Worldwide Geographical and Temporal Analysis of SARS-CoV-2 Haplotypes shows Differential Distribution Patterns

Santiago Justo Arévalo[1,2,*], Daniela Zapata Sifuentes[1,±], César Huallpa Robles[3,±], Gianfranco Landa Bianchi[1,±], Adriana Castillo Chávez[1,±], Romina Garavito-Salini Casas[1,±], Guillermo Uceda-Campos[2,4] Roberto Pineda Chavarría[1].

1.- Universidad Ricardo Palma, Facultad de Ciencias Biológicas, Lima – Perú.
2.- Universidade de Sao Paulo, Instituto de Química, Departamento de Bioquímica, São Paulo - Brasil
3.- Universidad Nacional Agraria La Molina, Facultad de Ciencias, Lima – Perú.
4.- Universidad Nacional Pedro Ruiz Gallo, Facultad de Ciencias Biológicas, Lambayeque - Perú
± These authors contributed equally to this work.
* Corresponding author: santiago.justo@urp.edu.pe

ABSTRACT:

Since the identification of SARS-CoV-2 in December 2019 a large number of SARS-CoV-2 genomes has been sequenced with unprecedented speed around the world and deposited in several databases. This marks a unique opportunity to study how a virus spread and evolve in a worldwide context. However, currently there is not a useful haplotype classification system to help tracking the virus evolution. Here we identified eleven mutations with 10 % or more frequency in a data set of 7848 genomes. Using these mutations, we identified 6 SARS-CoV-2 haplotypes or OTUs (Operational Taxonomic Unit) that correlate well with a phylogenomic tree. After that, we analyzed the geographical and temporal distribution of these OTUs, as well as their correlation with patient status. Our geographical analysis showed different OTUs prevalence between continents and the temporal distribution analysis revealed an evolution-like pattern in SARS-CoV-2. Finally, we observed a homogenous distribution of OTUs in mild and severe patients and a great prevalence of OTU 2 in asymptomatic patients. However, genomes in the asymptomatic category, comes from isolates on three consecutive days in February (15 to 17), weakening this observation and highlighting the need to increase genomic analyzes in asymptomatic and severe patients. Our classification system is phylogenetically consistent and allows us to easily track geographic and temporal distribution of important mutations around the world. In the next months, it could be updated using similar steps that we used here.

INTRODUCTION:

COVID-19 was declared a pandemic by the World Health Organization on March 11th 2020[1], with around 10 million cases and 500 thousand of deaths around the world[2], quickly becoming the most important health concern in the world. Several efforts to produce vaccines, drugs and diagnostic tests to help in the fight against SARS-CoV-2 are being mounted in a large number of laboratories all around the world.

Since the publication on January 24th of the first complete genome sequence of SARS-CoV-2 from China[3], thousands of genomes have been sequenced in a great number of countries on all 5 continents and were made available in several databases. This marks a milestone in scientific history and gives us an unprecedented opportunity to study how a specific virus evolves in a worldwide context. As of June 09, 2020, the GISAID database[4] contained 27542 genomes with at least 29000 sequenced bases.

At the moment, some analysis have been performed to identify SARS-CoV-2 variants around the world, most of them on a particular group of genomes and/or at the beginning of the pandemic using limited datasets. In March, 2020 two major lineages were proposed based in position 8782

and 28144 using a data set of 103 genomes[5] which was followed by a particularly interesting proposal that identified the same major lineages (named A and B) and others sublineages[6]. Identification of SARS-CoV-2 variants aids in the understanding the evolution of the virus and may improve our efforts to control the disease.

We first used a "high-quality" data set of 7848 genomes from GISAID to perform a phylogenetic analysis and correlated the major branches with SARS-CoV-2 variants which can be classified into six haplotypes or OTUs (Operational Taxonomic Units) based on the distribution of the eleven most frequent mutations. After that, we classified 27542 genomes in these OTUs and analyzed the current geographical and temporal worldwide distribution of OTUs and attempt to correlate them with patient status information. We also discuss the possible implications of the most frequent mutations on protein structure and function.

RESULTS AND DISCUSSION:

Mutations frequency analysis

The GISAID database contains around 27542 genomes with at least 29000 sequenced bases (as of June 09th). To obtain high-quality data to perform the mutation frequency analysis we first filtered 19694 genomes that not meet our "high-quality" criteria (see material and methods).

With an alignment of our "high-quality" data set of 7848 genomes we performed identity (respect to the reference sequence EPI_ISL_406801) and frequency analysis at each genomic position. We identified 11 positions with less than 0.9 identity (Fig. 1.A and S1.A) plus another other 15 positions with identities between 0.98 and 0.9 (Fig. S1.B) and many more positions with higher identities (Fig. S1.C).

The eleven most frequent mutations (less than 0.9 identity) are comprised of eight non-synonymous mutations, two synonymous mutations and one mutation in the 5'UTR region of SARS-CoV-2 genome (Fig. 1.A). All these mutations have been already identified in other studies[8,9,10], although with different frequencies.

OTUs identification

After the identity and frequency analysis, we estimated a maximum-likelihood tree using the whole-genome alignment of the 7848 genomes. Then, we associated the main branches of the whole-genome tree with an alignment of the 11 positions (241, 1059, 3037, 8782, 14408, 23403, 25563, 28144, 28881, 28882, 28883) and noted that combinations of those 11 positions represent 6 well-defined groups in the tree (Fig. 1.B). Using these combinations, we defined 6 haplotypes that allow us to classified more than 97 % of the analyzed genomes (Fig. 1.C). We named these haplotypes Operational Taxonomic Units (OTUs) and numbered them according to proximity to the root.

We were able to clearly track the mutations that originated each of these OTUs. OTU_1 is the ancestor haplotype with characteristic T8782 and C28144. Although a genome of this haplotype was not the first to be sequenced, these two positions (8782 and 28144) are also T and C, respectively, in the RaTG13 coronavirus[5]. This observation supports the idea that OTU_1 is the ancestor haplotype. OTU_2 differ from OTU_1 in position 8782 and 28144 (C and T, respectively in OTU_2). These two OTUs were already described in other analysis; in the begin of the pandemic, Tang et al (2020) shows linkage disequilibrium between those positions and named them as S (OTU_1) and L (OTU_2) lineages. Rambaut et al (2020) used these positions to discriminate between their proposed major lineages A (OTU_1) and B (OTU_2).

A SARS-CoV-2 isolated in February 20 was the first that shows simultaneously four mutations different to OTU_2 (C241T, C3037T, C14408T and A23403G) (Fig. S3). The haplotype containing the previous two mutations and these four mutations was named OTU_3. We can note in the phylogenetic tree that in the transition between the first clades and OTU_3 some unclassified tips were showed. These could be genomes containing some of these four mutations (C241T, C3037T, C14408T, A23403G) but not all, representing intermediate steps in the formation of this haplotype. OTU_3 is the first group containing the D614G mutation in the spike protein. Korber et al. 2020 analyzed the temporal and geographic distribution of this mutation separating SARS-CoV-2 populations into two groups, those with D614 and those with G614.

Almost at the same time (February 24), SARS-CoV-2 with three adjacent mutations (G28881A, G28882A and G28883C) (Fig. S3) in N protein was isolated. These three mutations characterize OTU_4. The maximum likelihood tree shows that OTU_5 comes from OTU_3. OTU_5 does not present mutations in N protein, instead, it presents a variation in Orf3a (G25563T). Finally, OTU_6 present all the mutations of OTU_5 plus Nsp2 mutation (C1059T).

These 11 mutations have been separately described in other reports but, to our knowledge, they have not yet used been used together to classify SARS-CoV-2 variants during the pandemic. The fact that we were able to classify more than 97 % of the genomes in our "high-quality" data set (Fig. 1.C) shows that, at least to the present date, this classification system covers almost all the currently known genomic information around the world. Thus, at the moment this system can be of practical use to analyze the geographical and temporal distribution of haplotypes during these first five months of 2020.

It is highly likely is that during the next months, some of these OTUs will disappear and others will appear when new mutations in these "parental" OTUs become fixed in the population. Thus, methodologies to actively update this classification system on a real-time basis need to be proposed. We propose that the best strategy will be to continually monitor the appearance of new lineages by tracking mutations that exceed a fixed percentage of genomes in the database (to allow tracking of relevant medically mutations) and with the association of these mutations to a phylogenomic tree to confirm its phylogenetical relevance.

Unfortunately, the constancy of viral genomes sequencing efforts is highly variable around the world and over time, and it is likely that over time the amount of new genetic information of SARS-CoV-2 isolates will decrease. As we noted above, the number of sequenced genomes in April was much less than in March (Fig.S2). Thus, likely the cut-off used here to track a mutation (0.1 identity) could be changed in the next months.

Worldwide geographic distribution of OTUs

Using our OTUs classification, we analyzed the worldwide geographic distribution during the first five months of 2020. We began by plotting continental information in the unrooted tree of our "high-quality" data set (Fig. 2.A) and observed some interesting patterns. For instance, OTU_2 seems to be more prevalent in Asia, OTU_3 and 4 in Europe, and OTU_6 in North America (Fig. 2.A). However, this approach does not allow us to evaluate continents with less sequenced genomes (Fig. S2), such as South America, Oceania and Africa.

To better analyze which were the most prevalent OTUs in each continent, we analyzed all the complete genomes in GISAID database (27542 genomes). In this analysis, we randomly separate the genomes into 6 groups of 4590 genomes each and classified them into OTUs. Then, we independently analyzed each continent and sample, taking the percentages of OTUs in each

month and multiplying this percentages by the number of cases reported during that month. Finally, these results were added to obtain a value for each OTU and the means of the six samples were statistically compared (see material and methods for details).

This approach more clearly illustrates that OTU_6 was the most prevalent in North America, followed by OTUs 1, 2, 3 and 5, the less prevalent was OTU_4 (Fig. 2.B). First genomes in North America belonged to OTU_1 and 2, March and April showed similar OTUs counts with a tendency to OTU_6 increment (Fig. S4). Finally, in May OTU_6 appear well established in North America (Fig. S4). OTU_6 has 8 of the 11 high frequency genomic variations described (all except those in N protein) (Fig. 5.C).

South America presents a more homogeneous distribution with OTU_4 in a statistically significant greater proportion (Fig. 6.C). Apparently, this greater OTU_4 proportion in South America was established in April (Fig. S4). Similarly, OTU_4 was most prevalent in Europe (Fig. 6.D). OTU_3 appear slightly less represented than OTU_4 but significatively than the other OTUs (Fig. 6.D). At the haplotype level, OTU_4 present mutations in N protein that could significantly increase the fitness of this group in comparison with OTU_3 that does not present mutations in N (Fig. 1.A). We therefore believe that is important to more deeply study the biological implications of these mutations.

Oceania presents OTU_2, 3 and 6 as the most prevalent (Fig. 6.F); in May, apparently, it shows a homogeneous distribution of OTUs in the genomes sequenced (Fig. S4). The analysis of Oceania is in part biased due the great percentage of genomes that we are not unambiguously classified most genomes sequenced in Australia do not cover position 241 (in the 5` UTR region).

OTU_3 was the more prevalent in Africa, with all other OTUs represented in similar percentages (Fig. 6.G).

In the case of Asia, we observed OTU_2 as the most prevalent during the months analyzed (Fig. 6.E). A detailed analysis of the sequenced genome counts by months in Asia (Fig. S4) shows that this great OTU_2 representation is due to a high number of isolates during the first months of the pandemic that correlated with the large number of cases in those months in Asia. Thus, the first epicenter of the pandemic (Asia) was dominated by OTU_2; after that, as in other continents, new haplotypes began to prevail.

Worldwide temporal distribution of OTUs

A rooted tree was estimated with our 7848 “high-quality” genomes data set and labeled by date (Fig. 3.A). Here we can clearly follow the evolution beginning with OTU_1 and 2 at the base of the tree (labeled with colors that correspond to the first months of the pandemic). The clade where OTU_3 is the most prevalent has intermediate temporal distribution (late February up to late April) and OTU_4, 5 and 6 have a more recent distribution pattern with representatives isolated up to May.

To gain more insight into these patterns, we estimated the most prevalent OTUs during each month of the pandemic following similar steps that those done for continents (see material and methods). In this analysis, we did not consider December that present genomes just belonging to OTU_2 (Fig.S4), or January that has representatives from OTU_1 and OTU_2 mainly from Asia (Fig.S4 and Fig.S5).

Analysis using the data of February from North America, Europe and Asia showed that OTU_1 and 2 continue as the most prevalent in the world but with presence of OTU_3, 4, 5 and 6 (Fig.

3.B). Analysis by continents showed that during this month Asia and North America still had higher proportions of OTU_1 and 2, but in Europe a more homogeneous distribution of OTUs 2-6 was observed (Fig. S4).

In March, when the epicenter of pandemic move to Europe and North America, but cases were still appearing in Asia, OTU_2, 3, 4 and 6 appear as the most prevalent (Fig. 3.C). Interestingly OTU_5 remained in relatively low proportions (Fig. 3.C). Apparently, this month contain the more homogenous OTUs distribution in a worldwide context, but with some OTUs more prevalent in each continent (Fig. S4).

During April, OTU_1 continued its downward while OTU_5 increased its presence (Fig. 3.D) probably due its higher representation (compared to March) in several continents such as South America, North America and Asia (Fig. S4). During this month we also witnessed the establishment of OTU_4 in South America and Europe and OTU_6 in North America (Fig. S4).

The last month analyzed (May) shows OTU_1 and 2 as the less prevalent; OTU_3, 4 and 5 in similar proportions and OTU_6 as the significantly most prevalent in the world (Fig. 6.E) probably because of its dominance in North America (Fig. S4). In this month, also we observed a preponderance of isolated genomes belonging to OTU_5 in Asia (Fig. S4).

Analysis of OTUs and patient status relation

We attempt to analyze the patient status information in the GISAID database for roughly two thousand genomes. Unfortunately, GISAID categories in patient status is not well organized and we had to reclassify its information into four categories, Not Informative, Asymptomatic, Mild and Severe (Fig. S6.A).

Using this classification scheme, we found that 44.83 % falls in the Not Informative category, 48 % in the Mild category, and just 3.22 % and 3.95 % of the genomes with patient status information could be classified as Asymptomatic and Severe, respectively (Fig. S6.B). We analyzed the group distribution in the three informative categories (Asymptomatic, Mild and Severe).

Isolates from patients with mild symptoms presented a relatively homogeneous distribution, with percentages between 21.7 % and 7.8 % from all six OTUs. The severe category was also relatively homogeneous with OTU_1 being the least prevalent (3.7 %). Conversely, 86.4 % (57 of the 66) of the genomes classified as Asymptomatic belong to OTU_2 (Fig. 4.B).

However, we have to interpret these observations with extreme caution since all the asymptomatic genomes that belong to OTU_2 were isolated in Asia in February (Fig. S7.A) during a short period of three days (Fig. S7.B). Other genomes from the asymptomatic group belong to other OTUs and were isolated in different months and different continents (Fig. S7.B). Thus, we currently require more robust information to obtain a better-defined distribution of asymptomatic cases.

Description of the most frequent mutations

C241T

The C241T mutation is present in the 5` UTR region. In coronaviruses 5`UTR region is important for viral transcription[11] and packaging[12]. Computational analysis showed that this mutation could create a TAR DNA-binding protein 43 (TDP43) binding site[13], TDP43 is a well-characterized RNA-binding protein that recognize UG-rich nucleic acids[14] described to regulate splicing of pre-

mRNA, mRNA stability and turnover, mRNA trafficking and can also function as a transcriptional repressor and protect mRNAs under conditions of stress[15]. Experimental studies are necessary to confirm different binding constants of TDP43 for the two variants of 5`UTR and its *in vivo* effects.

C1059T

Mutation C1059T lies on Nsp2. Nsp2 does not have a clearly defined function in SARS-CoV-2 since the deletion of Nsp2 from SARS-CoV has little effect on viral titers and so may be dispensable for viral replication[16]. However, Nsp2 from SARS-CoV can interact with prohibitin 1 and 2 (PBH1 and PBH2)[17], two proteins involved in several cellular functions including cell cycle progression[18], cell migration[19], cellular differentiation[20], apoptosis[21] and mitochondrial biogenesis[22].

C3037T and T8782C

Mutations C3037T and T8782C are synonymous mutations in Nsp3 and Nsp4 respectively, therefore, is more difficult to associate these changes to an evolutionary advantage for the virus. These two mutations occurred in the third position of a codon, one possibility is that this changes the frequency of codon usage in humans increasing expression or any other of the related effects caused by synonymous codon change (some of them reviewed[23]).

C3037T causes a codon change from TTC to TTT. TTT is more frequently present in the genome of SARS-CoV-2 and other related coronaviruses compared to TTC[24] but in humans the codon usage of TTT and TTC are similar[23]. The reason why TTT is more frequent in SARS-CoV-2 is unknown but seems that is a selection related to SARS-CoV-2 and not by the host.

For T8782C, the change in codon is from AGT to AGC and opposite to C3037T, the mutated codon (AGC) is less frequent than AGT in the SARS-CoV-2 genome[24] but AGC is more frequent than AGT in human genomes[23]. In this case, seems that the change is related to an adaptation to human host.

C14408T

The C14408T mutation changes P323 to leucine in Nsp12, the RNA-dependent RNA polymerase of SARS-CoV2 (Fig. 5.A and B). P323, along with P322 end helix 10 and generate a turn preceding a beta-sheet (Fig. 5.C). Leucine at position 323 could form hydrophobic interactions with methyl group of L324 and the aromatic ring of F396 creating a more stable variant of Nsp12 (Fig. 5.E). Protein dynamics simulations showed an increase in stability of the Nsp12 P323L variant[25]. In absence of P322, the mutation P323L would probably be disfavored due to flexibilization of the turn in the end of helix 10. Experimental evidence is necessary to confirm these hypotheses and to evaluate its impact on protein function.

A23403G

An interesting protein to track is spike protein (Fig. 6.A) due to its importance in SARS-CoV-2 infectivity. It has been suggested that the D614G change in the S1 domain that results from the A23403G mutation generates a more infectious virus, less spike shedding, greater incorporation in pseudovirions[26] and higher viral load[7].

How these effects occur at the structural level remains unclear, although some hypotheses have been put forward: 1) We think that there is no evidence for hydrogen-bond between D614 and T859 mentioned by Korber et al. 2020, distances between D614 and T859 are too long for an

hydrogen bond (Fig 6.B), 2) distances between Q613 and T859 (Fig. 6.C) could be reduced by increased flexibility due to D614G substitution, forming a stabilizing hydrogen bond, 3) currently available structures do not show salt-bridges between D614 and R646 as proposed by Zhang et al. 2020 (Fig. 6.D).

G25563T

Orf3a (Fig. 7.A) is required for efficient *in vitro* and *in vivo* replication in SARS-CoV[27], has been implicated in inflammasome activation[28], apoptosis[29], necrotic cell death[30] and has been observed in Golgi membranes[31] where pH is slightly acidic[32]. Kern et al. 2020 showed that Orf3a preferentially transports Ca^{+2} or K^{+} ions through a pore (Fig 7.B) of in which one constriction is formed by the side chain of Q57 (Fig.7.C).

Mutation G25563T produce a Q57H variant of Orf3a (Fig. 7.C) that did not show significant differences in expression, stability, conductance, selectivity or gating behavior[8]. We modelled Q57H mutation and we did not observe differences in the radius of constriction (Fig. 7.C) formed by aminoacid 57 but we observed slightly differences in the electrostatic surface due the ionizability of histidine side chain (Fig. 7.D).

C28144T

Orf8 is not well described in the literature and we cannot extrapolate results from SARS-CoV because most of SARS-CoV genomes present a 29-nucleotide deletion that divides Orf8 into Orf8a and Orf8b[33]. Recently, it was showed that Orf8a can directly interact with MHC-1 molecules and reduce their surface expression[34]. Also, they showed that in Orf8-expressing cells MHC-1 molecules are target for lysosomal degradation[34]. However, they observed the same downregulating effect for S84 (C28144) and L84 (T28144) Orf8 variants[34].

G28881A, G28882A, G28883C

N protein is formed by two domains and three disordered regions. The central disordered region named LKR was shown to interact directly with RNA[35] and other proteins[36], probably through positive side chains; also, this region contains phosphorylation sites able to modulate the oligomerization of N protein[37].

Mutation G28883C that introduces an arginine at position 204 contributes one more positive charge to each N protein. Mutations G28881A and G28882A produce a change from arginine to lysine, these two positive amino acids probably have low impact in the overall electrostatic distribution of N protein. However, change from R to K in this position could change the probability of phosphorylation in S202 or T205. Using the program NetPhoK[38], we observed different phosphorylation potential in S202 and T205 between G28881-G28882-G28883 (RG) and A28881-A28882-C28883 (KR) (Fig. S8)

CONCLUDING REMARKS:

Here, we present a complete geographical and temporal worldwide distribution of SARS-CoV-2 haplotypes during the first five months of the pandemic. We identified 11 high frequency mutations. These important variations (asserted mainly by their frequencies) need to be tracked during the pandemic.

Our proposed classification system, showed to be phylogenetically consistent, allows us to easily monitor the spatial and temporal changes of these mutations in a worldwide context. This was

only possible due the unprecedented worldwide efforts in genome sequencing of SARS-CoV-2 and the public databases that rapidly share the information.

In the next months, this classification system will need to be updated, identification of new haplotypes could be performed by combining identification of new frequent mutations (probably using less percentage cut-off than here, for example 1 %) and phylogenetic analysis. This classification can be useful to track SARS-CoV-2 evolution and dissemination of medical important mutations.

Our geographical analysis results showed a differential distribution pattern of OTUs in each continent. This could be due by different adaptations process in each continent, by evolutive characteristics and/or by lockdown policies. Temporal analysis showed an expected pattern of virus adaptation.

Our weak conclusion related to the patient status information is due the poorly organized metadata publicly available. Thus, we highlight the importance of correct management and organization genomic metadata.

Finally, although more studies need to be performed to increase our knowledge of the biology of SARS-CoV-2, we were able to make hypotheses about the possible effects of the most frequent mutations identified. This will help in the development of new studies that will impact vaccine development, diagnostic tests creation, among others.

MATERIAL AND METHODS:

Mutation frequency analysis:

To perform the mutation frequency analysis we first downloaded genomes from GISAID database (as of June 9th 2020) with more than 29000 nt, less than 1 % Ns, less than 0.05 % unique aminoacid mutations (not seen in other sequences in database), no insertion/deletion unless verified by submitter and with host description human or environment (27542 genomes downloaded). After that, we filtered genomes with any ambiguity using a python script (19278 genomes filtered). This set of genomes (8264 genomes) was aligned using MAFFT with FFT-NS-2 strategy and default parameter settings[39]. Finally, we manually removed genomes without sequence information from nt 203 to nt 29674 respect to the reference genome (EPI_ISL_406801), with a substitution in the stop codon of ORF10, with frameshifts in one or more ORFs, with any insertion and with 1, 2, 4 or more than 9 deletions. A list of all the manually removed genomes can be found in supplementary table 1. The final 7848 genomes ("high-quality" genomes data set) from nt 203 to 29674 (respect to the reference sequence) were aligned using MAFFT with FFT-NS-2 strategy and default parameter settings[39]. This alignment was used to calculate identity percentages based in the reference sequence and nucleotide frequency in each nucleotide genome position.

Phylogenetic tree construction:

Using the last alignment described in the previous section we estimated a maximum likelihood tree with IQ-TREE 2[40] using the GTR+F+R2 model of nucleotide substitution[41,42,43], default heuristic search options, ultrafast bootstrapping with 1000 replicates[44] and the reference genome as the outgroup.

OTUs determination:

Alignment of the identified mutations with 10 % or more frequency in our 7848 genomes dataset were associated with the whole-genome rooted tree using the MSA function from the ggtree package[45,46] in R, and visually examined to identify the major haplotypes based in these positions, designated as OTUs (Operational Taxonomic Units).

Analysis of OTUs geographical distribution:

To perform these analysis we align using MAFFT with FFT-N-2 strategy[39] the 27542 genomes with more than 29000 nt, less than 1 % Ns, less than 0.05 % unique aminoacid mutations (not seen in other sequences in database), no insertion/deletion unless verified by submitter and with host description human or environment from GISAID (as of June 9th 2020). After that, we randomly select genomes and grouped them in six samples with 4590 genomes each. Genomes in each sample were divided in continents and each continent in months. For each continent we used the months with 25 or more total genomes in that month in that continent. We calculated the percent of each OTU in each month analyzed. Then, we multiplied each OTU percent by the number of cases reported in the analyzed continent and month (number of cases by continent were obtained from European Centre for Disease Prevention and Control: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide) (To count the number of cases in each continent in each month we just used the data of the countries with at least one genome in our set of 27542 genomes) An example of the process of this analysis is showed in figure S9. Finally, we added by OTUs the contribution of each month and averaged the results of the six samples. Statistical difference between OTUs were calculated using the package "ggpubr" in R with the non-parametric Kruskal-Wallis test, and pairwise statistical differences were calculated using non-parametric Wilcoxon test from the same R package. Also, we plot the continent information of each genome used to estimate the maximum likelihood tree to generate a tree figure with this information using ggtree package in R[45,46].

Analysis of OTUs temporal distribution:

The same samples from the previous step, were now divided by months and each month by continents. For each month we used the continents with 25 or more total genomes in that continent in that month. The percentages of OTUs were calculated in each continent analyzed. After that, we multiplied each OTU percent by the number of cases reported in the analyzed month and continent (cases were obtained in the same form that in the previous section). Finally, we added by OTUs the contribution of each continent and averaged the results of the six samples. Final values were plotted and analyzed statistically as described in the previous section. Also, we plot the collection date information of each genome used to estimate the maximum likelihood tree to generate a tree figure with this information using ggtree package in R[45,46].

Analysis of patient status with OTUs:

2091 genomes with patient status information was downloaded from GISAID database. Genomes with ambiguities in any of the positions necessary to classify the genomes in OTUs were filtered (supplementary table 2 contain this information). The 2049 genomes remained were aligned using MAFFT with FFT-NS-2 strategy and default parameter settings[39]. Using this alingment we estimated a maximum likelihood tree with IQ-TREE 2[40] using the GTR+F+R5 model of nucleotide substitution[41,42,43], default heuristic search options, ultrafast bootstrapping with 1000 replicates[44]. Patient status information from GISAID was recategorized in four disease levels: No Informative, Asymptomatic, Mild and Severe. A table showing the GISAID patient status categorize comprising our categories can be found in supplementary table 3. We plot the

information of these categories to our maximum likelihood tree of these genomes using ggtree package in R[45,46], and calculate the frequency of OTUs in each patient status category.

DATA AVAILABILITY:

The data that support the findings of this study are available on request from the corresponding author upon reasonable request.

REFERENCES:

1. Cuccinotta D and Vanelli M. 2020. WHO declares COVID-19 a pandemic. Vol. 91, 157-160. Acta Biomedica.
2. WHO. 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019 Retrieved on 02 July 2020.
3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao G, Tan W. 2020. A novel coronavirus from patients with pneumonia in China, 2019. Vol. 382(8), 727-733. The New England Journal of Medicine.
4. Shu Y and McCauley J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Vol. 22(13), 1-3. Euro Surveillance.
5. Tang X, Wi C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. Vol. 7, 1012-1023. National Science Review.
6. Rambaut A, Holmes E, Hill V, O`Toole A, McCrone A, Ruis C, du Plessis L, Pybus O. 2020. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv preprint doi: https://doi.org/10.1101/2020.04.17.046086.
7. Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi E, Bhattacharya T, Foley B, Hastie K, Parker M, Partridge D, Evans C, Freeman T, de Silva T, McDanal C, Perez L, Tang H, Moon-Walker A, Whelan S, LaBranche C, Saphire E, Montefiori D. 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. https://doi.org/10.1016/j.cell.2020.06.043. Cell.
8. Kern D, Sorum B, Hoel C, Sridharan S, Remis J, Toso D, Brohawn S. 2020. Cryo-EM structure of the SARS-CoV-2 3a ion channel in lipid nanodiscs. bioRxiv preprint doi: https://doi.org/10.1101/2020.06.17.156554.
9. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo R, Zella D, Ippodrino R. 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Vol. 18(179), 1-9. Journal of Translational Medicine.
10. Yin C. 2020. Genotyping coronavirus SARS-CoV-2: methods and implication. Genomics. https://doi.org/10.1016/j.ygeno.2020.04.016
11. Madhugiri R, Fricke M, Marz M, Ziebuhr J. 2014. RNA structure analysis of alphacoronavirus terminal genome regions. Vol. 194, 76-89. Virus Research.
12. Masters P. 2019. Coronavirus genomic RNA packaging. Vol. 537, 198-207. Virology.
13. Mukherjee M and Goswami S. 2020. Global cataloguing of variations in untranslated regions of viral genome and prediction of key host RNA binding protein-microRNA interactions modulating genome stability in SARS-CoV-2. bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.134585

14. Kuo P, Chiang C, Wang Y, Doudeva L, Yuan H. 2014. The crystal structure of TDP-43 RRM1-DNA complex reveals the specific recognition for UG- and TG-rich nucleic acids. Vol. 42(7), 4712-4722. Nucleic Acids Research.
15. Lee E, Lee V, Trojanowski J. 2011. Gains or losses: molecular mechanisms of TDP43-mediated neurodegeneration. Vol. 13(1), 38-50. Nature Reviews Neuroscience.
16. Graham R, Sims A, Brockway S, Baric S, Denison M. 2005. The nsp2 replicase proteins of murine hepatitis virus and severe acute respirator syndrome coronavirus are dispensable for viral replication. Vol. 79(21), 13399-13411. Journal of Virology.
17. Cornillez-Ty C, Liao L, Yates J, Kuhn P, Buchmeier M. 2009. Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling. Vol. 83(19), 10314-10318. Journal of Virology.
18. Wang S, Nath N, Adlam M, Chellappan S. 1999. Prohibitin, a potential tumor suppressor, interacts with RB and regulates E2F function. Vol. 18, 3501-3510. Oncogene.
19. Rajalingam K, Wunder C, Brinkmann V, Churin Y, Hekman M, Sievers C, Rapp U, Rudel T. 2005. Prohibitin is required for RAS-induced RAF-MEK-ERK activation and epithelial cell migration. Vol. 7(8), 837-843. Nature Cell Biology.
20. Sun L, Liu L, Yang X, Wu Z. 2004. Akt binds prohibitin 2 and relieves its repression of MyoD and muscle differentiation. Vol. 117(14), 3021-3029. Journal of Cell Science
21. Fusaro G, Dasgupta P, Rastogi S, Joshi B, Chellappan S. 2003. Prohibitin induces the transcriptional activity of p53 and is exported from the nucleus upon apoptotic signaling. Vol. 278(48), 47853-47861. The Journal of Biological Chemistry.
22. Merkwirth C and Langer T. 2008. Prohibitin function within mitochondria: essential roles for cell proliferation and cristae morphogenesis. Vol. 1793, 27-32. Biochimica et Biophysica Acta.
23. Mauro V and Chapel S. 2014. A critical analysis of codon optimization in human therapeutics. Vol. 20(11), 604-613. Trends in Molecular Medicine.
24. Gu H, Chu D Peiris M, Poon L. 2020. Multivariate Analyses of Codon Usage of SARS-CoV-2 and other betacoronaviruses. bioRxiv preprint doi: https://doi.org/10.1101/2020.02.15.950568.
25. Chand G and Azad G. 2020. Identification of novel mutations in RNA-dependent RNA ploymerases of SARS-CoV-2 and their implications. bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.079939.
26. Zhang L, Jackson C, Mou H, Ojha A, Rangarajan E, Izard T, Farzan M, Choe H. 2020. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv preprint doi: https://doi.org/10.1101/2020.06.12.148726.
27. Castaño-Rodriguez C, Honrubia J, Gutierrez-Alvarez J, DeDiego M, Nieto-Torres J, Jimenez-Guardeño J,, Regla-Nava J, Fernandez-Delgado R, Verdia-Báguena C, Queralt-Martín M, Kochan G, Perlman S, Aguilella V, Sola I, Enjuanes L. 2018. Role of severe acute respiratory syndrome coronavirus viroporins E, 3a, and 8a in replication and pathogenesis. Vol. 9(3), 1-23. American Society for Microbiology.
28. Siu K, Yuen K, Castaño-Rodriguez C, Ye Z, Yeung M, Fung S, Yuan S, Chan C, Yuen K, Enjuanes L, Jin D. 2019. Severe acute respiratory syndrome coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. Vol. 33. 8865-8877. The FASEB Journal.
29. Chan C, Tsoi H, Chan W, Zhai S, Wong C, Yao X, Chan W, Tsui S, Chan H. 2009. The ion channel activity of the SARS-coronavirus 3a protein is linked to its pro-apoptotic function. Vol. 41, 2232-2239. The International Journal of Biochemistry and Cell Biology

30. Yue Y, Nabar N, Shi C, Kamenyeva O, Xiao X, Hwang I, Wang M, Kehrl J. 2018. SARS-Coronavirus open reading frame-3a drives multimodal necrotic cell death. Vol. 9, 1-15. Cell Death and Disease.
31. Padhan K, Tanwar C, Hussain A, Hui P, Lee M, Cheung C, Malik J, Jameel S. 2007. Severe acute respiratory syndrome coronavirus Orf3a protein interacts with caveolin. Vol. 88, 3067-3077. Journal of General Virology.
32. Griffiths G and Simons K. 1986. The trans Golgi network: sorting at the exit site of the golgi complex. Vol. 234, 438-443. Science.
33. Lau S, Fen Y, Cheng H, Luk H, Yang W, Li K, Zhang Y, Huang Y, Song Z, Chow W, Fan R, Ahmed S, Yeung H, Lam C, Cai J, Wong Chan J, Yuen K, Zhang H, Woo P. 2015. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. Vol. 89(20), 10532-10547. Journal of Virology.
34. Zhang Y, Zhang J, Chen Y, Luo B, Yuan Y, Huang F, Yang T, Yu F, Liu J, Liu B, Song Z, Chen J, Pan T, Zhang X, Li Y, Li R, Huang W, Xiao F, Zhang H. 2020. The Orf8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-1. bioRxiv preprint doi: https://doi.org/10.1101/2020.05.24.111823.
35. Chang C, Hsu Y, Chang Y, Chao F, Wu M, Huang Y, Hu C, Huang T. 2009. Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein implications for ribonucleocapsid protein packaging. Vol. 83(5), 2255-2264. Journal of Virology.
36. Luo H, Chen Q, Chen J, Chen K, Shen X, Jiang H. 2005. The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. Vol. 579, 2623-2628. FEBS letters.
37. Chang C, Chen C, Chiang M, Hsu Y, Huang T. 2013. Transient oligomerization of the SARS-CoV N protein – Implication for virus ribonucleoprotein packaging. Vol. 8(5), e65045. PlosONE.
38. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. 2004. Prediction of post-translationalo glycosylation and phophorylation of proteins from the aminoacid sequence. Vol. 4, 1633-1649. Proteomics.
39. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Vol. 30(14), 3059-3066. Nucleic Acids Research.
40. Minh B, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Vol. 37, 1530-1534. Molecular Biology and Evolution.
41. Tavaré S. 1986. Some mathematical questions in biology: DNA sequence analysis. Lectures on mathematics in the life sciences. Vol. 17, 57-86. American Mathematical society.
42. Soubrier J, Steel M, Lee M, Sarkissian C, Guindon S, Ho S, Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Vol. 29(11), 3345-3358. Molecular Biology and Evolution.
43. Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Vol. 139, 993-1005. Genetics
44. Hoang D, Chernomor O, Haeseler A, Minh B, Vinh L. 2017. UFBoot2: Improving the ultrafast bootstrap approximation. Vol. 35(2), 518-522. Molecular Biology and Evolution.

45. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Vol. 69, 1-18- Current Protocols in Bioinformatics.
46. Yu G, Smith D, Zhu H, Guan Y, Lam T. 2017. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Vol. 8, 28-36. Methods in Ecology and Evolution.

Acknowledgements:

We thank Professor Shaker Chuck Farah (Institute of Chemistry – University of Sao Paulo) for English writing corrections and helpful comments. Also, we thank Professors Aline Maria da Silva (Institute of Chemistry – University of Sao Paulo), Joao Renato Rebello Pinho (Albert Einstein Hospital – Sao Paulo) and PhD(c). Deyvid Amgarten (Albert Einstein Hospital – Sao Paulo) for its helpful comments. To the Ricardo Palma University High-Performance Computational Cluster (URPHPC) managers Gustavo Adolfo Abarca Valdiviezo and Roxana Paola Mier Hermoza at the Ricardo Palma Informatic Department (OFICIC) for their contribution in programs and remote use configuration of URPHPC.

Author contributions:

S.J.A., D.Z.S., C.H.R., G.L.B., A.C.C. and R.G.C. conceived, initiated and coordinated the project. S.J.A. performed the phylogenetic analyses, geographical and temporal analyses. G.U.C. wrote python scripts used in data processing and analyses. S.J.A., D.Z.S., C.H.R., G.L.B., A.C.C., R.G.C., R.P.C. performed the structural analysis. Manuscript was written by S.J.A., D.Z.S., C.H.R., G.L.B. All authors discussed the methodologies, results, read and approved the manuscript.

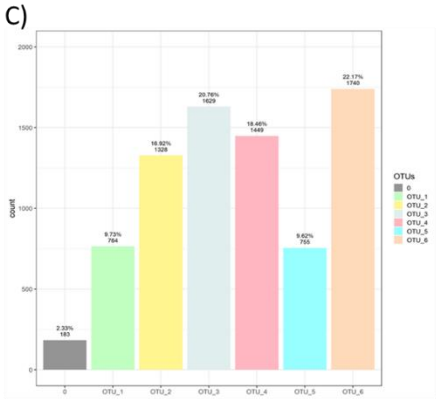
Competing interests:

The authors declare no competing interests

A)

Position	OTU						Region	AA change	Identity*
	1	2	3	4	5	6			
241	C	C	T	T	T	T	5' UTR	----	0.276
1059	C	C	C	C	C	T	Nsp2	T85I	0.773
3037	C	C	T	T	T	T	Nsp3	Syn	0.281
8782	T	C	C	C	C	C	Nsp4	Syn	0.100
14408	C	C	T	T	T	T	Nsp12	P323L	0.278
23403	A	A	G	G	G	G	S	D614G	0.275
25563	G	G	G	G	T	T	Orf3a	Q57H	0.676
28144	C	T	T	T	T	T	Orf8	S84L	0.103
28881	G	G	G	A	G	G	N	R203K	0.813
28882	G	G	G	A	G	G	N	R203K	0.813
28883	G	G	G	C	G	G	N	G204R	0.813

C)



B)

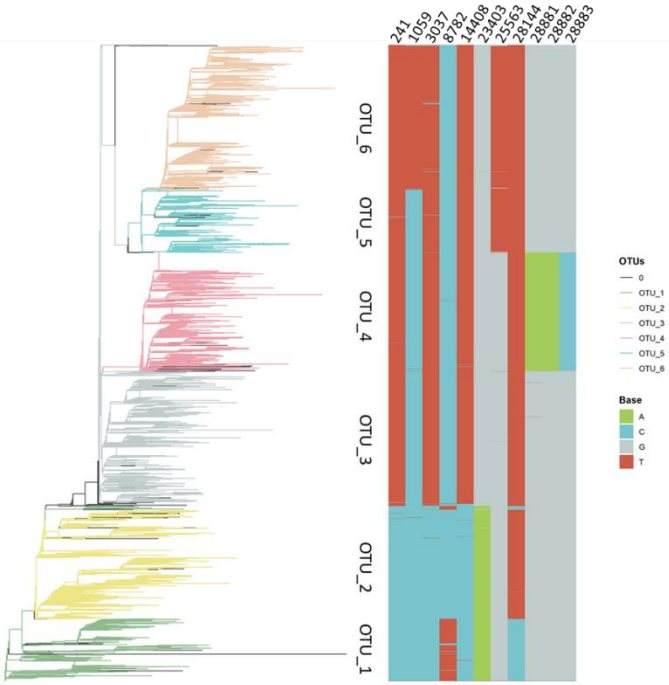


Figure 1. Six haplotypes based in eleven positions can classify more than 97 % of the genomes. A) Rooted tree of 7848 SARS-CoV-2 complete genomes associated to an alingment of eleven genomic positions (241, 1059, 3037, 8782, 14408, 23403, 25563, 28144, 28881, 28882, 28883) showing well correlation between haplotypes (OTUs) based in these eleven positions. Tips of the tree where colored based in the OTU. B) Bar diagram shows the OTUs distribution of the 7848 genomes (0 correspond to unclassified genomes). C) Table showing the haplotype of each OTU.

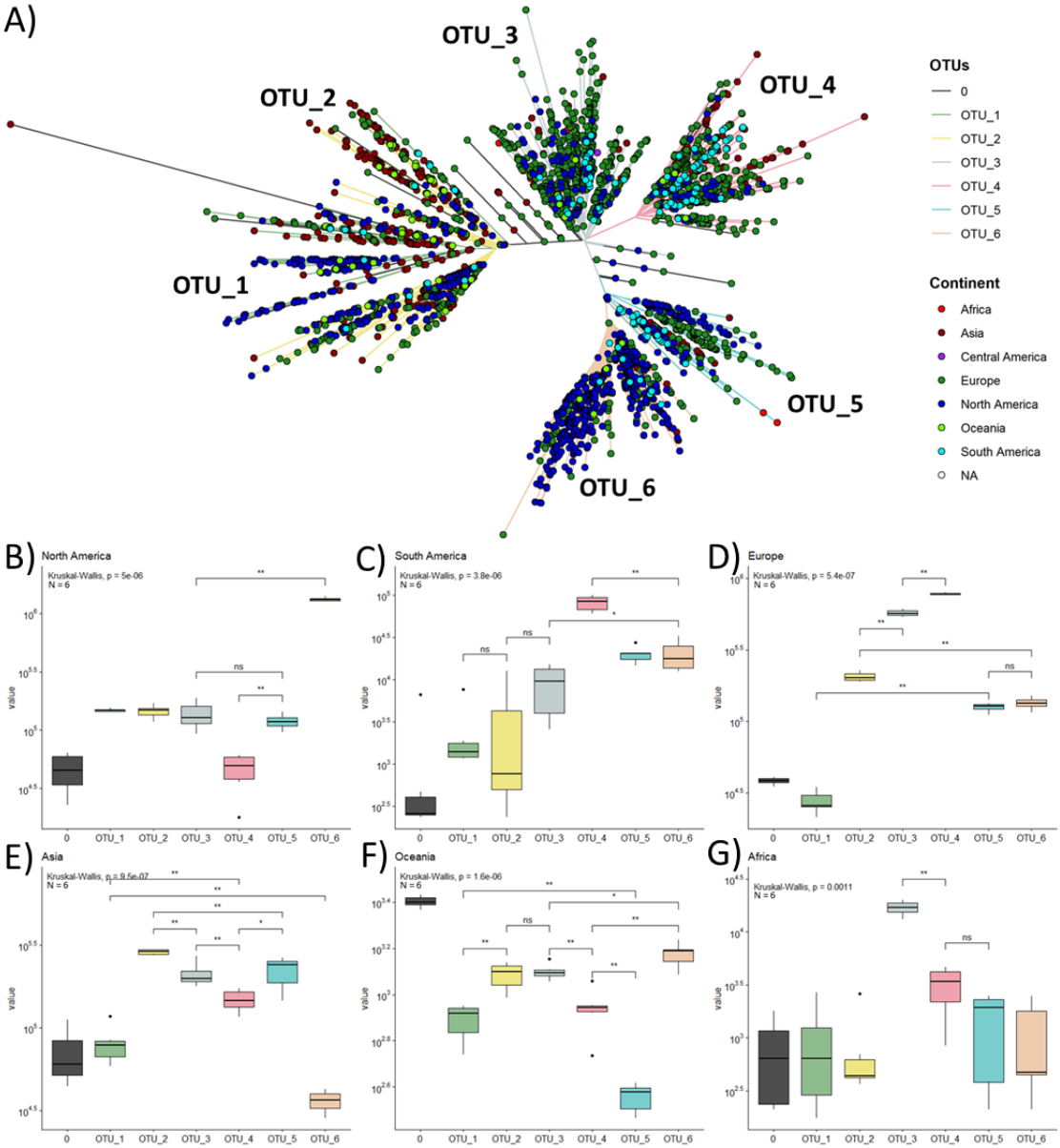


Figure 2. Differential distribution of genotypes showed by worldwide geographic analysis. A) Unrooted tree of the 7848 genomes, tips were colored according to OTUs and points in each tip were colored according to continent. B-G) Boxplots of OTUs distribution in each continent (B, North America; C, South America; D, Europe; E, Asia; F, Oceania; G, Africa).

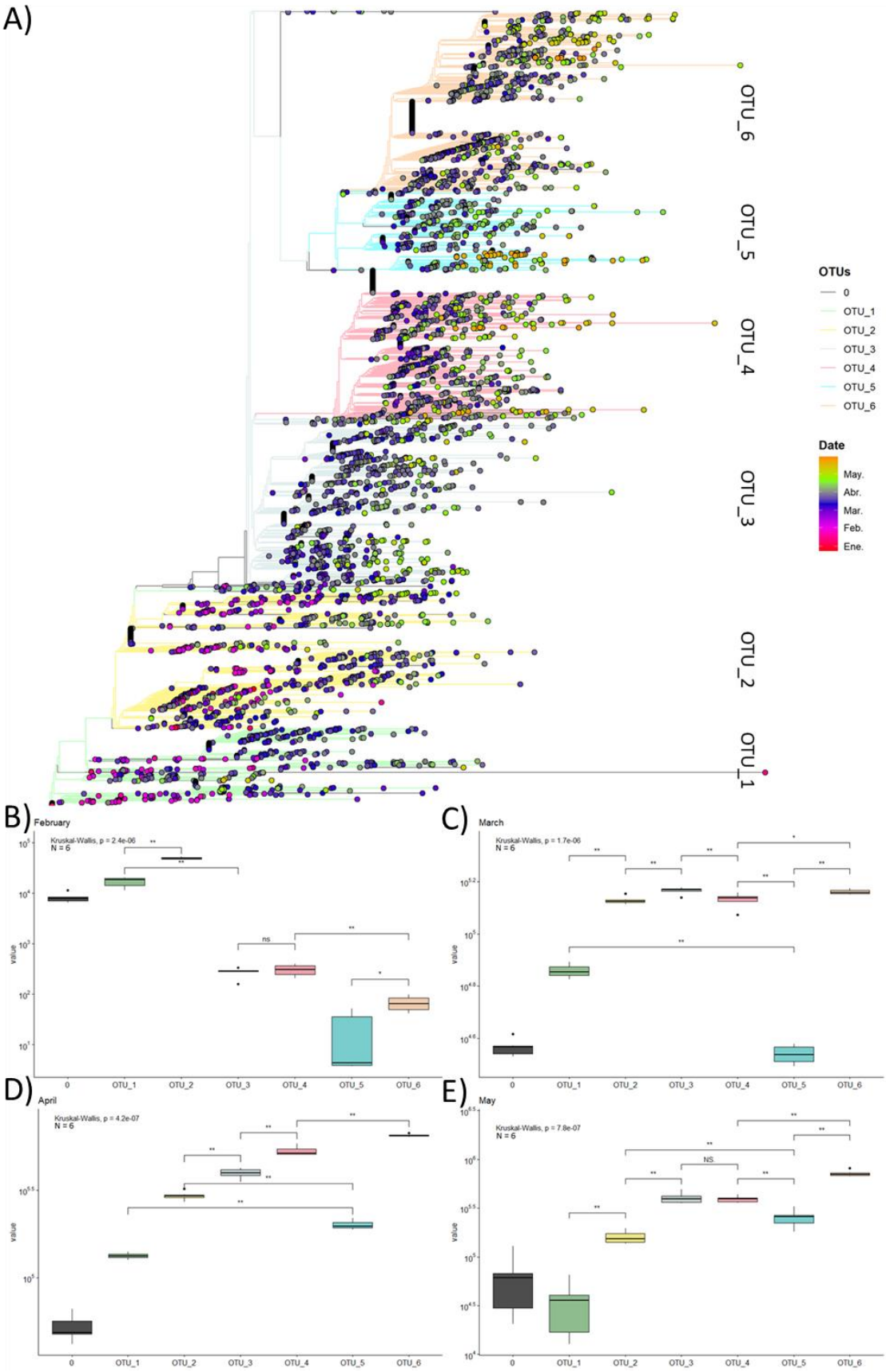

A)
OTU_6
OTU_5
OTU_4
OTU_3
OTU_2
OTU_1
OTUs
0
OTU_1
OTU_2
OTU_3
OTU_4
OTU_5
OTU_6
Date
May.
Abr.
Mar.
Feb.
Ene.
B)
February
Kruskal-Wallis, p = 2.4e-06
N = 6
C)
March
Kruskal-Wallis, p = 1.7e-06
N = 6
D)
April
Kruskal-Wallis, p = 4.2e-07
N = 6
E)
May
Kruskal-Wallis, p = 7.8e-07
N = 6
value
0
OTU_1
OTU_2
OTU_3
OTU_4
OTU_5
OTU_6

Figure 3. Evidence of evolution in SARS-CoV-2 showed by worldwide temporal distribution analysis. A) Rooted tree of the 7848 genomes showing temporal distribution. Tips were colored by OTUs and point in each tip were colored according to the isolation date of the

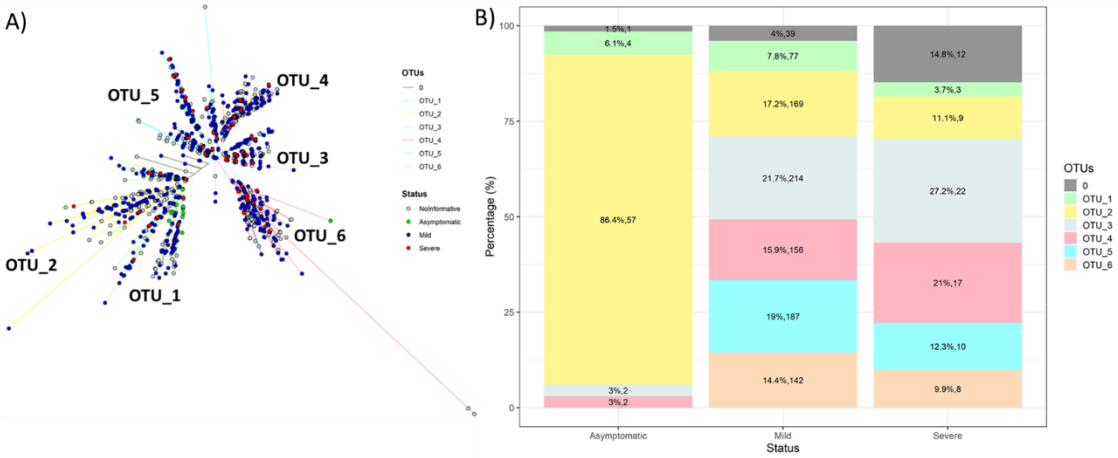


Figure 4. Scarce number of genomes with patient status avoid robust conclusions. A) Unrooted tree of 2050 genomes with patient status information in GISAID database. Tips were colored by OTUs and points in each tip were colored by Status. B) Percentages by OTUs by patient status categories.

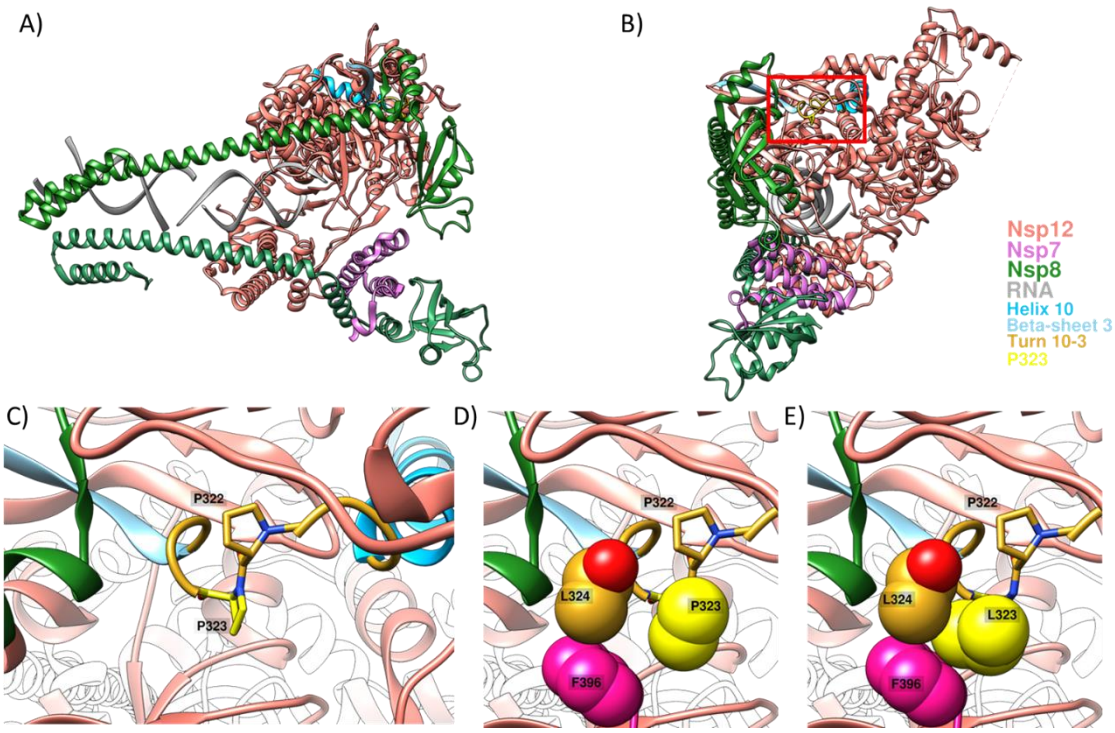


Figure 5. P323L could impact stability of Nsp12 without disturbing its overall structure. A) Structure of RNA-dependent RNA polymerase complex (PDB ID: 6YYT). Chains (Nsp12, Nsp7, Nsp8, RNA) are distinguished by colors. Helix 10, Beta-sheet 3, Turn 10-3 and P323 also are differentially colored. B) Structure in A rotated 90 degrees. C) Zoom of the red box in B showed P322 and P323 in the center of Turn 10-3. D) Turn 10-3 with side chains of P323, L324 and F396 in sphere representation to highlight the distance between side chains of P323 and L324. E) P323 in D was computationally replaced by L323. Now, distances between methyl group of leucine are shorter with L323.

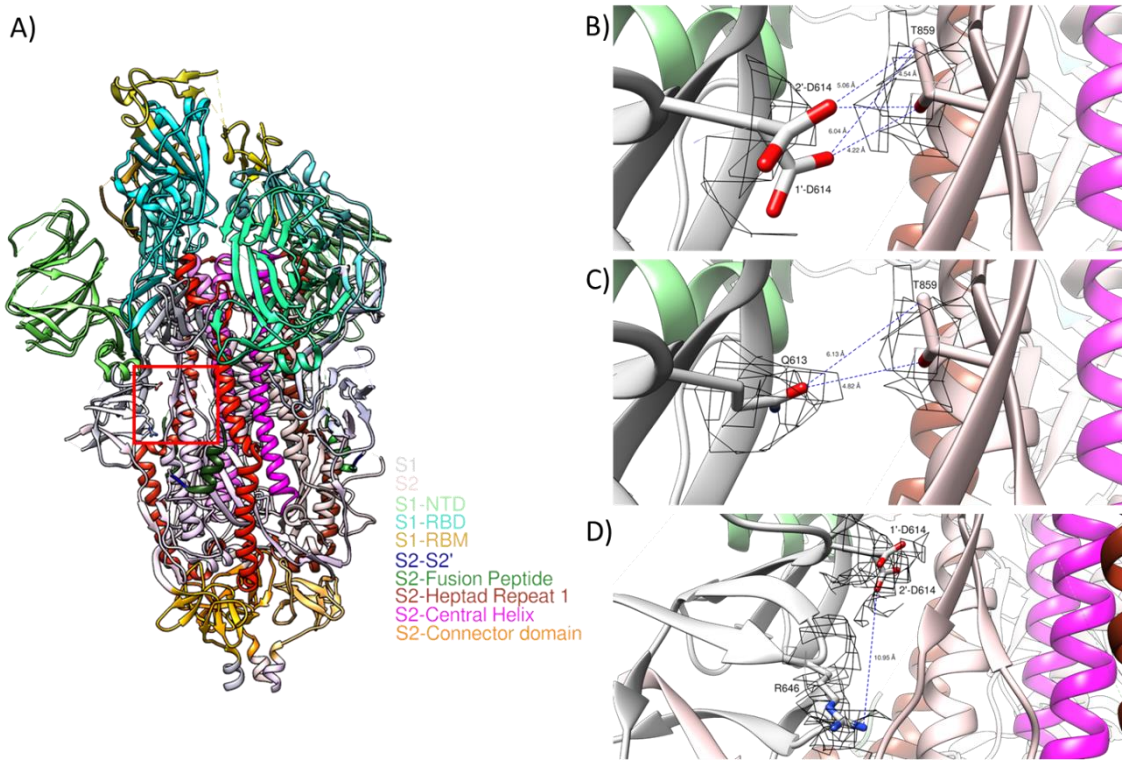


Figure 6. Structural hypotheses about D614G mutation in Spike protein. A) Structure of the open state of Spike trimer (PDB ID: 6YVB) colored by domains. B) Distances between side chains of two possible rotamers of D614 (1`-D614 and 2`-D614) and T859. Except for 1`-D614 and carbonyl group of T859, the other distances seems to be large to form a hydrogen bond. C) Distances between side chains Q613 and T859. These distances are also large to form hydrogen bonds. D) R646 point to the opposite side of D614 showing that there is no salt bridge. B,C and D shows electron density maps of the side chains of the labeled residues.

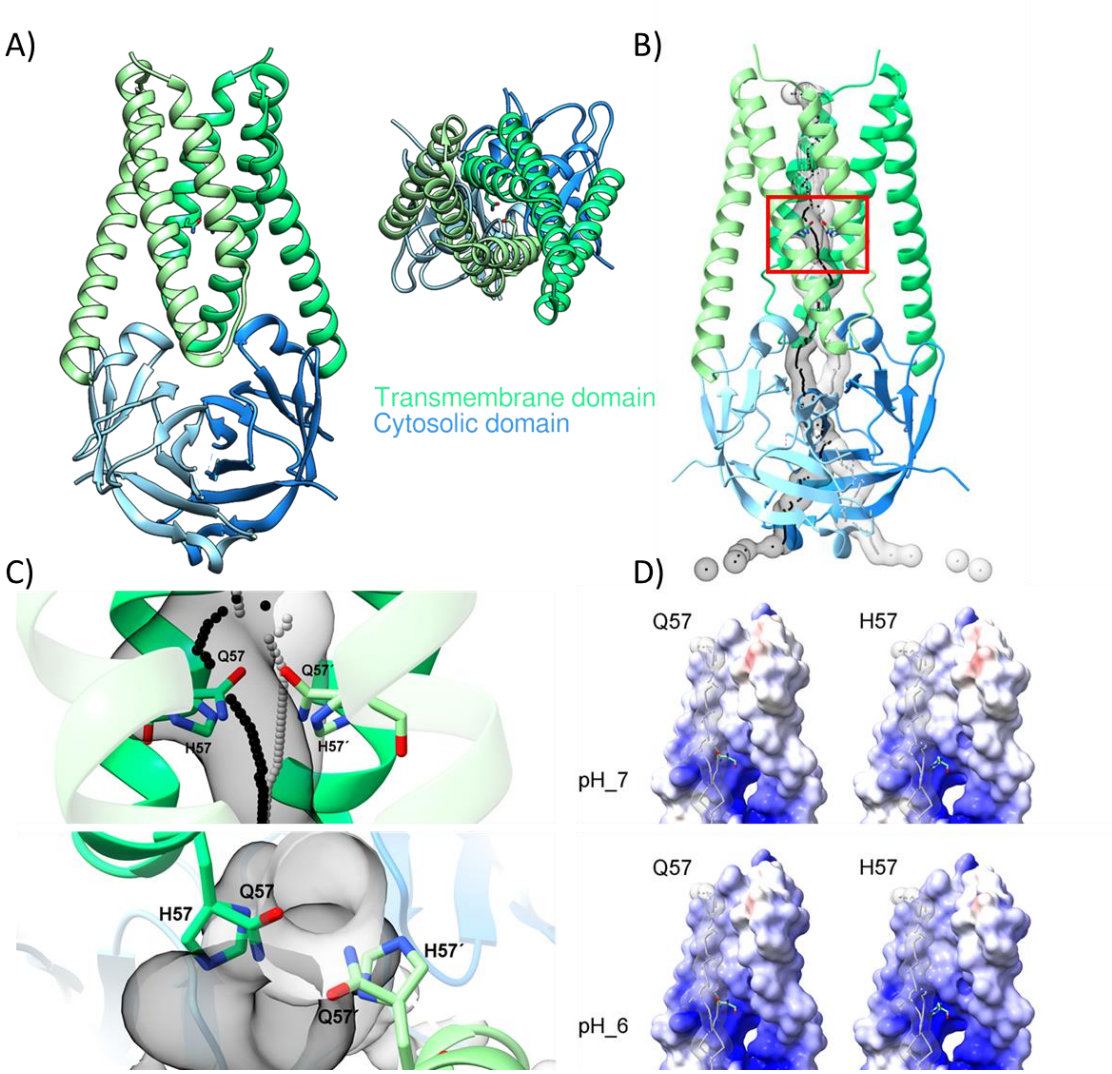


Figure 7. Orf3a Q57H apparently does not modify pore constriction distances but electrostatics distribution. A) Structure of the Orf3a dimer (PDB ID: 6XDC) colored by domains. Right of A shows the same structure but in an upper view. B) Orf3a showing the central pore, in red box the section corresponding to the fifth pore constriction. C) zoom of the red box in B, above we showed Q and H variants superpossed. Below we show a transversal cut of the pore near to the fifth. Pore radius in two variants are similar. D) Electrostatic surface maps of Q57 and H57 variants in two different pHs (7 and 6). Residues Q57 and H57 are shown in stick representations to point the fifth constriction. We show a slighlty more positive region at the height of the fifth constriction.