/112 | 0/97 | 0/99 | 26/117 | unique |
| 18,486nt | | $CA \rightarrow TA$ noneffective | 0/112 | 0/97 | 13/99 | 2/117 | unique |
| 18,877nt | | $CT \rightarrow TT$ noneffective | 13/112 | 1/97 | 6/99 | 3/117 | BR,DE,ES,FR,IN |
| 19,677nt | | $GG \rightarrow TG$ CAGGGT (Glutamine Glycine) \rightarrow CATGGT (Histidine Glycine) | 0/112 | 0/97 | 26/99 | 0/117 | unique |
| 19,839nt | endoRNAse | $TA \rightarrow CA$ noneffective | 0/112 | 0/97 | 11/99 | 7/117 | CN,DE,ES,FR,RU |
| 20,268nt | | $AG \rightarrow GG$ noneffective | 2/112 | 5/97 | 15/99 | 29/117 | FR,ES,RU,ZA |
| 21,304nt | 2'-O-ribose methyltransferase | $CG \to TG$ $CGC (Arginine) \to TGC (Cysteine)$ | 0/112 | 0/97 | 0/99 | 25/117 | ES |
| 22,162nt | | $\begin{array}{c} TT \rightarrow CT \\ noneffective \end{array}$ | 0/112 | 0/97 | 13/99 | 2/117 | unique |
| 23,403nt | Spike glycoprotein | $AT \rightarrow GT$ GAT (Aspartic Acid) \rightarrow GGT (Glycine) | 77/112 | 72/97 | 99/99 | 117/117 | prevalent |
| 23,707nt | | CA → TA noneffective | 0/112 | 0/97 | 11/99 | 3/117 | unique |
| 25,907nt | ORF3a protein | $GT \to TT$ $GGT (Glycine) \to GTT (Valine)$ | 0/112 | 0/97 | 0/99 | 26/117 | unique |
| 25,563nt | Protom | $GA \rightarrow TA$ CAGAGC (Glutamine Serine) \rightarrow CATAGC (Histidine Serine) | 65/112 | 54/97 | 37/99 | 66/117 | prevalent |
| 27,964nt | ORF8 protein | $CA \rightarrow TA$ $TCA (Serine) \rightarrow TTA (Leucine)$ | 13/112 | 6/97 | 4/99 | 31/117 | unique |
| | | | | US-II 06/12 - | US-III 07/09 - | US-IV 08/01 - | 1 |

| Table 5A - | - US | | US-I 02/29 - 04/26/2020* | 07/07/2020* | 07/22/2020 | 12/01/2020 | |
|------------|-----------------------------|---|-----------------------------|-------------|------------|------------|-------------------|
| Position | Location | Mutation | Count | Count | Count | Count | Incidence |
| 28,144nt | ORF8 protein | $TA \to CA$ $TTA (Leucine) \to TCA (Serine)$ | 15/112 | 15/97 | 0/99 | 0/117 | CN,DE,ES,IN |
| 28,472nt | | $CC \to TC$ CC T (Proline) \to TC T (Serine) | 0/112 | 0/97 | 0/99 | 22/117 | unique |
| 28,821nt | | $CT \rightarrow AT$ $TCT (Serine) \rightarrow TAT (Tyrosine)$ | 0/112 | 0/97 | 9/99 | 5/117 | unique |
| 28,854nt | | $CA \rightarrow TA$ $TCA (Serine) \rightarrow TTA (Leucine)$ | 3/112 | 0/97 | 13/99 | 28/117 | CN,DE,ES,FR,IN,RU |
| 28,869nt | Nucleocapsid phosphoprotein | $CA \rightarrow TA$ CCA (Proline) $\rightarrow CTA$ (Leucine) | 0/112 | 0/97 | 0/99 | 25/117 | DE |
| 28,881nt | | $GGG \rightarrow AAC$ AGGGGA (Arginine Glycine) $\rightarrow AAACGA$ (Lysine Arginine) | 3/112 | 1/97 | 17/99 | 17/117 | prevalent |
| 28,887nt | | $CT \to TT$ ACT (Threenine) $\to ATT$ (Isoleucine) | 0/112 | 1/97 | 1/99 | 10/117 | BR,CN,FR,IN,RU |
| 28,977nt | | $CT \rightarrow TT$ $TCT (Serine) \rightarrow TTT (Phenylalanine)$ | 0/112 | 0/97 | 29/99 | 4/117 | CN |

| | Table 6B - U | 18 | | 01/19/2020 | - 01/20/2021 | |
|-----------|---|---|--|---------------------------|------------------------------|-------------------|
| | Position | Location | Mutation | Total Count | Percentage | |
| nedRxiv p | reprint@toi: htt | ps://doi.org/10.1101/2021.02.04.21251111; this version po | sted February 6 | 202 ^{4,1} The co | pyrig ht211 older | for this preprint |
| (which | was2Abttcert | fied by peer review) is the author/funder, who has grante | d me d R x il⁄ a lic | ense48082ti6splay | the pareparint in | n perpetuity. |
| | 833nt | It is made available under a CC-BY-NC-ND 4. | 0 International II | cense,171 | 2.21 | |
| | 1,059nt | 11392 | C→T | 28,844 | 54.49 | |
| | 3,037nt | nsn3 | C→T | 49,077 | 92.71 | |
| | 8,083nt | 11315 | G→A | 2,779 | 5.25 | |
| | 8,782nt | nsp4 | C→T | 2,798 | 5.29 | |
| | 10,319nt | | $C \rightarrow T$ | 8,465 | 15.99 | |
| | 10,323nt | 3C-like proteinase | A→G | 1,176 | 2.22 | |
| | 10,741nt | | $C \rightarrow T$ | 1,120 | 2.12 | |
| | 11,083nt | nsp6 | $G \rightarrow T$ | 1,612 | 3.05 | |
| | 11,916nt | nsp7 | $C \rightarrow T$ | 1,670 | 3.15 | |
| | 14,408nt | DNA dependent DNA polymerese | $C \rightarrow T$ | 49,140 | 92.83 | |
| | 14,805nt | KINA-uependent KINA polymerase | $C \rightarrow T$ | 3,176 | 6 | |
| | 16,260nt | | $C \rightarrow T$ | 1,797 | 3.39 | |
| | 17,747nt | Helicase | $C \rightarrow T$ | 2,049 | 3.87 | |
| | 17,858nt |] | A→G | 2,084 | 3.94 | |
| | 18,060nt | | $C \rightarrow T$ | 2,135 | 4.03 | |
| | 18,424nt | 3'-to-5' exonuclease | A→G | 6,708 | 12.67 | |
| | 18,877nt | | $C \rightarrow T$ | 1,517 | 2.87 | |
| | 19,839nt | | T→C | 1,955 | 3.69 | |
| | 20,268nt | endoKNAse | A→G | 6,742 | 12.74 | |
| | 21,304nt | 2'-O-ribose methyltransferase | $C \rightarrow T$ | 6,603 | 12.47 | |
| | 23,403nt | · · · · · · · · · · · · · · · · · · · | A→G | 49,154 | 92.86 | |
| | 23,604nt | Spike glycoprotein | C→A | 1,238 | 2.34 | |
| | 24,076nt | | T→C | 2,148 | 4.06 | |
| | 25,563nt | ODE2 | G→T | 31,241 | 59.02 | |
| | 25,907nt | OKF3a | G→T | 6,369 | 12.03 | |
| | 27,964nt | ODE0 | $C \rightarrow T$ | 12,002 | 22.67 | |
| | 28,144nt | ORF8 | T→C | 2,790 | 5.27 | |
| | 28,472nt | | $C \rightarrow T$ | 6,473 | 12.23 | |
| | 28,821nt | | C→A | 1,821 | 3.44 | |
| | 28,842nt | | G→T | 1,152 | 2.18 | |
| | 28.854nt | | C→T | 6.694 | 12.65 | |
| | 28.869nt | 1 | C→T | 6.640 | 12.54 | |
| | 28,881nt | Nucleocapsid phosphoprotein | G→A | 6,887 | 13.01 | 1 |
| | 28,882nt | 1 | G→A | 6,848 | 12.94 | 1 |
| | 28,883nt | 1 | G→C | 6,847 | 12.93 | 1 |
| | Position eprinf@tbi: https://doi.o was24bbtcertified by 1 833nt 1,059nt 3,037nt 8,083nt 8,782nt 10,319nt 10,323nt 10,741nt 11,083nt 11,916nt 14,408nt 14,408nt 14,805nt 16,260nt 17,747nt 17,858nt 18,060nt 18,424nt 18,877nt 19,839nt 20,268nt 21,304nt 23,403nt 25,563nt 25,907nt 27,964nt 28,81nt 28,869nt 28,881nt 28,883nt 28,883nt 28,883nt 28,883nt 28,887nt 29,784nt 29,870nt | 1 | C→T | 1,090 | 2.06 | 1 |
| | 29,402nt | 1 | G→T | 1,630 | 3.08 | 1 |
| | 29,784nt | 0.47 JPD | C→T | 1,062 | 2.01 | 1 |
| | 29.870nt | 3´UTR | C→A | 1,990 | 3.76 | 1 |

| Table 7A - | India | | IN-I 01/27 - 05/27/2020* | IN-II 06/03 07/04/2020 | |
|------------|------------------------------|---|-----------------------------|---------------------------|----------------------------------|
| Position | Location | Mutation. | Count | Count | Incidence |
| (24high w | as not certifiet kby peer re | view) is the author/funder, who has granted medF | xiv a koense to | displays/one pre | prin piovaena etuity. |
| 2,292nt | nsp2 | $AG \rightarrow CG$ $CAG (Glutamine) \rightarrow CCG (Proline)$ | 0/99 | 22/98 | unique |
| 2,836nt | | $CT \rightarrow TT$ noneffective | 23/99 | 44/98 | unique |
| 3,037nt | | $CT \rightarrow TT$ noneffective | 81/99 | 96/98 | prevalent |
| 3,634nt | | $CA \rightarrow TA$ noneffective | 8/99 | 17/98 | ZA |
| 4,084nt | nsp5 | $CA \rightarrow TA$ noneffective | 12/99 | 1/98 | ZA |
| 4,300nt | | $GC \rightarrow TC$ noneffective | 0/99 | 16/98 | unique |
| 6,312nt | | $CA \rightarrow AA$ ACA (Threonine) $\rightarrow AAA$ (Lysine) | 10/99 | 0/98 | US |
| 11,083nt | nsp6 | GT → TT TTGTAT (Leucine Tyrosine) → TTT (Phenylalanine) | 13/99 | 0/98 | BR,CN,DE, ES,FR,US,Z A |
| 14,408nt | RNA-dependent RNA | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT (Leucine)$ | 80/99 | 91/98 | prevalent |
| 15,324nt | polymerase | $CA \rightarrow TA$ noneffective | 7/99 | 18/98 | B,C,G,F |
| 16,512nt | helicase | $AT \rightarrow GT$ noneffective | 0/99 | 11/98 | unique |
| 18,568nt | | $CT \rightarrow TT$ CT C (Leucine) \rightarrow TT C (Phenylalanine) | 0/99 | 22/98 | unique |
| 18,877nt | 3'- to - 5'exonuclease | $CT \rightarrow TT$ noneffective | 45/99 | 51/98 | BR,DE,ES, FR,US |
| 19,154nt | | $CA \rightarrow TA$ ACA (Threonine) $\rightarrow ATA$ (Isoleucine) | 0/99 | 12/98 | unique |
| 21,724nt | | GT → TT TTGTTC (Leucine Phenylalanine) → TTTTTC (Phenylalanine Phenylalanine) | 6/99 | 23/98 | RU |
| 22,444nt | Spike glycoprotein | $CC \rightarrow TC$ noneffective | 26/99 | 48/98 | US |
| 23,403nt | | $AT \rightarrow GT$ GAT (Aspartic Acid) \rightarrow GGT (Glycine) | 80/99 | 96/98 | prevalent |
| 23,929nt | | $CA \rightarrow TA$ noneffective | 10/99 | 0/98 | FR,RU,US |
| 25,563nt | ORF3a protein | GA → TA CAGAGC (Glutamine Serine) → CATAGC (Histidine Serine) | 43/99 | 51/98 | prevalent |
| 26,735nt | Membrane glycoprotein | CA → TA noneffective | 39/99 | 49/98 | DE,ES,FR, US |
| 28,311nt | | $CC \rightarrow TC$ $CCC (Proline) \rightarrow CTC (Leucine)$ | 10/99 | 0/98 | unique |
| 28,854nt | phosphoprotein | $CA \rightarrow TA$ $TCA \text{ (Serine)} \rightarrow TTA \text{ (Leucine)}$ | 29/99 | 41/98 | CN,DE,ES, FR,RU,US,Z A |

| Table 7B - Ind | ia | | 01/19/2020 | - 01/20/2021 | |
|----------------------|---|-----------------------|-----------------|-------------------|--------------|
| Position | Location | Mutation | Total Count | Percentage | |
| prepriat deli: http: | s://doi.org/10.1101/2021.52.07.21251111: this version | n posted February 6 | 2024.81the co | pvridh holder | for this pre |
| h was3nontcertif | ed by peep Review polyphetauthold under owno has gr | anted me@Rxit/ a lic | ense tø4displag | the 2016 point in | n perpetuit |
| 1,947nt | It is made available under a CC-BY-NC-N | D 4.0 International I | cense100 | 3.05 | |
| 2,292nt | nsp2 | A→C | 73 | 2.23 | |
| 2,836nt | | C→T | 281 | 8.57 | |
| 3,037nt | | $C \rightarrow T$ | 2,824 | 86.18 | |
| 3,634nt | | $C \rightarrow T$ | 283 | 8.64 | |
| 4,300nt | | G→T | 70 | 2.14 | |
| 4,354nt | nsp3 | G→A | 227 | 6.93 | |
| 4.372nt | 1 | A→G | 72 | 2.2 | |
| 5,700nt | | C→A | 949 | 28.96 | |
| 6.312nt | | C→A | 302 | 9.22 | |
| 6,573nt | | C→T | 228 | 6.96 | |
| 8,782nt | | C→T | 74 | 2.26 | |
| 8.917nt | nsp4 | $C \rightarrow T$ | 122 | 3.72 | |
| 9.693nt | Ĩ | C→T | 156 | 4.76 | |
| 11.083nt | nsp6 | G→T | 369 | 11.26 | |
| 13.730nt | nop o | $C \rightarrow T$ | 332 | 10.13 | |
| 14.408nt | RNA-dependent RNA polymerase | $C \rightarrow T$ | 2.768 | 84.47 | |
| 15.324nt | | C→T | 285 | 8.7 | |
| 16.626nt | Helicase | C→T | 143 | 4.36 | |
| 18,568nt | | $C \rightarrow T$ | 71 | 2.17 | |
| 18,877nt | 3'-to-5' exonuclease | $C \rightarrow T$ | 654 | 19.96 | |
| 19.524nt | | C→T | 69 | 2.11 | |
| 21.550nt | | A→C | 115 | 3.51 | |
| 21,551nt | 2'-O-ribose methyltransferase | $A \rightarrow T$ | 112 | 3.42 | |
| 21,724nt | | $G \rightarrow T$ | 109 | 3.33 | |
| 22.444nt | | $C \rightarrow T$ | 507 | 15.47 | |
| 22.468nt | Spike glycoprotein | G→T | 76 | 2.32 | |
| 23.403nt | ~F 9-7 - F | A→G | 2.832 | 86.42 | |
| 23.929nt | | C→T | 298 | 9.09 | |
| 25.528nt | | C→T | 222 | 6.77 | |
| 25.563nt | ORF3a | G→T | 652 | 19.9 | |
| 26.735nt | Membrane glycoprotein | $C \rightarrow T$ | 654 | 19.96 | |
| 27.384nt | ORF6 | T→C | 77 | 2.35 | |
| 28,144nt | ORF8 | T→C | 73 | 2.23 | |
| 28.311nt | | $C \rightarrow T$ | 299 | 9.12 | |
| 28,854nt | | C→T | 541 | 16.51 | |
| 28.878nt | | G→A | 70 | 2.14 | |
| 28.881nt | Nucleocapsid phosphoprotein | G→A | 1.434 | 43.76 | |
| 28.882nt | | $G \rightarrow A$ | 1.430 | 43.64 | |
| 28,883nt | | $G \rightarrow C$ | 1.430 | 43.64 | |
| 29,474nt | | $G \rightarrow T$ | 72 | 2.2 | |
| 29.750nt | 3´UTR | $C \rightarrow T$ | 74 | 2.26 | |
| 29.868nt | | G→A | 351 | 10.71 | |
| 20.870nt | at downstream region of ORF10 | $C \rightarrow A$ | 154 | 47 | |

| | Table 8A - | Russia | | 03/24 - 06/07/2020 | | |
|-----------|----------------|--------------------------------|---|------------------------------------|----------------|-------------------|
| | Position | Location | Mutation | Count | Incidence | f |
| meakxiv p | reprint doi: r | ttps://doi.org/10.1101/2021.02 | | <u>, 2021. The co</u> | pyright holder | for this preprint |
| (which | was4not ce | rtified by peerreview) is the | author/funder, who has granted medRxiv a lice | ens <u>e</u> tto∕ <u>¢tis</u> play | the preprint i | h perpetuity. |
| | | It is made availa | ble under a CC-BNONCERTY 4.0 International li | cense | 1 | 1 |
| | 0.005 | | $CT \rightarrow TT$ | 224/226 | | 1 |

| 3,037nt | | $CT \rightarrow TT$ noneffective | 224/226 | prevalent |
|----------|------------------------------|--|---------|-----------------|
| 3,140nt | nsp3 | $CC \rightarrow TC$ $CCT (Proline) \rightarrow AATCTT (Asparagine Leucine)$ | 13/226 | unique |
| 14,408nt | RNA-dependent RNA polymerase | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT (Leucine)$ | 225/226 | prevalent |
| 20,268nt | endoRNAse | $AG \rightarrow GG$ noneffective | 32/226 | ES,FR,US, ZA |
| 23,403nt | Spike glycoprotein | $AT \rightarrow GT$ GAT (Aspartic Acid) \rightarrow GGT (Glycine) | 226/226 | prevalent |
| 25,563nt | ORF3a protein | GA → TA CAGAGC (Glutamine Serine) → CATAGC (Histidine Serine) | 10/226 | prevalent |
| 26,750nt | Membrane glycoprotein | $CA \rightarrow TA$ noneffective | 45/226 | unique |
| 27,415nt | ORF6 protein | $GC \to TC$ $GCA \text{ (Alanine)} \to TCA \text{ (Serine)}$ | 10/226 | unique |
| 28,881nt | Nucleocapsid phosphoprotein | $GGG \rightarrow AAC$ AGGGGA (Arginine Glycine) $\rightarrow AAACGA$ (Lysine Arginine) | 172/226 | prevalent |

| | Table 8B - R | ussia | | 01/19/2020 | | |
|-----------|-------------------------------|---|---------------------------|-----------------------------|-------------------------------|------------------|
| | Position | Location | Mutation | Total Count | Percentage | |
| medRxiv p | reprin ^{a0} toi: htt | ps://doi.org/10.1101/2021.02.04.21251111; this version po | sted ₽eb fuarv 6 | . 2021 ⁵ 7The co | pvriatht ⁷ faolder | for this preprin |
| (which | was2Aottcert | fied by peer review) is the author/funder, who has grante | d me ɗR xíiv a lic | ense to dī splay | the <i>preprint</i> i | n perpetuity. |
| | 1,059nt | It is made available under a CC-BY-NC-ND 4. |) Intematipnal li | cense 31 | 2.59 | |
| | 3,037nt | | $C \rightarrow T$ | 1,188 | 99.08 | |
| | 3,177nt | | $C \rightarrow T$ | 28 | 2.34 | |
| | 3,373nt | nsp3 | C→A | 43 | 3.59 | |
| | 6,874nt | | T→G | 72 | 6.01 | |
| | 6,883nt | | $C \rightarrow T$ | 38 | 3.17 | |
| | 8,887nt | nsp4 | A→G | 108 | 9.01 | |
| | 11,029nt | nané | G→A | 41 | 3.42 | |
| | 11,083nt | lispo | $G \rightarrow T$ | 32 | 2.67 | |
| | 12,316nt | nsp8 | A→G | 28 | 2.34 | |
| | 12,886nt | nsp9 | A→G | 39 | 3.25 | |
| | 13,599nt | - | T→C | 63 | 5.25 | |
| | 14,408nt | RNA-dependent RNA polymerase | $C \rightarrow T$ | 1,180 | 98.42 | |
| | 15,540nt | | $C \rightarrow T$ | 29 | 2.42 | |
| | 19,839nt | 1 DN14 | T→C | 105 | 8.76 | |
| | 20,268nt | endoKNAse | A→G | 47 | 3.92 | |
| | 21,724nt | | G→A | 38 | 3.17 | |
| | 21,772nt | | $C \rightarrow T$ | 41 | 3.42 | |
| | 22,020nt | Spike glycoprotein | T→C | 73 | 6.09 | |
| | 23,403nt | | A→G | 1,195 | 99.67 | |
| | 25,563nt | ORF3a | $G \rightarrow T$ | 43 | 3.59 | |
| | 26,750nt | Membrane glycoprotein | $C \rightarrow T$ | 53 | 4.42 | |
| | 27,415nt | ORF7a | G→T | 34 | 2.84 | |
| | 28,253nt | ORF8 | $C \rightarrow T$ | 32 | 2.67 | |
| | 28,881nt | | G→A | 1,079 | 89.99 | |
| | 28,882nt | | G→A | 1,079 | 89.99 | |
| | 28,883nt | Nucleocapsid phosphoprotein | G→C | 1,075 | 89.66 | |
| | 28.905nt | | C→T | 62 | 5.17 | |
| | 28,975nt | | G→T | 24 | 2 | |
| | 29,518nt | ORF10 | $C \rightarrow T$ | 49 | 4.09 | |

| Table 9A - | France | | 04 - 09/12/2020 | |
|-----------------|---|---|-----------------------|--------------------|
| Position | Location | Mutation | Count. | Incidence. |
| (walich wa | rint doi: https://doi.org/10.1101/2 as not certified by peer review) | is the author/funder. who has granted medRxiv a license to | ldisplanvithe pr | eprinteinaperpetui |
| 2.1110 | It is made | available under a CC-BYONET OTY 0. International license | · · µ • • • • • • • • | 1 provident and |
| 1,059nt | | ACC (Threenine) $\rightarrow ATC$ (Isoleucine) | 16/116 | prevalent |
| 2.41 Cat | nsp2 | $CA \rightarrow TA$ | 25/110 | CN,ESP,RU, |
| 2,410nt | | noneffective | 25/116 | US,ZA |
| 3.037nt | | $CT \rightarrow TT$ | 115/116 | prevalent |
| | | | | 1 |
| 4,543nt | | $CA \rightarrow TA$ CAC (Histidine) $\rightarrow TAC$ (Tyrosine) | 15/116 | DE,ES |
| 5 (20) | nsp3 | $GT \rightarrow TT$ | 15/110 | DEEC |
| 5,629nt | - | noneffective | 15/116 | DE,ES |
| 0.051 | | $GG \rightarrow TG$ | 22/11/ | FOR |
| 8,371nt | | CAGGTA (Glutamine Valine) \rightarrow CATGTA (Histidine Valine) | 23/116 | ES,RU |
| | | $GT \rightarrow TT$ | | |
| 9,526nt | nsp4 | $ATGTCA$ (Methionine Serine) $\rightarrow ATTTCA$ (Isoleucine | 15/116 | DE,ES |
| | | Serine) | | |
| 11,497nt | nsp6 | $CT \rightarrow TT$ | 15/116 | DE,ES |
| | - | $\frac{\text{noneffective}}{\text{GC} \rightarrow \text{TC}}$ | - | |
| 13,993nt | | GCT (Alanine) \rightarrow TC T (Serine) | 15/116 | DE,ES |
| 14 408nt | | $CT \rightarrow TT$ | 114/116 | prevalent |
| 14,40811 | RNA-dependent RNA | $\mathbf{CCT} (\mathbf{Proline}) \to \mathbf{CTT} (\mathbf{Leucine})$ | 114/110 | prevalent |
| 15,324nt | polymerase | $CA \rightarrow TA$ | 22/116 | BR,CN,IN |
| | | GT → TT | | |
| 15,766nt | | GT G (Valine) \rightarrow TT G (Leucine) | 15/116 | DE,ES |
| 16 889nt | | $AA \rightarrow GA$ | 15/116 | DE ES |
| 10,000110 | halianna | $AAA (Lysine) \rightarrow AGA (Arginine)$ | 10,110 | 22,25 |
| 17.019nt | nencase | $GI \rightarrow II$ GAGTTT (Glutamine Acid Phenylalanine) \rightarrow GATTTT | 15/116 | DE ES |
| | | (Aspartic Acid Phenylalanine) | | ,_~~ |
| 20.268nt | endoRNAse | $AG \rightarrow GG$ | 13/116 | ES,RU,US, |
| 20,200 | | noneffective | 10,110 | ZA |
| 22,992nt | | AGC (Serine) $\rightarrow AAC$ (Asparagine) | 15/116 | DE,US |
| 22 402 4 | Spike glycoprotein | $AT \rightarrow GT$ | 116/116 | 1 / |
| 23,403nt | | GAT (Aspartic Acid) $\rightarrow GGT$ (Glycine) | 116/116 | prevalent |
| 25.562 4 | | $GA \rightarrow TA$ | 57/116 | 1 / |
| 25,563nt | ORE3a protein | $CAGAGC$ (Glutamine Serine) $\rightarrow CATAGC$ (Histidine Serine) | 57/116 | prevalent |
| 25.710 at | ord su protein | $CT \rightarrow TT$ | 16/116 | DEEC |
| 25,710nt | | noneffective | 10/110 | DE,ES |
| 26,735nt | | $CA \rightarrow TA$ | 15/116 | DE,ES,IN, |
| | Membrane glycoprotein | $TC \rightarrow CC$ | - | US |
| 26,876nt | | noneffective | 15/116 | DE,ES |
| 28 833nt | | $CA \rightarrow TA$ | 12/116 | FS |
| 20,05511 | | $TCA (Serine) \rightarrow TTA (Leucine)$ | 12/110 | LS |
| 28,851nt | | $GT \to TT$ $AGT (Serine) \to ATT (Isoleucine)$ | 10/116 | IN |
| | | $GGG \rightarrow AAC$ | | |
| 28,881nt | Nucleopaneid theorhometer | AGGGGA (Arginine Glycine) \rightarrow AAACGA (Lysine | 17/116 | prevalent |
| | ructeocapsia prosproprotein | Arginine) | | |
| 28 075 | | $GT \rightarrow CT$ | 15/110 | DEEGDI |
| 28,973nt | | AIGICI (Methionine Serine) \rightarrow AICICI (Isoleucine Serine) | 15/116 | DE,ES,IN |
| 20.200 | | $GC \rightarrow AC$ | 15/11/ | DEEG |
| 29,399nt | | GCT (Alanine) \rightarrow ACT (Threonine) | 15/116 | DE,ES |

| | Table 9B - Fi | rance | 01/19/2020 - 01/20/2021 | | | |
|-----------|-----------------------|---|--|--------------------------|-----------------|-------------------|
| | Position | Location | Mutation | Total Count | Percentage | |
| medRxiv p | epriateddi: htt | ps://doi.org/10.1101/2021.92,94,21251111; this version po | sted February 6 | 2021 ⁰⁰ he co | pyright7holder | for this preprint |
| (which | was2Abttcert | fied by peer review) is the author/funder, who has grante | d me@R x ii/ a lic | ense2to600isplay | the ppeprint in | n perpetuity. |
| | 313nt | $ORF1ab polyprotein \rightarrow leader protein$ | unte (navipnani | cense 55 | 2.07 | |
| | 445nt | | $I \rightarrow C$ | 163 | 6.14 | |
| | 1,059nt | nsp2 | $C \rightarrow I$ | 385 | 14.51 | |
| | 2,410nt | | $C \rightarrow I$ | 320 | 08.22 | |
| | 3,03/III 3,090nt | | $C \rightarrow T$ | 2,000 | 96.25 | |
| · | 4.543nt | | $C \rightarrow T$ | 666 | 2.0 | |
| | 4 960nt | | $G \rightarrow T$ | 69 | 25.1 | |
| | 4.965nt | | $C \rightarrow T$ | 69 | 2.6 | |
| | 5.170nt | | $C \rightarrow T$ | 53 | 2 | |
| | 5.629nt | nsp3 | G→T | 666 | 25.1 | |
| | 6.070nt | | C→T | 70 | 2.64 | |
| | 6,286nt | | $C \rightarrow T$ | 168 | 6.33 | |
| | 7,303nt | | $C \rightarrow T$ | 70 | 2.64 | |
| | 7,564nt | | $C \rightarrow T$ | 71 | 2.68 | |
| | 8,371nt | | $G \rightarrow T$ | 233 | 8.78 | |
| | 9,246nt | non 1 | $C \rightarrow T$ | 69 | 2.6 | |
| | 9,526nt | lisp4 | $G \rightarrow T$ | 667 | 25.14 | |
| | 10,279nt | | $C \rightarrow T$ | 70 | 2.64 | |
| | 10,301nt | | C→A | 69 | 2.6 | |
| | 10,525nt | 3C-like proteinase | $C \rightarrow T$ | 70 | 2.64 | |
| | 10,582nt | | $C \rightarrow T$ | 113 | 4.26 | |
| | 10,688nt | | $G \rightarrow T$ | 69 | 2.6 | |
| | 11,083nt | | G→T | 99 | 3.73 | |
| | 11,132nt | nsp6 | $G \rightarrow T$ | 54 | 2.04 | |
| | 11,497nt | a | $C \rightarrow T$ | 666 | 25.1 | |
| | 11,851nt | nsp7 | $G \rightarrow T$ | 96 | 3.62 | |
| | 13,993nt | | $G \rightarrow I$ | 664 | 25.03 | |
| | 14,230nt | | $C \rightarrow A$ | 68 | 2.56 | |
| | 14,408nt | RNA-dependent RNA polymerase | $C \rightarrow I$ | 2,000 | 98.23 | |
| | 15,524III 15,728nt | | $C \rightarrow I$ | 62 | 2 27 | |
| | 15,756nt | | $C \rightarrow T$ | 667 | 2.37 | |
| | 16.889nt | | $A \rightarrow G$ | 665 | 25.07 | |
| | 17.019nt | Helicase | $G \rightarrow T$ | 665 | 25.07 | |
| | 18.877nt | 3'-to-5' exonuclease | $C \rightarrow T$ | 675 | 25.44 | |
| | 20.268nt | endoRNAse | A→G | 111 | 4.18 | |
| | 21,255nt | 2'-O-ribose methyltransferase | G→C | 167 | 6.29 | |
| | 21,800nt | | G→T | 72 | 2.71 | |
| | 22,227nt | Cuille alternation | $C \rightarrow T$ | 172 | 6.48 | |
| | 22,992nt | Spike glycoprotein | G→A | 666 | 25.1 | |
| | 23,403nt | | A→G | 2,607 | 98.27 | |
| | 25,563nt | | $G \rightarrow T$ | 1,474 | 55.56 | |
| | 25,688nt | ORF3a | $C \rightarrow T$ | 56 | 2.11 | |
| | 25,710nt | | C→T | 677 | 25.52 | |
| | 26,735nt | | C→T | 670 | 25.25 | |
| | 26,801nt | Membrane glycoprotein | $C \rightarrow G$ | 167 | 6.29 | |
| | 26,876nt | ODEZ | $T \rightarrow C$ | 667 | 25.14 | |
| | 27,032nt | | $G \rightarrow I$ | 08 | 2.56 | |
| | 27,804nt | UKF/D | $C \rightarrow I$ | 83 95 | 3.2 | |
| | 20,03UII | | $C \rightarrow A$ | 62 | 3.2 2.24 | |
| | 20,033m | | $C \rightarrow 1$ $G \rightarrow \Lambda$ | 280 | 2.34 | |
| | 20,00111 28 882nt | | $G \rightarrow A$ | 230 | 10.33 | |
| | 28,883nt | Nucleocansid phosphoprotein | $G \rightarrow C$ | 276 | 10.44 | |
| | 28,932nt | rueresouperd prosproprotoin | $C \rightarrow T$ | 167 | 6.29 | |
| | 28.975nt | | G→C | 664 | 25.03 | |
| | 29,399nt | | G→A | 662 | 24.95 | |
| | 29,402nt | | G→T | 73 | 2.75 | |
| | 29,645nt | ORF10 | $G \rightarrow T$ | 169 | 6.37 | |
| | 29,779nt | 3′UTR | G→T | 67 | 2.53 | |

| Table 10A | - Spain | | | 06/(09/20/ |)1 - /2020 | | |
|-----------|---------|---|--|----------------|---------------|---|--|
| T | | - | | ~ | | - | |

 Position
 Location
 Mutation
 Count
 Incidence

 medRxiv preprint doi: https://doi.org/10.1101/2021.02.04.21251111; this version posted February 6, 2021. The copyright holder for this preprint (which was not gettified by peer review) is the author/funder, who has granted medRxiv graptse to display the preprint in perpetuity. It is made available under a CC. Porce (SCM) 4.0 International license.
 Incidence

| 445nt | ORF1ab polyprotein \rightarrow | $TT \rightarrow CT$ | 88/135 | CN,DE,FR |
|----------|----------------------------------|---|---------|-------------|
| 3 037nt | | $CT \rightarrow TT$ | 131/135 | prevalent |
| 3,037III | | noneffective | 131/133 | prevalent |
| 5,572nt | nsp3 | $GT \rightarrow TT$ ATGTAC (Methionine Tyrosine) \rightarrow ATTTAC (Isoleucine Tyrosine) | 11/135 | unique |
| 5,784nt | | $CT \rightarrow TT$ $ACT (Threonine) \rightarrow ATT$ (Isoleucine) | 13/135 | unique |
| 6,286nt | | $CT \rightarrow TT$ noneffective | 89/135 | DE,FR,ZA |
| 14,408nt | RNA-dependent RNA polymerase | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT$ (Leucine) | 132/135 | prevalent |
| 20,268nt | endoRNAse | $AG \rightarrow GG$ noneffective | 26/135 | FR,RU,US,ZA |
| 21,255nt | 2´-O-ribose methyltransferase | $GT \rightarrow CT$ noneffective | 84/135 | DE,FR |
| 22,227nt | · · · · | $CT \rightarrow TT$ noneffective | 89/135 | DE,FR,ZA |
| 22,297nt | | $TA \rightarrow CA$ noneffective | 11/135 | RU |
| 25,049nt | Spike glycoprotein | $GA \rightarrow TA$ GA T (Aspartic Acid) \rightarrow TA T (Tyrosine) | 18/135 | DE |
| 25,062nt | | $\begin{array}{c} \mathrm{GT} \to \mathrm{TT} \\ \mathrm{GGT} \ (\mathrm{Glycine}) \to \mathrm{GTT} \\ (\mathrm{Valine}) \end{array}$ | 18/135 | unique |
| 26,801nt | Membrane glycoprotein | $CA \rightarrow GA$ noneffective | 89/135 | DE,FR,ZA |
| 27,944nt | | $CC \rightarrow TC$ noneffective | 56/135 | FR |
| 27,982nt | ORF8 protein | $CA \rightarrow TA$ $CCA (Proline) \rightarrow CTA$ (Leucine) | 13/135 | unique |
| 28,657nt | | $CG \rightarrow TG$ noneffective | 19/135 | unique |
| 28,881nt | Nucleocapsid phosphoprotein | $GGG \rightarrow AAC$ $AGGGGGA (Arginine Glycine)$ $\rightarrow AAACGA (Lysine Arginine)$ | 14/135 | prevalent |
| 28,932nt | | $CT \rightarrow TT$ $GCT (Alanine) \rightarrow GTT$ (Valine) | 89/135 | unique |
| 29,645nt | ORF10 protein | $GT \rightarrow TT$ noneffective | 89/135 | DE,FR |

| | Table 10B - S | Spain | | 01/19/2020 | | |
|-----------|-----------------|---|----------------------------|----------------------|----------------|-------------------|
| | Position | Location | Mutation | Total Count | Percentage | |
| nedRxiv p | repriated to it | ps://doi.org/10.1101/2021.52.07.21251111: this version pc | sted Februarv 6 | 2024,690 e co | pvria%+holder | for this preprint |
| (which | was3nontcert | fied by peer review) is the author/funder, who has grante | d me ɗR ≭iî⁄ a Íice | ense tbl@isplay | the preprint i | n perpetuity. |
| | 445nt | It is made available under a CC-BY-NC-ND 4. | 0 International li | cense ₈₅₈ | 25.03 | |
| | 1,059nt | nen? | $C \rightarrow T$ | 122 | 3.56 | |
| | 1,987nt | nsp2 | A→G | 75 | 2.19 | |
| | 3,037nt | | C→T | 2,717 | 79.26 | |
| | 5,170nt | nom2 | C→T | 141 | 4.11 | |
| | 6,286nt | lisp5 | C→T | 861 | 25.12 | |
| | 6,294nt | | T→C | 82 | 2.39 | |
| | 8,782nt | non4 | $C \rightarrow T$ | 601 | 17.53 | |
| - | 9,477nt | nsp4 | T→A | 379 | 11.06 | |
| | 11,083nt | nanG | $G \rightarrow T$ | 166 | 4.84 | |
| | 11,132nt | nspo | $G \rightarrow T$ | 137 | 4 | |
| | 13,006nt | nsp9 | T→C | 77 | 2.25 | |
| | 14,408nt | DNIA danan dané DNIA malansanan | $C \rightarrow T$ | 2,708 | 79 | |
| | 14,805nt | KINA-dependent KINA porymerase | $C \rightarrow T$ | 408 | 11.9 | |
| | 20,268nt | endoRNAse | A→G | 1,223 | 35.68 | |
| | 21,255nt | 2'-O-ribose methyltransferase | G→C | 780 | 22.75 | |
| | 22,227nt | | $C \rightarrow T$ | 843 | 24.59 | |
| | 23,403nt | Spike glycoprotein | A→G | 2,731 | 79.67 | |
| | 25,049nt | | $G \rightarrow T$ | 71 | 2.07 | |
| | 25,563nt | | G→T | 147 | 4.29 | |
| | 25,688nt | | C→T | 78 | 2.28 | |
| | 25,979nt | ORF3a | G→T | 371 | 10.82 | |
| | 26,088nt | | C→T | 215 | 6.27 | |
| | 26,144nt | | G→T | 100 | 2.92 | |
| | 26,801nt | Membrane glycoprotein | C→G | 855 | 24.94 | |
| | 27,944nt | ODE0 | C→T | 456 | 13.3 | |
| | 28,144nt | ORF8 | T→C | 599 | 17.47 | |
| | 28,657nt | | C→T | 441 | 12.86 | |
| | 28,863nt | | C→T | 378 | 11.03 | |
| | 28,881nt | | G→A | 398 | 11.61 | |
| | 28,882nt | Nucleocapsid phosphoprotein | G→A | 396 | 11.55 | |
| | 28,883nt | | G→C | 395 | 11.52 | |
| | 28,932nt | | $C \rightarrow T$ | 850 | 24.8 | |
| | 29,645nt | ORF10 | $G \rightarrow T$ | 840 | 24.5 | |
| | 29,734nt | | G→C | 302 | 8.81 | |
| | 29,870nt | 3 UTR | C→A | 107 | 3.12 | |

| Table 11A - | Cable 11A - Germany | | | DE-II 02 - 06/17/2020 | DE-III 06/24 - 08/28/2020 | DE-IV 09/10 - 10/13/2020 | |
|-------------|-------------------------------|--|-------|--------------------------|---------------------------------|-----------------------------|--------------------|
| Position | Location | Mutation | Count | Count | Count | Count | Incidence |
| 241nt | 5'UTR | $CG \rightarrow TG$ noneffective | 4/62 | 112/138 | 17/17 | 70/70 | prevalent |
| 445nt | nsp1 | $TT \rightarrow CT$ $TTG (Leucine) \rightarrow GTCTTG (Valine Leucine)$ | 0/62 | 0/138 | 1/17 | 17/70 | CN,FR |
| 1,059nt | | $CC \rightarrow TC$ $ACC \text{ (Threonine)} \rightarrow ATC \text{ (Isoleucine)}$ | 21/62 | 27/138 | 0/17 | 2/70 | prevalent |
| 1,440nt | nsn? | $GC \rightarrow AC$ $GGC (Glycine) \rightarrow GAC (Aspartic Acid)$ | 15/62 | 18/138 | 0/17 | 0/70 | US |
| 1,513nt | 11512 | $CC \rightarrow TC$ noneffective | 0/62 | 0/138 | 0/17 | 13/70 | unique |
| 2,891nt | | $GC \to AC$ $GCA (Alanine) \to ACA (Threonine)$ | 15/62 | 18/138 | 0/17 | 0/70 | US |
| 3,037nt | | $CT \rightarrow TT$ noneffective | 41/62 | 114/138 | 17/17 | 70/70 | prevalent |
| 3,602nt | | $CA \rightarrow TA$ CAC (Histidine) $\rightarrow TAC$ (Tyrosine) | 0/62 | 0/138 | 5/17 | 6/70 | unique |
| 4,543nt | nsp3 | $CA \rightarrow TA$ noneffective | 0/62 | 0/138 | 5/17 | 2/70 | ES,FR,US |
| 6,286nt | | $CT \rightarrow TT$ noneffective | 0/62 | 0/138 | 1/17 | 17/70 | ES,FR,ZA |
| 6,941nt | | $CT \rightarrow TT$ noneffective | 0/62 | 0/138 | 5/17 | 6/70 | unique |
| 14,408nt | | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT (Leucine)$ | 39/62 | 114/138 | 17/17 | 70/70 | prevalent |
| 15,324nt | RNA-dependent RNA polymerase | $CA \rightarrow TA$ noneffective | 1/62 | 1/138 | 5/17 | 6/70 | BR,CN,FR, IN |
| 16,075nt | | $GA \rightarrow TA$ GAT (Aspartic Acid) $\rightarrow TAT$ (Tyrosine) | 0/62 | 0/138 | 0/17 | 11/70 | FR |
| 19,839nt | endoRNAse | $TA \rightarrow CA$ noneffective | 0/62 | 0/138 | 2/17 | 11/70 | CN,ES,FR, IN,US |
| 21,255nt | 2'-O-ribose methyltransferase | $GT \rightarrow CT$ noneffective | 0/62 | 0/138 | 1/17 | 17/70 | ES,FR |
| 21,855nt | | $CT \rightarrow TT$ $TCT (Serine) \rightarrow TTT (Phenylalanine)$ | 0/62 | 0/138 | 5/17 | 6/70 | ZA |
| 22,227nt | | $CT \rightarrow TT$ noneffective | 0/62 | 0/138 | 1/17 | 18/70 | ES,FR,ZA |
| 22,346nt | Spike glycoprotein | $GC \rightarrow TC$ GC T (Alanine) \rightarrow TC T (Serine) | 0/62 | 0/138 | 0/17 | 13/70 | unique |
| 22,377nt | | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT (Leucine)$ | 0/62 | 0/138 | 0/17 | 13/70 | unique |
| 23,403nt | | $AT \rightarrow GT$ GAT (Aspartic Acid) \rightarrow GGT (Glycine) | 1/62 | 112/138 | 17/17 | 70/70 | prevalent |
| 25,505nt | | $AA \rightarrow GA$ CAA (Glutamine) $\rightarrow CGA$ (Arginine) | 0/62 | 0/138 | 5/17 | 6/70 | unique |
| 25,563nt | ORF3a protein | $GA \rightarrow TA$ CAGAGC (Glutamine Serine) $\rightarrow CATAGC$ (Histidine Serine) | 21/62 | 27/138 | 2/17 | 5/70 | prevalent |
| 25,906nt | | $GG \to CG$ $GGT (Glycine) \to CGT (Arginine)$ | 0/62 | 0/138 | 5/17 | 6/70 | unique |
| 26,801nt | Membrane glycoprotein | $CA \rightarrow GA$ noneffective | 1/62 | 0/138 | 1/17 | 17/70 | ES,FR,ZA |
| 27,046nt | Weinbrane grycoprotein | $CG \rightarrow TG$ ACG (Threonine) $\rightarrow ATG$ (Methionine) | 1/62 | 16/138 | 3/17 | 0/70 | BR,RU |
| 28,651nt | | $CA \rightarrow TA$ noneffective | 0/62 | 0/138 | 5/17 | 6/70 | FR,RU |
| 28,706nt | | $CA \rightarrow TA$ CAC (Histidine) $\rightarrow TAC$ (Tyrosine) | 0/62 | 0/138 | 0/17 | 11/70 | unique |
| 28,869nt | Nucleocapsid phosphoprotein | $CA \rightarrow TA$ $CCA (Proline) \rightarrow CTA (Leucine)$ | 0/62 | 0/138 | 5/17 | 6/70 | unique |
| 28,881nt | | $GGG \rightarrow AAC$ $AGGGGA (Arginine Glycine) \rightarrow AAACGA (Lysine Arginine)$ | 9/62 | 35/138 | 9/17 | 38/70 | prevalent |
| 28,932nt | | $\frac{\text{CT} \rightarrow \text{TT}}{\text{GCT} \text{ (Alanine)} \rightarrow \text{GTT} \text{ (Valine)}}$ | 0/62 | 0/138 | 1/17 | 17/70 | FR |
| 29,645nt | ORF10 protein | $GT \rightarrow TT$ noneffective | 0/62 | 0/138 | 1/17 | 17/70 | ES,FR |
| 29,751nt | 3'UTR | $GA \rightarrow CA$ noneffective | 0/62 | 0/138 | 0/17 | 11/70 | unique |

| Table 11B - | Germany | | 01/19/2020 | | |
|--|---|--|--------------------------|---------------------------------|---------------|
| Position | Location | Mutation | Total Count | Percentage | |
| preprinte 2001: htt | ps://doi.org/10.1101/2021.02.04.21251111; this version po | sted Petruary 6 | 2021 ⁴ The co | pyright ¹ folder | for this prep |
| h was2004tcert | fied by peer review) is the all the review has granted | d me dR ≭il∕ a lic | ense tel&lisplay | the preprint i | n perpetuity. |
| 241nt | It is made available under a CC-BY-NC-ND 4. | 0 International li | <mark>cens∉</mark> ,790 | 86.64 | |
| 313nt | OPF1ab polymentain 1 loader protain | $C \rightarrow T$ | 53 | 2.57 | |
| 445nt | OKI Tab polyprotein → leader protein | T→C | 159 | 7.7 | |
| 1,059nt | | $C \rightarrow T$ | 399 | 19.31 | |
| 1,440nt | nsp2 | G→A | 76 | 3.68 | |
| 2,891nt | | G→A | 76 | 3.68 | |
| 3,037nt | | C→T | 1,796 | 86.93 | |
| 3,373nt | | C→A | 53 | 2.57 | |
| 3,602nt | | C→T | 77 | 3.73 | |
| 4.543nt | nsp3 | C→T | 42 | 2.03 | |
| 6.286nt | | C→T | 155 | 7.5 | |
| 6.406nt | | $C \rightarrow T$ | 57 | 2.76 | |
| 6.941nt | 1 | $C \rightarrow T$ | 79 | 3.82 | |
| 8 782nt | nsp/ | $C \rightarrow T$ | 128 | 62 | |
| 11 083nt | nspf | $C \rightarrow T$ | 01 | 0.2 | |
| 14.408nt | пъро | | 1 782 | 4.4 86.25 | |
| 14,408ht | PNA dependent PNA polymerase | $C \rightarrow T$ | 1,782 | 2 37 | |
| 15 224nt | KivA-dependent KivA porymetase | $C \rightarrow T$ | 129 | 6.69 | |
| 19,524III | | $C \rightarrow T$ | 130 | 0.08 | |
| 18,87/III 18,072mt | 3'-to-5' exonuclease | $C \rightarrow 1$ | 59 | 2.00 | |
| 18,972nt | | $G \rightarrow A$ | 58 | 2.81 | |
| 19,839nt | endoRNAse | $1 \rightarrow C$ | 52 | 2.52 | |
| 20,268nt | | A→G | /8 | 3.78 | |
| 21,255nt | 2-O-ribose methyltransferase | $G \rightarrow C$ | 162 | 7.84 | |
| 21,614nt | 4 | $C \rightarrow T$ | 45 | 2.18 | |
| 21,855nt | | $C \rightarrow T$ | 76 | 3.68 | |
| 22,227nt | Spike glycoprotein | C→T | 166 | 8.03 | |
| 22,468nt | | G→T | 116 | 5.61 | |
| 23,403nt | | A→G | 1,800 | 87.12 | |
| 25,505nt | | A→G | 74 | 3.58 | |
| 25,550nt | | T→A | 53 | 2.57 | |
| 25,563nt | | $G \rightarrow T$ | 492 | 23.81 | |
| 25,906nt | ORF3a | $G \rightarrow C$ | 74 | 3.58 | |
| 25,922nt | | $G \rightarrow T$ | 50 | 2.42 | |
| 25,996nt | | $G \rightarrow T$ | 75 | 3.63 | |
| 26,144nt | | $G \rightarrow T$ | 44 | 2.13 | |
| 26,530nt | | A→G | 55 | 2.66 | |
| 26,735nt | Mambara aluanantain | $C \rightarrow T$ | 43 | 2.08 | |
| 26,801nt | Memorane grycoprotein | C→G | 145 | 7.02 | |
| 27,046nt | | C→T | 68 | 3.29 | |
| 27,944nt | 0.0.00 | C→T | 89 | 4.31 | |
| 28.144nt | ORF8 | T→C | 131 | 6.34 | |
| 28.651nt | | C→T | 74 | 3.58 | |
| 28.854nt | 1 | C→T | 59 | 2.86 | |
| 28,869nt | 1 | C→T | 75 | 3.63 | |
| | 1 | $G \rightarrow A$ | 124 | 6 | |
| 28.878nt | Nucleocapsid phosphoprotein | C A | 589 | 28 51 | |
| 28,878nt 28,881nt | r (ueroceupoid priosprioproteini | $(\rightarrow A)$ | // . / | 2.0 | |
| 28,878nt 28,881nt 28,882nt | | $G \rightarrow A$ | 585 | 28.31 | |
| 28,878nt 28,881nt 28,882nt 28,882nt | | $G \rightarrow A$ $G \rightarrow A$ $G \rightarrow C$ | 585 | 28.32 | |
| 28,878nt 28,881nt 28,882nt 28,883nt 28,932nt | | $ \begin{array}{c} G \rightarrow A \\ \hline G \rightarrow A \\ \hline G \rightarrow C \\ \hline C \rightarrow T \end{array} $ | 585 585 162 | 28.31 28.32 28.32 7.84 | |

| Table 12A | - China | | CN-I 12/23/2019 - 03/18/ 2020* | CN-II 03/20 - 07/22/2020 | | |
|-----------|------------------------------|--|--------------------------------------|---|--------------------------------|-------|
| Position | Location by poor | review) is the author/fullation has granted medBy | iv a license to d | ne copyright hold count splay the proprie | ier for this prep Incidence | orint |
| 241nt | 5'UTR I | is made available under a CCBY-NC-ND 4.0 Internat noneffective | ional lipegse . | 23/33 | prevalent | • |
| 3,037nt | nsp3 | $CT \rightarrow TT$ noneffective | 2/99 | 23/33 | prevalent | |
| 8,782nt | nsp4 | $CC \rightarrow TC$ noneffective | 29/99 | 0/33 | DE,ES,IN, US | |
| 14,408nt | RNA-dependent RNA polymerase | $CT \rightarrow TT$ $CCT (Proline) \rightarrow CTT (Leucine)$ | 2/99 | 19/33 | prevalent | |
| 23,403nt | Spike glycoprotein | $AT \rightarrow GT$ GAT (Aspartic Acid) \rightarrow GGT (Glycine) | 2/99 | 22/33 | prevalent | |
| 28,144nt | ORF8 protein | $TA \rightarrow CA$ TTA (Leucine) \rightarrow TCA (Serine) | 29/99 | 0/33 | DE,ES,IN, US | |
| 28,881nt | Nucleocapsid phosphoprotein | GGG → AAC AGGGGA (Arginine Glycine) → AAACGA (Lysine Arginine) | 2/99 | 11/33 | prevalent | |

| | Table 12B - | China | 01/19/2020 - | | | |
|-----------|-------------------------------|--|----------------------------|---------------------------|------------------------------|-------------------|
| | Position | Location | Mutation | Total Count | Percentage | |
| medRxiv p | reprint ⁴ ଅoi: htt | ps://doi.org/10.1101/2021.9 <mark>2.64</mark> ,21251111; this version po | sted February 6 | 2021 ¹⁵ The co | pyright ⁴ flolder | for this preprint |
| (which | was2Aottcert | fied by peer review) is the author/funder, who has grante | d me @R xílv a lice | ense to&display | the preprint i | n perpetuity. |
| | 1,397nt | It is made available under a CC-BY-NC-ND 4. |) Inte matio nal li | cense 18 | 2.99 | |
| | 2,392nt | nsp2 | T→C | | | |
| | 3,037nt | | $C \rightarrow T$ | 65 | 10.78 | |
| | 6,354nt | nen? | $C \rightarrow T$ | 14 | 2.32 | |
| | 7,075nt | nsp5 | T→C | 14 | 2.32 | |
| | 8,022nt | | T→G | 15 | 2.49 | |
| | 8,782nt | nsp4 | $C \rightarrow T$ | 191 | 31.67 | |
| | 10,747nt | 3C-like proteinase | $C \rightarrow T$ | 14 | 2.32 | |
| | 11,083nt | | $G \rightarrow T$ | 40 | 6.63 | |
| | 11,794nt | nspo | A→G | 14 | 2.32 | |
| | 14,408nt | | $C \rightarrow T$ | 55 | 9.12 | |
| | 15,324nt | | $C \rightarrow T$ | 13 | 2.16 | |
| | 15,342nt | RNA-dependent RNA polymerase | $C \rightarrow T$ | 14 | 2.32 | |
| | 15,360nt | | $C \rightarrow T$ | 14 | 2.32 | |
| | 15,666nt | | G→A | 14 | 2.32 | |
| | 16,733nt | | $C \rightarrow T$ | 14 | 2.32 | |
| | 17,373nt | Helicase | $C \rightarrow T$ | 26 | 4.31 | |
| | 18,060nt | 3'-to-5' exonuclease | $C \rightarrow T$ | 16 | 2.65 | |
| | 21,707nt | | C→T | 24 | 3.98 | |
| | 21,727nt | | C→T | 14 | 2.32 | |
| | 22,020nt | spike glycoprotein | T→C | 16 | 2.65 | |
| | 23,403nt | | A→G | 67 | 11.11 | |
| | 25,416nt | 0772 | C→T | 14 | 2.32 | |
| | 26,144nt | ORF3a | G→T | 39 | 6.47 | |
| | 27.213nt | ORF6 | C→T | 15 | 2.49 | |
| | 28,144nt | ORF8 | T→C | 212 | 35.16 | |
| | 28.688nt | | T→C | 13 | 2.16 | |
| - | 28.854nt | | C→T | 14 | 2.32 | |
| | 28.881nt | | G→A | 33 | 5.47 | |
| | 28.882nt | Nucleocapsid phosphoprotein | G→A | 31 | 5.14 | |
| | 28,883nt | | $G \rightarrow C$ | 31 | 5.14 | |
| | 29.095nt | | $C \rightarrow T$ | 31 | 5.14 | |
| | 29.742nt | | G→T | 19 | 3.15 | |
| | 29,835nt | 3°UTR | $C \rightarrow T$ | 14 | 2.32 | |

| Country | Total number mutations | Novel Unique mutations | Novel Shared mutations | Sum novel mutations | Prevalent mutations | C to T transitions [in % of mutants] | RNA replication | Spike glycoprotein | Nucleocapsid phosphoprotein | Covid-19 cases | Covid-19 deaths |
|----------------------------------|------------------------------|--|---------------------------------------|--|---|---|--|---|---|---------------------------------|-----------------------|
| United Kingdom | medRxiv pre (\utagebich v | print doi: http ∕ as n⊉0 certi | s://doi.org/10 fied bys peer It | 0.1101/2021. review) is th is made ava | 02.04.21251 e author/fun ilable under | 111; this vers der, who@has a CC-BY-NC- | ion posted Fe grantes medF ND 4.0 Intern | bruary 6, 2021 xiv a license to ational license | The copyright hol display the prepri | derforthisp nt3n,7pe,≰peetui | eprint ty. (2.77%) |
| South Africa | 28 | 9 | 12 | 21 (75%) | 7 | 46.4 | 4 | 7 | 3 | 1,443,939 | 43,633 (3.02%) |
| USA | 39 | 17 | 13 | 30 (76.9%) | 7 | 59 | 13 | 3 | 7 | 26,514,275 | 447,490 (1.69%) |
| India | 23 | 9 | 9 | 18 (78.3%) | 5 | 69.6 | 6 | 4 | 2 | 10,740,309 | 154,202 (1.44%) |
| Brazil | 9 | 2 | 2 | 4 (44.4%) | 5 | 44.4 | 1 | 2 | 2 | 9,119,477 | 222,775 (2.44%) |
| Russia | 10 | 3 | 1 | 4 (40%) | 6 | 50 | 2 | 1 | 1 | 3,832,080 | 72,697 (1.90%) |
| France | 27 | 0 | 20 | 20 (74.1%) | 7 | 40.7 | 7 | 2 | 5 | 3,153,487 | 75,620 (2.40%) |
| Spain | 20 | 6 | 10 | 16 (80%) | 4 | 50 | 3 | 4 | 3 | 2,830,478 | 58,319 (2.06%) |
| Germany | 33 | 11 | 15 | 26 (78.8%) | 7 | 51.5 | 5 | 5 | 5 | 2,209,057 | 57,150 (2.59%) |
| People's Republic of China | 7 | 0 | 2 | 2 (28.6%) | 5 | 57.1 | 1 | 1 | 1 | 89,430 | 4,636 (5.18%) |

| Table 13 – A Survey | Covid-19 numbers – January 30, 2021 |
|---------------------|-------------------------------------|
|---------------------|-------------------------------------|