 doi:10.20944/preprints202101.0173.v1

Molecular epidemiology surveillance of SARS-CoV-2: mutations and genetic diversity one year after emerging.

Alejandro Flores-Alanis[1], Armando Cruz-Rangel[2], Flor Rodríguez-Gómez[3], James González[4], Carlos Alberto Torres-Guerrero[5], Gabriela Delgado[1], Alejandro Cravioto[1] and Rosario Morales-Espinosa[1]*.

[1]Departamento de Microbiología y Parasitología, Facultad de Medicina, Universidad Nacional Autónoma de México, Mexico City 04360, Mexico. bioalejandrofa@gmail.com; delgados@unam.mx; dracravioto@hotmail.com

[2]Laboratorio de Bioquímica de Enfermedades Crónicas, Instituto Nacional de Medicina Genómica, Mexico City 14610, Mexico. acruz@inmegen.gob.mx

[3]Departamento de Ciencias Computacionales, Centro Universitario de Ciencias Exactas e Ingenierías, Universidad de Guadalajara. Guadalajara 44430, Jalisco, Mexico. fiores.flor@gmail.com

[4]Departamento de Biología Celular, Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico. james@ciencias.unam.mx

[5]Independet consultant. cartogue86@gmail.com

*Correspondance: marosari@unam.mx

 doi:10.20944/preprints202101.0173.v1

Abstract

In December 2019, the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in the province of Wuhan, China. Since then, it has spread worldwide with new mutations being reported. We performed genomic analysis to identify the changes in genetic diversity of SARS-CoV-2 between December 2019 and November 2020, and through molecular surveillance, we monitored the mutations that could be involved in viral fitness. We analyzed 2,213 complete genomes from 6 geographical regions worldwide, which were downloaded from GenBank and GISAID databases. Although SARS-CoV-2 presented low genetic diversity, there has been an increase over time, with the presence of several hotspot mutations throughout its genome. We identified 7 frequent mutations that resulted in non-synonymous substitutions (*dN*). Two of them, C14408T>P323L and A23403G>D614G, located in the nsp12 and Spike protein, respectively, emerged early in the pandemic and showed a considerable increase in frequency over time. Two other mutations, A1163T>I120F in nsp2 and G22992A>S477N in the Spike protein emerged recently and have spread in Oceania and Europe. Continuous molecular surveillance of SARS-CoV-2 will be necessary to detect and describe the transmission dynamics of new variants of the virus with clinical relevance. This information is important to improve programs to control the virus.

Keywords: SARS-CoV-2, genetic diversity, molecular surveillance, natural selection, non-synonymous substitution.

 doi:10.20944/preprints202101.0173.v1

1. Introduction

Following reports of a new infectious disease in the province of Wuhan, China, in December 2019, the subsequent global pandemic has led to 82,579,768 confirmed cases and 1,818,849 deaths up to January 2, 2021[1]. The infectious agent responsible for this pandemic was identified as a virus of the Coronavirus family (CoVs), which was named as severe acute respiratory syndrome CoV 2 (SARS-CoV-2) [2], while the disease caused by this virus was called COVID-19 (*Coronavirus disease 2019*). SARS-CoV-2 is a positive, simple-strand RNA virus with a genome of approximately 29 Kb in size that is organized into 11 open reading frames (ORFs) [3]. The first ORF represents approximately 85% of the viral genome, which is composed of two overlapping ORFs (ORF1a and ORF1b). These ORFs encode two polypeptides that are processed into 16 non-structural proteins (nsp1-16). The main non-structural proteins include RNA-dependent RNA polymerase (RdRp or nsp12) and a 3'>5' exonuclease (ExonN or nsp14) [4]. The remaining ORFs encode the following 4 structural proteins: the Spike surface glycoprotein (S); an envelope protein (E); a membrane protein (M); and the nucleocapsid protein (N); as well as other accessory proteins (ORF3a, ORF6, ORF7a/b, ORF8 and ORF10) [5,6].

An important factor in the evolution of RNA viruses is their high mutation rate (10^{-6} to 10^{-4} substitutions/nucleotide/cell infection) [7]. This phenomenon can be explained partially because the RNA polymerase cannot correct mistakes during genome replication [8]. However, CoVs possess an ExonN with the capacity to correct mistakes that occur during replication [9]. This feature has contributed to the low mutation rate of CoVs compared to others RNA viruses [10–12].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

When a virus is well adapted to its environment, the establishment of new mutations in the virus population is not favored because most mutations become deleterious (purifying selection). In general, mutations that increase in frequency could be advantageous and fixed by positive selection, or they could be neutral and fixed by genetic drift. It is important to note that when neutral mutations increase in frequency, they can be confused with positive natural selection [8,13]. Therefore, in studies involving the evolutionary dynamics of a new pathogenic virus, like SARS-CoV-2, it is important to know if the increase in the frequency of mutations is due to natural selection, in order to determine the possible consequences for its fitness, such as increased infectiousness and pathogenicity, or due to adaptation thereby becoming drug resistant or having the ability to evade the immune system.

The aim of the present study was to use molecular epidemiology to track non-synonymous substitutions (*dN*) that could be implicated in the fitness of SARS-CoV-2 and its spread in different regions between December 2019 and November 2020. The information generated will be useful to understand the evolutionary dynamics of SARS-CoV-2 better in order to improve intervention measures against it.

2. Results

2.1. Global genetic diversity of SARS-CoV-2

Comparison among the 2,213 SARS-CoV-2 genomes showed high nucleotide identity (99.9-100%), with an average pairwise difference of 12.78 nucleotides between any two genomes. The global nucleotide diversity (π) of the 2,213 whole genomes was low (π=0.00044±0.00001). This diversity was not evenly distributed throughout the virus genome, with several high diversity peaks or hotspot mutations in ORF1ab, S gene and N

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

gene being detected. N gene showed the highest peak of nucleotide diversity (π=0.02934) (Figure 1).

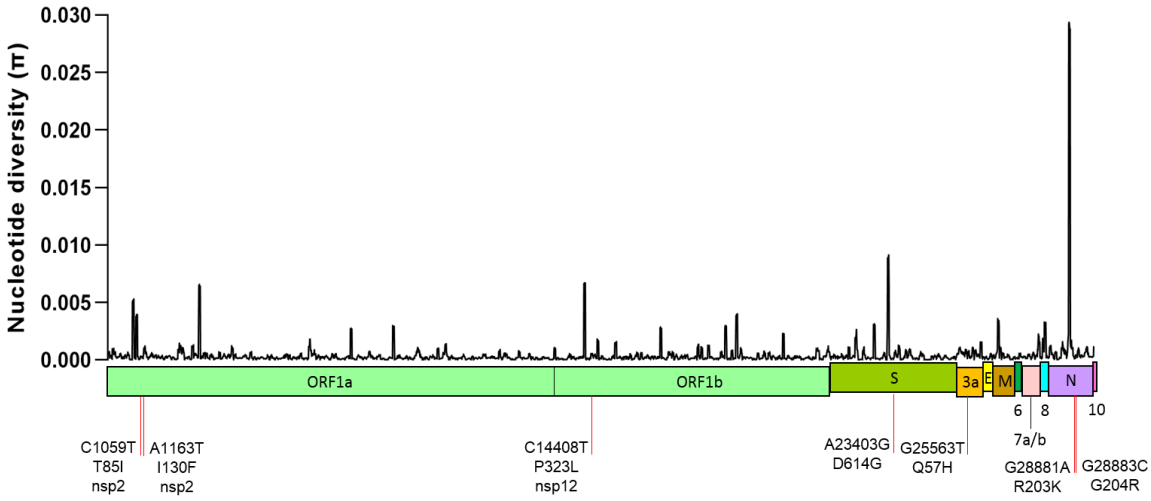


Figure 1. Nucleotide diversity (π) in a total of 2,213 SARS-CoV-2 genomes. Several hotspot mutations were detected along the genome. Seven nucleotide substitutions with frequencies >10% in the sample population are indicated, all of which resulted in amino acid non-synonymous (*dN*) substitutions. The π values were calculated within a sliding window of 50 bp moving with 10 bp steps.

2.2. Spatial-temporal genetic diversity of SARS-CoV-2

Over time, an increase in the global π values was observed, which coincided with the increase in COVID-19 cases from December 2019 to October 2020 (Figure 2). There was a slight decrease in π values in November but we only sampled until 13th November so a decrease for the month as a whole was expected. Regional analysis around the world showed that π values were low and similar to each other (United States of America (US) π=0.00044±0.00001, Latin America (LA) π=0.00037±0.00002, Europe (EU) π=0.00043±0.00002, Africa (AF) π=0.00047±0.00002, Asia (AS) π=0.00042±0.00001 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

Oceania (OC) π=0.00046±0.00001), although AF and OC regions showed the highest diversities.

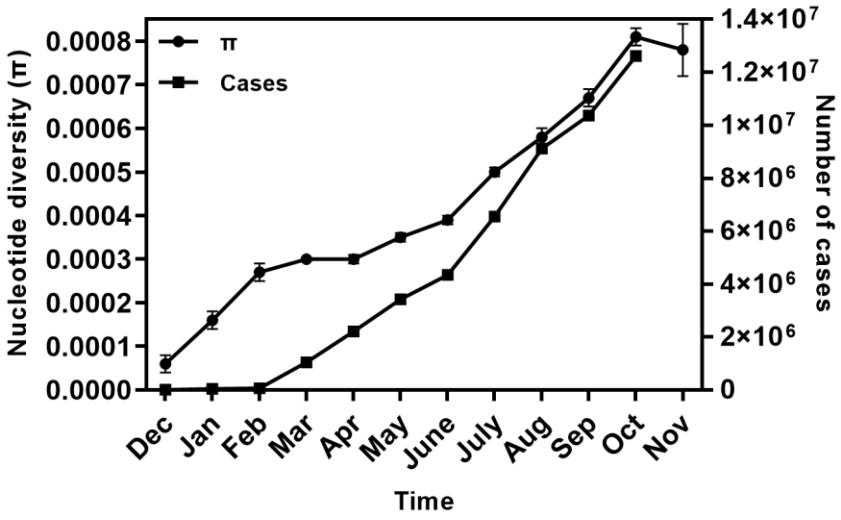


Figure 2. Temporal changes of SARS-Cov-2 nucleotide diversity (π) and monthly incidence of COVID-19 cases according to confirmed global cases from December 2019 to November 2020. The number of cases in November was not considered because we recorded until 13th November 2020 only. The number of isolates analyzed per month was as follows: Dec=15, Jan=103, Feb=84, Mar=628, Apr=222, May=221, June=118, July=233, Aug=196 and Sep=179, Oct=171, Nov=43. Dec, December; Jan, January, Feb, February, Mar, March; Apr, April; Aug, August; Sep, September; Oct, October; Nov, November.

Fluctuations in the π values with a tendency to increase over time were observed in US (January π=0.00025±0.00007 - October π=0.00071±0.00003), EU (January π=0.00008±0.00002 - November π=0.00072±0.00004), AF (February π=0.00038±0.00019 - November π=0.00100±0.00004) and AS (December π=0.00006±0.00002 - October π=0.00062±0.00005). In LA, there was an increase in the π values from March to August

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

(π=0.00028±0.00002 - π=0.00046±0.00004) but in September, a drastic decrease in the π value (π=0.00025±0.00012) was detected. While OC showed low diversity between February and September (π=0.00014±0.00006 - π=0.00020±0.00002), the diversity increased dramatically (π=0.00074±0.00003 and π=0.00080±0.00022, respectively) during the months of October and November (Figure 3).

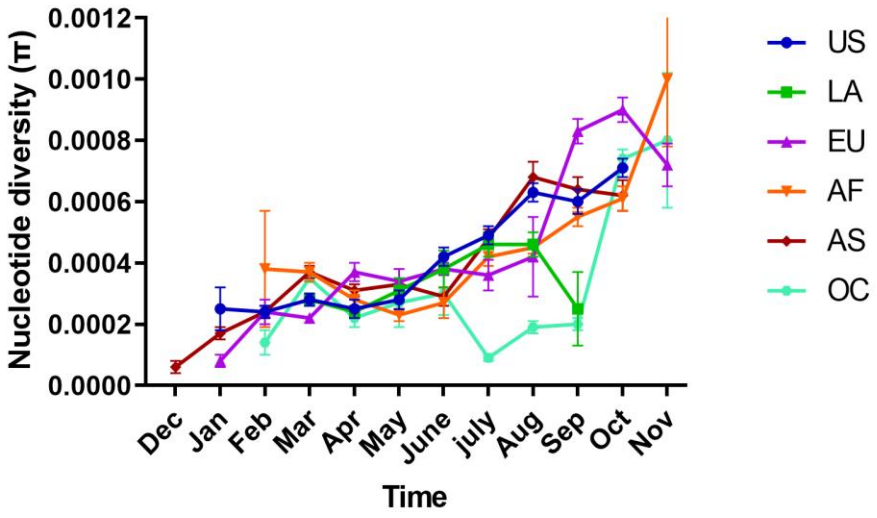


Figure 3. Temporal changes of SARS-Cov-2 nucleotide diversity (π) by region. Dec, December; Jan, January, Feb, February, Mar, March; Apr, April; Aug, August; Sep, September; Oct, October; Nov, November. US, United States of America; LA, Latin America; EU, Europe; AF, Africa; AS, Asia; OC, Oceania.

2.3. Non-synonymous substitutions and natural selection

Among the 2,213 whole genomes analyzed, we found 3,178 polymorphic sites (*S*), of which a high proportion (58.5%, 1861 sites) were non-synonymous (*dN*) when compared with the reference strain, Wuhan-Hu-1. Although there were a large number of *dN* substitutions, the majority were neutral (*dN/dS* values were between -22.85 and 7.96 but

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

not statistically significant). In general, it appears that the global population of SARS-CoV-2 is under purifying selection (*dN/dS*=-3.533; p<0.01).

When we analyzed *dN* substitutions in total, we identified 7 with frequencies >10% in the global population of SARS-CoV-2 (Table 1). These 7 frequencies varied by region: T85I and Q57H (nsp2) were the most frequent in US; I120F (nsp2) in OC, and R203K and G204R (N protein) in LA, AF and OC; while P323L (nsp12) and D614G (S protein) were highly frequent in all regions. Positive selection was seen in T85S (*dN/dS*=5.89; p<0.01) and P323L (*dN/dS*=7.49; p<0.01), while I120F, D614G, Q57H and G204R had positive values of *dN/dS* but there were not significant. Meanwhile, R203K presented a *dN/dS* negative value, but again, this was not significant (Table 1).

Table 1. Non-synonymous substitutions (*dN*) of medium-high frequency in the global population of SARS-CoV-2.

Nucleotide change	Amino acid change	Genomic location	*dN/dS* (p value)	Distribution and frequency (%)						
				US	LA	EU	AF	AS	OC	Global
C1059T	T85I	ORF1a (nsp2)	5.89 (0.009)	49.2	12.5	9.80	5.80	2.40	11.4	14.41
A1163T	I120F	ORF1a (nsp2)	4.79 (0.052)	0.00	0.00	0.20	0.00	11.4	43.2	10.08
C14408T	P323L	ORF1b (nsp12)	7.49 (0.002)	80.6	92.3	81.8	88.4	68.2	81.1	79.58
A23403G	D614G	S gene	2.42 (0.153)	80.3	90.9	85.3	92.6	69.0	81.1	80.80
G25563T	Q57H	ORF3	7.13 (0.105)	59.8	18.7	21.1	18.6	25.9	16.2	27.47
G28881A	R203K	N gene	-0.43 (0.805)	7.90	41.8	34.2	48.8	26.9	54.0	33.44
G28883C	G204R	N gene	1.79	7.90	41.8	34.0	48.3	26.7	54.0	33.30

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

			(0.285)							

US, United States; LA, Latin America; EU, Europe; AF, Africa; OC, Oceania.

2.4. Phylogeny and dynamics of the highly frequent global *dN* substitutions

Phylogenetic analysis, using the Nexstrain nomenclature [14], showed that the 2,213 genomes were grouped into 7 clusters (Figure 4). G614 was related to clade 19A and the emergence of clade 20A, while L323 was related to the emergence of clade 20A. Clades 20B and 20C, and the subclades 20A.EU1 and 20A.EU2 arose from clade 20A. K203 and R204 were related to the emergence of clade 20B, while I85 was related to the emergence of clade 20C. H57 and F120 emerged into clades 20A and 20B, respectively. Finally, subclade 20A.EU1 was related to G614 and L323, and subclade 20A.EU2 to G614, L323 and H57 (Figure 4).

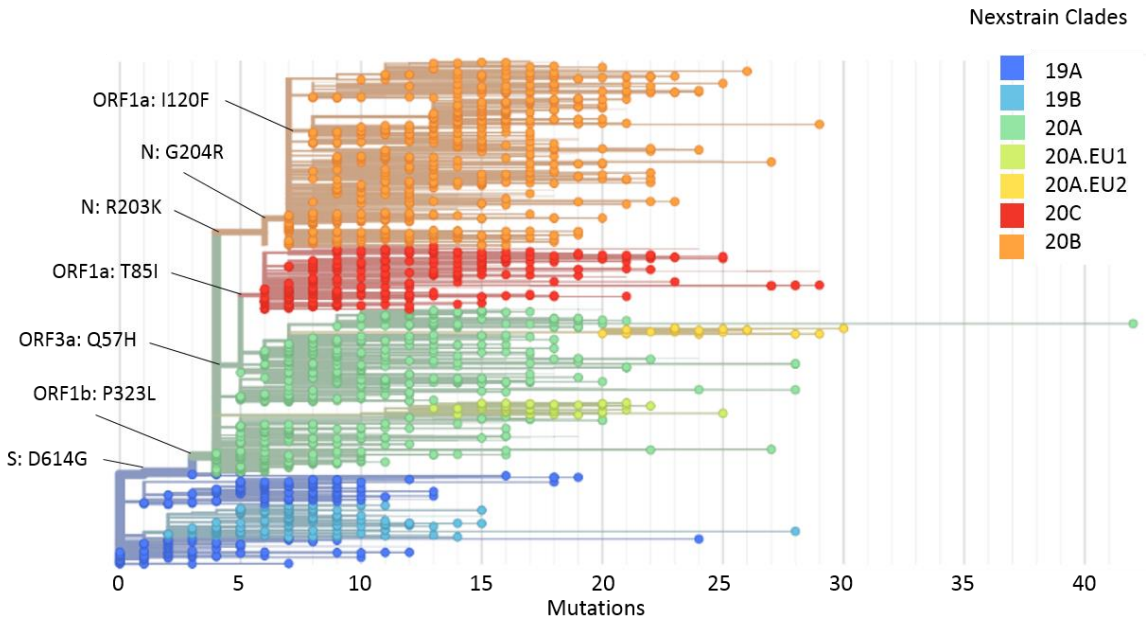
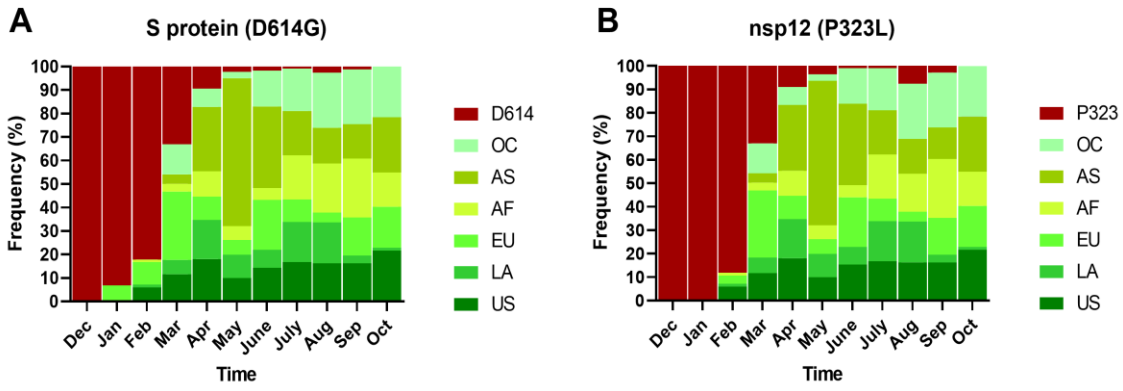


Figure 4. A Maximum-likelihood phylogeny of 2,213 SARS-CoV-2 genomes. The branches with tip circles represent the 2,213 genomes analyzed in the present work, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

branches without a tip circle represent the 1,888 reference genomes. The 7 *dN* substitutions analyzed here are located in the base of the nodes. The colour of each circle indicates the clades to which it belongs according to Nexstrain nomenclature.

Subsequently, we performed a spatial-temporal analysis of the *dN* substitutions with the highest global frequencies (>75%), G614 and L323 (Table 1). G614 was detected for the first time in January, and L323 in February, with both substitutions presenting a high increase in their frequencies between February and March. From April to September, these substitutions were present in >90% of the isolates analyzed each month, and in October, they presented in 100% of the isolates (Figure 5). By region, we observed fluctuations in their frequencies over time, but they were persistent in all regions. In the US and LA regions, both substitutions were detected from February to October; in EU, G614 was detected from January to October and L323 from February to October; in AF, both substitutions were detected from February to October, while in AS and OC, they were detected from March to October (Figure 5). November was not included in the analysis because we only sampled until 13th November and genomes could not be obtained from all regions. However, 43 isolates from EU, AF and OC were recovered in this month, and all of these presented the G614 and L323 substitutions.



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

Figure 5. Spatial-temporal frequencies of D614G (A) and P323L (B) substitutions. Dec, December; Jan, January, Feb, February, Mar, March; Apr, April; Aug, August; Sep, September; Oct, October. OC, Oceania; AS, Asia; AF, Africa; EU, Europe; LA, Latin America; US, United States of America. The number of isolates analyzed per month was as follows: Dec=15, Jan=103, Feb=84, Mar=628, Apr=222, May=221, June=118, July=233, Aug=196, Sep=179, and Oct=171.

Interestingly enough, we found that L323 and G614 showed similar frequencies and distributions, and both substitutions presented a strong linkage disequilibrium (LD) (R^2=0.944; p<0.001). The Nextrain phylogenetic tree indicated that these substitutions emerged early in the pandemic (G614, 2020-01-06 [IC 2019-12-27 – 2020-01-16]; L323, 2020-01-20, [IC 2019-01-11 – 2020-01-21]) and have spread all over the world (Figure S1).

2.5. Emergence and transmission of new variants of SARS-CoV-2

In addition to those previously described in the section 2.3 above, we investigated if there were *dN* substitutions with a significant increase in frequency by region. We found a *dN* substitution in the S gene (G22992A>S477N) with a *dN/dS* value of 1.92 (p=0.485) and a frequency of 42.6% (n=153) in the virus population from the OC region. An I120F substitution was also present in high frequency in OC (43.2%, n=155) (Table 1).

The Nexstrain phylogenetic tree showed that F120 emerged in late March (2020-03-21; IC 2020-03-12 - 2020-03-27) in AS and it spread in AS and OC regions (Figure S1). We detected F120 with moderate frequency (11.4%, n=67) in AS (Bangladesh) from April to July and again in October, and in a genome from EU (Wales) in September. Meanwhile,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

N477 emerged in late May (2020-05-27; IC 2020-05-08 - 2020-06-05) in OC and it spread throughout this region (Figure S1 and Figure 6A).

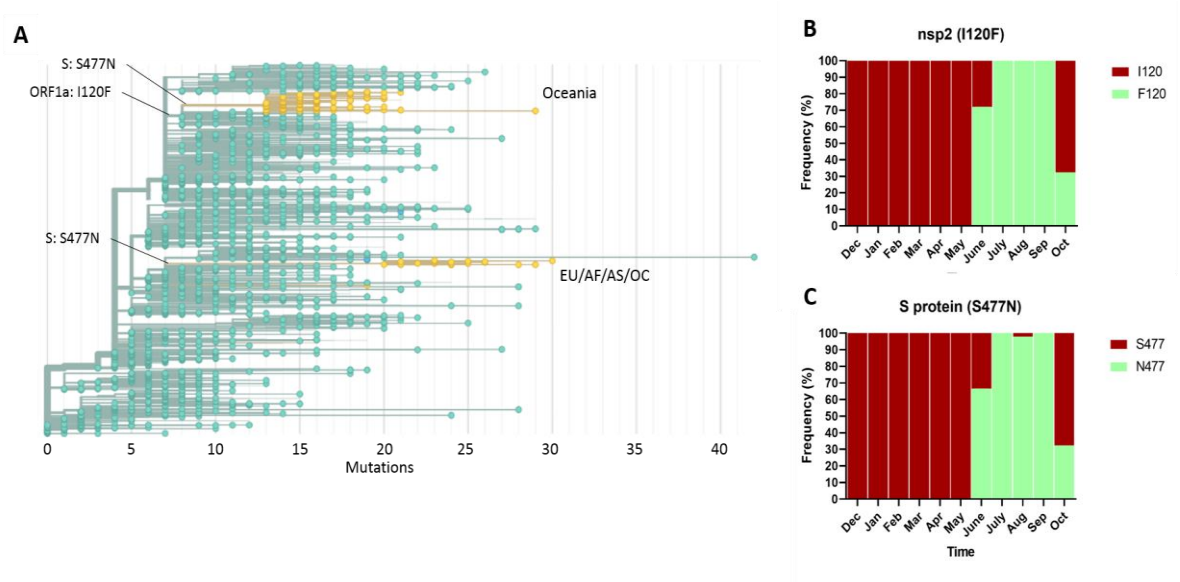


Figure 6. Maximum-likelihood phylogeny of 2,213 SARS-CoV-2 genomes (A), and temporal frequencies of I120F (B) and S477N (C) substitutions in Oceania. F120 and N477 are located in the base of the nodes. The presence of N477 is indicated in yellow. Dec, December; Jan, January, Feb, February, Mar, March; Apr, April; Aug, August; Sep, September; Oct, October. In Figures B and C, the number of isolates analyzed per month was as follows: Jan=1, Feb=3, Mar=144, Apr=17, May=6, June=18, July=42, Aug=46, Sep=42 and Oct=37.

F120 and N477 presented similar distributions and frequencies over time in OC, and were under strong linkage disequilibrium (R^2=-0.977; $p<0.001$). Both substitutions were detected from June to October, with their highest frequencies of 98-100% being seen in July-September. However, between September and October there was a dramatic decrease in their frequencies from 100% to 32.4% (Figure 6B and C).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

A second cluster carrying the S477N substitution that included genomes from EU (France, Netherlands, Norway, Belgium and Denmark; n=15, 78.95%), AF (Tunisia; n=2, 10.52), AS (Hong Kong; n=1, 5.26%) and OC (New Zealand; n=1, 5.26%) (Figure 6A), was detected during September and November. The phylogenetic trees showed that this cluster corresponded to subclade 20A.EU2 (Figure 4 and 6A). Moreover, one genome from AF (Ivory Coast) located in clade 20A also carried this substitution. The Nexstrain phylogenetic tree suggests that subclade 20A.EU2 emerged in EU during July (2020-07-24; IC 2020-07-09 - 2020-08-03) (Figure S1).

3. Discussion

Our results showed that the nucleotide diversity of the global population of SARS-CoV-2 has increased over time. Genome diversity was not homogeneous with regions showing high and low diversity. We found more than 3,000 mutations have emerged in the whole genome of the virus, and half of these have resulted in non-synonymous substitutions (*dN*), with most of them being neutral or likely neutral substitutions. The P323L and D614G substitutions in the global SARS-CoV-2 population have increased dramatically in their frequency over time. By October and November 2020, they were present in 100% of the virus population analyzed. Moreover, we detected 2 *dN* substitutions that spread in Oceania from July to October.

Analysis of the 2,213 SARS-CoV-2 genomes revealed that they shared a high nucleotide identity, suggesting that the genetic variation is limited withing the global population of the virus. In the whole genome, we detected genomic regions of high and low nucleotide diversity, implying that some genomic regions are evolving faster than others [15,16]. This difference between genomic region may be useful because regions with low diversity could

 doi:10.20944/preprints202101.0173.v1

be considered more suitable in which to develop and test new antiviral drugs, vaccines and detection methods (RT-PCR), to reduce the possibility of rapid drug resistance, immune system evasion and high numbers of false negatives when testing [17–19].

Global nucleotide diversity (π) varied over time and coincided with the increase of COVID-19 cases from December 2019 to October 2020. Previous studies have reported a positive association between sampling time and the evolution of the virus, indicating that more recent isolates have accumulated additional mutations more than older ones [15,19]. Although the number of samples per month in the current study was not homogeneous, the increase in diversity over time suggests that the global effective population size of SARS-CoV-2 is relatively high. Regionally, we also found a tendency for diversity to increase over time, however, we observed fluctuations in the π values, which could be explained by the sample size per month per region and the over representation of a few genotypes in a given time. Infection patterns during outbreaks that might occur in a region over a determined time period could result in the over-representation of some mutations, resulting in a decrease in genetic diversity and a similar effect to that of natural selection [12,13].

Although we found a great number of *dN* substitutions, it is still unclear if they play a significant role since most of them are neutral or likely neutral. Seven of them presented frequencies >10% in the global SARS-CoV-2 population, and were detected in nsp2 (T85I and I120F), nsp12 (P323L), S protein (D614G), ORF3a (Q57H) and N protein (R203K and G204R). Additionally, we found a substitution in the S protein (S477N) with a high frequency in OC. Although the *dN/dS* values were positive for T85I, I120F, G614, L323, Q57H, G204R and S477N, only F120 and L323 presented statistical significance, indicating positive natural selection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

Interestingly, we found that G614 and L323, and F120 and N477 presented a strong LD, suggesting that this LD is the result of natural selection, and that the average fitness of isolates that carry both mutations could overcome the adequacy of each substitution [20], and therefore, the LD among these two substitutions could persist over time [21]. However, more detailed bioinformatics and experimental analyses of LD, epistasis and natural selection will be needed to understand the evolution of these substitutions in more detail.

Our results, together with the Nexstrain phylogenetic analysis, showed that L323 and G614 emerged in early pandemic in EU and AS, respectively, and these have spread worldwide with dramatic increases in frequency over time. Other substitutions were more frequent by region, for example, F120 and N477 that were highly frequent in OC. F120 emerged in AS and was then introduced to OC where it spread, while N477 emerged and spread in OC. The phylogenetic analysis showed that N477 has also been detected in other regions, principally in EU where it formed a well-defined clade (20A.EU2). In OC, N477 could be the result of an outbreak during June-October with cryptic transmission of SARS-CoV-2 in the region, as with other outbreaks that have been reported in US [22]. A recent study reported the presence of N477 in EU during June-September increasing its frequency over time, principally in France [23]. Our results indicate that the presence of this substitution in OC and EU was the result of two independent events (homoplasy), but the few cases observed in AF, AS and OC in clade 20A.EU2 suggest genetic flow from EU to those regions.

Another homoplasy event was the emergence of the mutation A23063T>N501Y in England and South Africa. In the middle of December 2020, a new outbreak in England of a new SARS-CoV-2 strain (named linage B.1.1.7) was reported. The more significant changes in

 doi:10.20944/preprints202101.0173.v1

this strain were the mutations A23063T and C23604A that resulted in substitutions of N501Y and P681H, respectively, in the S protein. Although this strain was detected in late September, a rapid increase in its frequency occurred during December in England and its spread to other countries from the UK, Europe, Asia and America. Its rapid spread has been associated with the N501Y and P681H substitutions, which could be implicated in viral infectivity, but this is still under investigation [24–27]. In South Africa, the N501Y substitution was first detected in October [28], but recently, the British government reported 2 imported cases from South Africa [29]. The genomes shared the N501Y mutation but did not share the same mutations in the B1.1.7 linage.

The combination of several mutations and phylogenetic associations provides information to help determine the origin of the viral genotypes, and so theoretically, if we know the origin of the genotypes, the local and imported cases can be detected leading us to track the dynamics of viral spread at a local and global level. Thanks to molecular epidemiology, it has been possible to detect the emergence, introduction and transmission of new variants of the virus in different regions during this current pandemic [10,30–33]. This information is vital for developing public health interventions and policy to control viral spread.

Given its function, nsp12 is essential for the replication/transcription of the SARS-CoV-2 genome, and this protein is one of the targets for the treatment of COVID-19. The P323L substitution is located in an interphase region of nsp12, which has been reported to play an important role in the formation of a protein complex with nsp7 and nsp8 [34], which provides structural stability to nsp12 for its processivity [35,36]. However, a previous report suggests that L323 could have structural alterations [37] and an adverse effect on proofreading during the genomic replication of the virus [15]. Meanwhile, the P323L

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

substitution is located in a pocket that has been predicted as a possible druggable site [38], however, will be necessary to analyze if the mutation could affect these properties.

The S protein is a key factor for the entry of the virus to the host cell [39]. This protein has a receptor binding motif (RBM) that interacted with the ACE-2 receptor of the host cell [40]. The S447N substitution is located in the RBM and a recent study showed that S477 increases the affinity for the ACE-2 receptor [41]. Moreover, this substitution is part of an epitope recognized by human neutralizing antibodies [42], but further analysis is required to determine if N477 altered recognition by the human antibodies.

G614 has gained relevance since a correlation was shown to exist between the presence of this substitution and a higher capacity for infection by SARS-CoV-2 [43–45]. Moreover, studies *in vitro* have shown that this substitution is responsible for making the virus 2.4 times more infectious [46]. It has also been reported that the viral load in COVID-19 patients is higher than in those patients with isolates that do not present this substitution [47]. Furthermore, the S protein is one of the elements identified as being useful in the development of vaccines, while peptide 614-621 has been identified as an epitope of B cells [48]. However, neutralization of the virus by antibodies from convalescent plasma decreases when this substitution is present [46], which suggests that these behavioral characteristics affect the antigenicity of the S protein. Currently, there are several vaccines based on the S protein [49]. An effective COVID-19 vaccine could be the proximal solution to the SARS-CoV-2, nevertheless, the genetic diversity in the S protein and its implication in host immune evasion must be taken into account in the development of future vaccines.

Finally, nsp2 is a helical transmembrane protein implicated in the modulation of the host cell environment [50], although its precise function remains unknow. Previous studies have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

reported that a stabilizing mutation in the endosome-associate-protein-like domain of the nsp2 could be associated with the more contagious phenotype of SARS-CoV-2 when compared with SARS-CoV [51]. The I120F substitution occurs in the N-terminal of nsp2, which is located in the extracellular region of the protein. Further study of its possible implications in virus fitness is needed.

One year after the emergence of the SARS-CoV-2, the virus continues to mutate, and it will keep accumulating novel mutations with possible clinical and therapy repercussions forcing the development of new strategies to reduce the burden of COVID-19. Molecular epidemiologic surveillance needs to continue in order to detect genetic changes that might be involved in pathogenesis, host immune system evasion or future drug resistance, as well as its worldwide spread. Such information will contribute greatly to develop more efficient intervention practices for SARS-CoV-2, as well as provide a solid foundation for tackling other viral pandemics in the future.

4. Materials and methods

4.1. Sequences, alignments and quality control

A total of 2,500 complete genomes of SARS-CoV-2 from six regions around the world [United States of America (US), Latin America (LA), Europe (EU), Africa (AF), Asia (AS) and Oceania (OC)] were obtained randomly from NCBI [52] and GISAID [53] databases up to November 13, 2020. The genomes were aligned using MAFFT v7.3 [54] and revised by BioEdit v7.2 software [55], using the isolate from Wuhan, Hu-1 (GenBank: NC045512) as the reference strain. Non-coding regions were eliminated, as were all genomes that presented more than 15 non-determined (N) or other ambiguous nucleotides according to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 January 2021 doi:10.20944/preprints202101.0173.v1

the IUPAC nucleotide code. For final analysis, we included 2,213 genomes from 29,256 nucleotides distributed throughout the period between December 2019 and November 2020 (Supplementary Table S1).

4.2. Genetic analyses

We used DnaSP v5.1 software [56] to determine the number of polymorphisms (*S*), nucleotide diversity (π), the number of non-synonymous (*dN*) substitutions and linkage disequilibrium (LD) given by the R^2 index. The variations of π throughout the genome were estimated using a 50 bp window at 10 bp steps. To determine if diversity moves away from neutrality, the difference between synonymous and non-synonymous substitutions (*dN/dS*) was evaluated using the software MEGA v6.0 [57]. This estimation was based on the maximum joint likelihood of ancestral reconstruction states under the the Muse-Gaut models [58] and Felsenstein's codon substitution [59]. Moreover, the software calculates the probability of rejecting the null hypothesis of neutral evolution (p value). We obtained *dN* frequencies using *Jalview* v2.11 software [60].

4.3. Phylogenetic analysis

A maximum likelihood phylogenetic tree of the 2,213 SARS-CoV-2 genomes analyzed in this study on a background of 1,888 reference genomes was constructed in Nextstrain [28] on November 30, 2020. Additionally, we obtained a maximum likelihood phylogenetic tree (named Nexstrain phylogenetic tree) of SARS-CoV-2 obtained from Nextstrain [28] on the same date in order to localize the nucleotide changes and *dN* substitutions into their respective clades together with divergence times.

Supplementary Materials

 doi:10.20944/preprints202101.0173.v1

Table S1. Accession numbers, database, region, country and collection date of the 2,213 genomes analyzed.

Figure S1. Maximum-likelihood phylogeny of 3,156 SARS-CoV-2 genomes deposited in the GISAID database. The 8 *dN* substitutions analyzed here are located in the base of the nodes. The tree was constructed in the Nexstrain website [28], which allowed us to consult the divergence time of each node associated with each *dN* substitution.

Acknowledgments

We thank Luisa Sandner Miranda, PhD. for helpful discussions.

Funding

This work was supported by DGAPA-PAPIIT grant IN213816.

Author Contributions

R.M.E. and A.F.A. contributed to the study design; A.F.A. performed the data curation, genetic and phylogenetic analyses, and the interpretation of results; F.R.G. contributed to the interpretation of results; A.F.A. and R.M.E. wrote the original manuscript; R.M.E., F.R.G., A.C.R., J.G., C.A.T.G., G.D. and A.C. made a critical review and edited the final manuscript. All authors read and approved the final version of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard | WHO Coronavirus Disease

(COVID-19) Dashboard Available online: https://covid19.who.int/ (accessed on Jan 2, 2021).

2. Gorbalenya, A.E.; Baker, S.C.; Baric, R.S.; de Groot, R.J.; Drosten, C.; Gulyaeva, A.A.; Haagmans, B.L.; Lauber, C.; Leontovich, A.M.; Neuman, B.W.; et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **2020**, *5*, 536–544, doi:10.1038/s41564-020-0695-z.

3. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269, doi:10.1038/s41586-020-2008-3.

4. von Brunn, A.; Teepe, C.; Simpson, J.C.; Pepperkok, R.; Friedel, C.C.; Zimmer, R.; Roberts, R.; Baric, R.; Haas, J. Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. *PLoS One* **2007**, *2*, doi:10.1371/journal.pone.0000459.

5. Khailany, R.A.; Safdar, M.; Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene Reports* **2020**, *19*, 100682, doi:10.1016/j.genrep.2020.100682.

6. Cui, J.; Li, F.; Shi, Z.L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 2019, *17*, 181–192.

7. Sanjuán, R.; Nebot, M.R.; Chirico, N.; Mansky, L.M.; Belshaw, R. Viral Mutation Rates. *J. Virol.* **2010**, *84*, 9733–9748, doi:10.1128/jvi.00694-10.

8. Peck, K.M.; Lauring, A.S. Complexities of Viral Mutation Rates. *J. Virol.* **2018**, *92*, 1–8, doi:10.1128/jvi.01031-17.

9. Ogando, N.S.; Ferron, F.; Decroly, E.; Canard, B.; Posthuma, C.C.; Snijder, E.J. The Curious Case of the Nidovirus Exoribonuclease: Its Role in RNA Synthesis and Replication Fidelity. *Front. Microbiol.* 2019, *10*.

10. Bedford, T.; Riley, S.; Barr, I.G.; Broor, S.; Chadha, M.; Cox, N.J.; Daniels, R.S.; Gunasekaran, C.P.; Hurt, A.C.; Kelso, A.; et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift Europe PMC Funders Group. *Nature* **2015**, *523*, 217–220, doi:10.1038/nature14460.

11. Holmes, E.C.; Dudas, G.; Rambaut, A.; Andersen, K.G. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* **2016**, *538*, 193–200, doi:10.1038/nature19790.

12. Liu, Q.; Zhao, S.; Shi, C.M.; Song, S.; Zhu, S.; Su, Y.; Zhao, W.; Li, M.; Bao, Y.; Xue, Y.; et al. Population Genetics of SARS-CoV-2: Disentangling Effects of Sampling Bias and Infection Clusters. *Genomics, Proteomics Bioinforma.* **2020**, 4–11, doi:10.1016/j.gpb.2020.06.001.

13. Vitti, J.J.; Grossman, S.R.; Sabeti, P.C. Detecting natural selection in genomic data. *Annu. Rev. Genet.* 2013, *47*, 97–120.

14. Nextclade Available online: https://clades.nextstrain.org/ (accessed on Dec 16, 2020).

15. Pachetti, M.; Marini, B.; Benedetti, F.; Giudici, F.; Mauro, E.; Storici, P.; Masciovecchio, C.; Angeletti, S.; Ciccozzi, M.; Gallo, R.C.; et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **2020**, *18*, 1–9, doi:10.1186/s12967-020-02344-6.

16. Day, T.; Gandon, S.; Lion, S.; Otto, S.P. On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* **2020**, *30*, R849–R857, doi:10.1016/j.cub.2020.06.031.

17. Ramírez, J.D.; Florez, C.; Muñoz, M.; Hernández, C.; Castillo, A.; Gomez, S.; Rico, A.; Pardo, L.; Barros, E.C.; Castañeda, S.; et al. The arrival and spread of SARS-CoV-2 in Colombia. *J. Med. Virol.* **2020**, jmv.26393, doi:10.1002/jmv.26393.

18. Wright, E.; Lakdawala, S.; Cooper, V. SARS-CoV-2 genome evolution exposes early human adaptations. *bioRxiv* **2020**, 2020.05.26.117069, doi:10.1101/2020.05.26.117069.

19. van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T.; et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *83*, 104351, doi:10.1016/j.meegid.2020.104351.

20. Felsenstein, J. THE EFFECT OF LINKAGE ON DIRECTIONAL SELECTION. *Genetics* **1965**, *52*.

21. Karlin, S.; Feldman, M.W. Linkage and selection: Two locus symmetric viability model. *Theor. Popul. Biol.* **1970**, *1*, 39–71, doi:10.1016/0040-5809(70)90041-9.

22. Bedford, T.; Greninger, A.L.; Roychoudhury, P.; Starita, L.M.; Famulare, M.; Huang, M.L.; Nalla, A.; Pepper, G.; Reinhardt, A.; Xie, H.; et al. Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **2020**, *370*, 571–575, doi:10.1126/science.abc0523.

23. Hodcroft, E.B.; Zuber, M.; Nadeau, S.; Crawford, K.H.D.; Bloom, J.D.; Stadler, T.; Neher, R.A. Emergence and spread of a SARS-CoV-2 variant through Europe in the

summer of 2020 SeqCOVID-SPAIN consortium, 14. *medRxiv* **2020**, 2020.10.25.20219063, doi:10.1101/2020.10.25.20219063.

24. Wise, J. Covid-19: New coronavirus variant is identified in UK. *BMJ* **2020**, m4857, doi:10.1136/bmj.m4857.

25. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology - Virological Available online: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed on Dec 23, 2020).

26. 32 More Countries Have Found the New Covid-19 Variant First Seen in Britain - The New York Times Available online: https://www.nytimes.com/live/2021/01/01/world/covid-19-coronavirus-updates (accessed on Jan 2, 2021).

27. Gu, H.; Chen, Q.; Yang, G.; He, L.; Fan, H.; Deng, Y.Q.; Wang, Y.; Teng, Y.; Zhao, Z.; Cui, Y.; et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science (80-.).* **2020**, *369*, 1603–1607, doi:10.1126/science.abc4730.

28. Nextstrain / ncov / global Available online: https://nextstrain.org/ncov/global (accessed on Nov 30, 2020).

29. What do we know about the two new Covid-19 variants in the UK? | World news | The Guardian Available online: https://www.theguardian.com/world/2020/dec/23/what-do-we-know-about-the-two-

new-covid-19-variants-in-the-uk (accessed on Dec 23, 2020).

30. Puenpa, J.; Suwannakarn, K.; Chansaenroj, J.; Nilyanimit, P.; Yorsaeng, R.; Auphimai, C.; Kitphati, R.; Mungaomklang, A.; Kongklieng, A.; Chirathaworn, C.; et al. Molecular epidemiology of the first wave of severe acute respiratory syndrome coronavirus 2 infection in Thailand in 2020. *Sci. Rep.* **2020**, *10*, 1–8, doi:10.1038/s41598-020-73554-7.

31. Laiton-Donato, K.; Villabona Arenas, C.J.; Usme Ciro, J.; Franco Munoz, C.; Alvarez-Diaz, D.A.; Villabona-Arenas, L.; Echeverria-Londono, S.; Franco-Sierra, N.; Cucunuba, Z.; Florez-Sanchez, A.C.; et al. Genomic epidemiology of SARS-CoV-2 in Colombia. *medRxiv* **2020**, 2020.06.26.20135715, doi:10.1101/2020.06.26.20135715.

32. Lemieux, J.E.; Siddle, K.J.; Shaw, B.M.; Loreth, C.; Schaffner, S.F.; Gladden-Young, A.; Adams, G.; Fink, T.; Tomkins-Tinch, C.H.; Krasilnikova, L.A.; et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science (80-.).* **2020**, eabe3261, doi:10.1126/science.abe3261.

33. Lai, A.; Bergna, A.; Caucci, S.; Clementi, N.; Vicenti, I.; Dragoni, F.; Cattelan, A.M.; Menzo, S.; Pan, A.; Callegaro, A.; et al. Molecular Tracing of SARS-CoV-2 in Italy in the First Three Months of the Epidemic. *Viruses* **2020**, *12*, 798, doi:10.3390/v12080798.

34. Gao, Y.; Yan, L.; Huang, Y.; Liu, F.; Zhao, Y.; Cao, L.; Wang, T.; Sun, Q.; Ming, Z.; Zhang, L.; et al. Structure of the RNA-dependent RNA polymerase from

COVID-19 virus. *Science (80-.).* **2020**, *368*, 779–782, doi:10.1126/science.abb7498.

35. Subissi, L.; Posthuma, C.C.; Collet, A.; Zevenhoven-Dobbe, J.C.; Gorbalenya, A.E.; Decroly, E.; Snijder, E.J.; Canard, B.; Imbert, I. One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, E3900–E3909, doi:10.1073/pnas.1323705111.

36. Wang, R.; Chen, J.; Gao, K.; Hozumi, Y.; Yin, C.; Wei, G.W. Characterizing SARS-CoV-2 mutations in the united states. *arXiv* **2020**, doi:10.21203/rs.3.rs-49671/v1.

37. Chand, G.B.; Banerjee, A.; Azad, G.K. Identification of novel mutations in RNA-dependent RNA polymerases of SARS-CoV-2 and their implications on its protein structure. *PeerJ* **2020**, *8*, e9492, doi:10.7717/peerj.9492.

38. Ruan, Z.; Liu, C.; Guo, Y.; He, Z.; Huang, X.; Jia, X.; Yang, T. SARS-CoV-2 and SARS-CoV: Virtual screening of potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12). *J. Med. Virol.* **2020**, jmv.26222, doi:10.1002/jmv.26222.

39. Ye, Z.-W.; Yuan, S.; Yuen, K.-S.; Fung, S.-Y.; Chan, C.-P.; Jin, D.-Y. Zoonotic origins of human coronaviruses. *Int. J. Biol. Sci.* **2020**, *2020*, 1686–1697, doi:10.7150/ijbs.45472.

40. Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **2020**, *581*, 215–220, doi:10.1038/s41586-020-2180-5.

41. Starr, T.N.; Greaney, A.J.; Hilton, S.K.; Ellis, D.; Crawford, K.H.D.; Dingens, A.S.; Navarro, M.J.; Bowen, J.E.; Tortorici, M.A.; Walls, A.C.; et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **2020**, *182*, 1295-1310.e20, doi:10.1016/j.cell.2020.08.012.

42. Barnes, C.O.; Jette, C.A.; Abernathy, M.E.; Dam, K.M.A.; Esswein, S.R.; Gristick, H.B.; Malyutin, A.G.; Sharaf, N.G.; Huey-Tubman, K.E.; Lee, Y.E.; et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **2020**, doi:10.1038/s41586-020-2852-1.

43. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812-827.e19, doi:10.1016/j.cell.2020.06.043.

44. Brufsky, A. Distinct viral clades of SARS-CoV-2: Implications for modeling of viral spread. *J. Med. Virol.* **2020**, *92*, 1386–1390, doi:10.1002/jmv.25902.

45. Brufsky, A.; Lotze, M.T. DC/L-SIGNs of hope in the COVID-19 pandemic. *J. Med. Virol.* **2020**, *92*, 1396–1398, doi:10.1002/jmv.25980.

46. Hu, J.; He, C.L.; Gao, Q.; Zhang, G.J.; Cao, X.X.; Long, Q.X.; Deng, H.J.; Huang, L.Y.; Chen, J.; Wang, K.; et al. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity. *bioRxiv* **2020**, 2020.06.20.161323, doi:10.1101/2020.06.20.161323.

47. Yao, H.; Lu, X.; Chen, Q.; Xu, K.; Chen, Y.; Cheng, M.; Chen, K.; Cheng, L.;

Weng, T.; Shi, D.; et al. Patient-derived SARS-CoV-2 mutations impact viral replication dynamics and infectivity in vitro and with clinical implications in vivo. *Cell Discov.* **2020**, *6*, 1–16, doi:10.1038/s41421-020-00226-1.

48. Kim, S.J.; Nguyen, V.G.; Park, Y.H.; Park, B.K.; Chung, H.C. A novel synonymous mutation of SARS-COV-2: Is this possible to affect their antigenicity and immunogenicity? *Vaccines* **2020**, *8*, doi:10.3390/vaccines8020220.

49. Covid-19 Vaccine Tracker: Latest Updates - The New York Times Available online: https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html (accessed on Dec 11, 2020).

50. Davies, J.; Almasy, K.; McDonald, E.; Plate, L. Comparative multiplexed interactomics of SARS-CoV-2 and homologous coronavirus non-structural proteins identifies unique and shared host-cell dependencies. *bioRxiv Prepr. Serv. Biol.* **2020**, doi:10.1101/2020.07.13.201517.

51. Angeletti, S.; Benvenuto, D.; Bianchi, M.; Giovanetti, M.; Pascarella, S.; Ciccozzi, M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. *J. Med. Virol.* **2020**, *92*, 584–588, doi:10.1002/jmv.25719.

52. SARS-CoV-2 Resources - NCBI Available online: https://www.ncbi.nlm.nih.gov/sars-cov-2/ (accessed on Nov 30, 2020).

53. GISAID - Initiative Available online: https://www.gisaid.org/ (accessed on Nov 30, 2020).

54. Katoh, K.; Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 2010, *26*, 1899–1900.

55. Hall, T.A. BioEdit: a user-friendly biological sequences alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **1999**, *41*, 95–98.

56. Librado, P.; Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinforma. Appl. NOTE* **2009**, *25*, 1451–1452, doi:10.1093/bioinformatics/btp187.

57. Tamura, K.; Stecher, G.; Peterson, D.; Filipski, A.; Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **2013**, *30*, 2725–2729, doi:10.1093/molbev/mst197.

58. Muse, S. V.; Gaut, B.S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **1994**, *11*, 715–724, doi:10.1093/oxfordjournals.molbev.a040152.

59. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **1981**, *17*, 368–376, doi:10.1007/BF01734359.

60. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.A.; Clamp, M.; Barton, G.J. Sequence analysis Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinforma. Appl. NOTE* **2009**, *25*, 1189–1191, doi:10.1093/bioinformatics/btp033.