Global Geographic and Temporal Analysis of SARS-CoV-2 Haplotypes Normalized by COVID-19 Cases during the First Seven Months of the Pandemic

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, a large number of genomes have been sequenced with unprecedented speed around the world. This marks a unique opportunity to analyze virus spreading and evolution in a worldwide context. However, currently, there is not a useful haplotype description to help to track important and globally scattered mutations. Also, differences in the number of sequenced genomes between countries and/or months make it difficult to identify the emergence of haplotypes in regions where few genomes are sequenced but a large number of cases are reported. We proposed an approach based on the normalization by COVID-19 cases of relative frequencies of mutations using all the available data to identify major haplotypes. Thus, we can use a similar normalization approach to tracking the global temporal and geographic haplotypes distribution in the world. Using 48 776 genomes, we identify 5 major haplotypes based on 9 high-frequency mutations. Normalized global geographic and temporal analysis is presented here highlighting the current importance of nucleocapsid mutations (R203K, G204R) above the highly discussed D614G in spike protein. Also, we analyzed age, gender, and patient status distribution by haplotypes, but scarce and not well-organized information about this is publicly available. For that, we create a web-service to continuously update our normalized analysis of mutations and haplotypes, and to allow researchers to voluntarily share patient status information in a well-organized manner to improve analyses and making possible monitor the emergence of mutations and/or haplotypes with patients preferences or different pathogenic features. Finally, we discuss currently structural and functional hypotheses in the most frequently identified mutations.

INTRODUCTION:

COVID-19 was declared a pandemic by the World Health Organization on March 11th 2020[1], with around 23 million cases and 800 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 July 30, 2020, the GISAID database[4] contained 48 776 genomes with at least 29 000 sequenced bases.

At the moment, some analysis has 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].

To complement these current classification systems, we believe that haplotypes description and nomenclature could help to better track important mutations that are currently circulating in the world. Identification of SARS-CoV-2 haplotypes aids in understanding the evolution of the virus and may improve our efforts to control the disease.

To perform a reasonable analysis of the worldwide temporal and geographical distribution of SARS-CoV-2 haplotypes, we need to take into account the differences in the number of sequenced genomes in months and countries-continents. Thus, we first used a data set of 48776 complete genomes to estimate the worldwide relative frequency of nucleotides in each SARS-CoV-2 genomic position and found nine positions with normalized relative frequencies (NRF_p) greater than 0.1 and lesser than 0.9. After that, using a total of 19486 complete genomes with any ambiguous nucleotide position from GISAID we performed a phylogenetic analysis and correlated the major branches with SARS-CoV-2 variants which can be classified into five haplotypes or Operational Taxonomic Units (OTUs) based on the distribution of the nine identified nucleotide positions in our NRF_p analysis. After that, we analyzed the geographical and temporal worldwide distribution of OTUs normalized by the number of COVID-19 cases. Also, we attempt to correlate these OTUs with patient status, age, and gender information. Finally, we discuss the current hypothesis of the most frequent mutations on protein structure and function.

RESULTS AND DISCUSSION:

Mutations frequency analysis

The GISAID database contains around 48 776 genomes with at least 29 000 sequenced bases and from these 19 486 genomes does not contain any ambiguity (as of July 30th). With an alignment of the 48 776 genomes, we performed a normalized relative frequency analysis of each nucleotide in each genomic position (NRF_p) (see material and methods for details), this normalization was performed to reduce the bias due that the number of sequenced genomes in continents and months are not correlated with the number of cases in these continents and months. Using these NRF_p analyses, we identified 9 positions with greater than 0.1 and less than 0.9 NRF_p (Fig. 1.A and S1.A) plus many other positions with frequencies between 0.900-0.995 and 0.005-0.100 (Fig. S1.B and S1.C).

The nine most frequent mutations (NRF_p between 0.1 and 0.9) are comprised of seven non-synonymous mutations, one synonymous mutation and one mutation in the 5'UTR region of the SARS-CoV-2 genome (Fig. 1.A). All these mutations have been already identified in other studies[7,8,9,10], although with different frequencies.

OTUs identification

After NRF_p analysis, we estimated a maximum-likelihood tree using the whole-genome alignment of the 19 846 complete genomes without ambiguities. Then, we associated the main branches of the whole-genome tree with an alignment of the 9 positions (241, 1059, 3037, 14408, 23403, 25563, 28881, 28882, 28883) and noted that combinations of those 9 positions

represent 5 well-defined groups in the tree (Fig. 1.B). Using these combinations, we defined 5 haplotypes that allow us to classified 96.5 % of the analyzed genomes (Fig. 1.C), a great part of the remaining not classified genomes are due to the absence of sequencing corresponding to position 241. 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 C241, C3037, C14408, and A23403. This OTU_1 comprised genomes with T or C in position 8782 and C or T in 28144. In other analyses, these mutations divide SARS-CoV-2 strains into two lineages. For instance; at the beginning of the pandemic, Tang et al (2020) shows linkage disequilibrium between those positions and named them as S and L lineages. Rambaut et al (2020) used these positions to discriminate between their proposed major lineages A and B. After, seven months of pandemic NRF_p of T and C in positions 8782 and 28144, respectively, are not in the range of 0.1-0.9, indicating a small quantity of these genomes presented during the pandemic in comparison with other variations.

A SARS-CoV-2 isolated on February 20 was the first belonging to OTU_2 (Fig. S2) that shows simultaneously four mutations different to OTU_1 (C241T, C3037T, C14408T, and A23403G). We can note in the phylogenetic tree that in the transition between the first clade and OTU_2 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_2 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. S2) in N protein was isolated. These three mutations characterize OTU_3. The maximum likelihood tree shows that OTU_4 comes from OTU_2. OTU_4 does not present mutations in N protein, instead, it presents a variation in Orf3a (G25563T). Finally, OTU_5 presents all the mutations of OTU_5 plus one Nsp2 mutation (C1059T).

These 9 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 haplotypes during the pandemic. The fact that we were able to classify 96.5 % of the complete genomes 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. Also, most of the unclassified tips appear within a clade allowing us to easily establish their phylogenetic relationships to a haplotype. Thus, at the moment this system can be of practical use to analyze the geographical and temporal distribution of haplotypes during these seven months of 2020.

It is highly likely 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 circulating haplotypes on a real-time basis need to be proposed. We propose that the best strategy will be to continually monitor the appearance of new haplotypes by tracking mutations that exceed a fixed NRF_p in the world (to allow tracking relevant medical mutations) and associating these mutations to a phylogenomic tree to confirm its phylogenetical relevance, we will perform this task at least one time per month and update this information in our website.

Worldwide geographic distribution of OTUs

Using our OTUs classification, we analyzed the worldwide geographic distribution during the first seven months of 2020. We began by plotting continental information in the unrooted tree of the unambiguous complete genomes (Fig. 2.A) and observed some interesting patterns. For instance, all continents contain all OTUs; however, is relative clearly that most isolates belonging to OTU_5 come from North America (Fig. 2.A). This approach does not allow us to evaluate continents with less sequenced genomes (Fig. S4), such as South America, Oceania, and Africa; also, it is possible that fine differences can be found in the frequency of one OTU with respect to another in each continent. These differences are not observed at this level of analysis.

To better analyze which were the most prevalent OTUs in each continent, we analyzed all the complete genomes in the GISAID database (48 776 genomes). In this analysis, we compared the mean of the frequency of OTUs normalized by cases in each continent of six randomly selected groups of genomes (see material and methods for more details).

This approach more clearly illustrates that OTU_5 was the most prevalent in North America, followed by OTUs 2, 3, the least prevalent were OTU_1 and OTU_4 (Fig. 2.B). First genomes in North America belonged to OTU_1 (Fig. S4). March and April were dominated by OTU_5, but in June OTU_3 seems to have similar counts to OTU_5 (Fig. S4). OTU_5 has 6 of the 9 high-frequency genomic variations described (all except those in N protein) (Fig. 1.A).

South America presents a greater OTU_3 frequency (Fig. 2.C) that was established in April (Fig. S4). Unfortunately, few genomes were sequenced in South America in May, June, and July (44 genomes in total in the three months), hindering a correct analysis of frequencies in these months. Similarly, OTU_3 was most prevalent in Europe, Africa, and Asia (Fig. 2.D, 2.E, and 2.G). Followed by OTU_2 in Europe and Africa, and by OTU_1 in Asia (Fig. 2.D, 2.E, and 2.G). At the haplotype level, OTU_3 present mutations in N protein that apparently increase the fitness of this group in comparison with OTU_3 (OTU_3 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 in N protein.

Oceania presents a more homogeneous distribution of OTUs, with OTU_1 in slightly higher but statistically significant frequency among other OTUs (Fig. 2.F). The analysis of Oceania is in part biased due to the great percentage of genomes without information of position 241 (in the 5` UTR region), hindering unambiguously classification of several sequenced in Australia.

Worldwide temporal distribution of OTUs

A rooted tree was estimated with the 19 846 genomes data set and labeled by date (Fig. 3.A). Here we can clearly follow the evolution beginning with OTU_1 at the base of the tree (mostly labeled with colors that correspond to the first months of the pandemic). Clades, where OTU_2, 4, and 5 are the most prevalent, have intermediate temporal distribution (mostly late February up to late April). OTU_3 has a similar distribution pattern to OTU_2, 4, and 5 but with more representatives isolated in May, June, and July.

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 for details). In this analysis, we did not consider December and January that present genomes just belonging to OTU_1 mainly from Asia (Fig.S4 and S5).

Analysis using the data of February from North America, Europe, and Asia showed that OTU_1 continues as the most prevalent in the world but with the presence of OTU_2, 3, 4, and 5 (Fig.

3.B). Analysis by continents showed that during this month Asia and North America still had higher proportions of OTU_1, but in Europe, a more homogeneous distribution of OTUs 2-5 was observed (Fig. S4).

In March, when the epicenter of pandemic moves to Europe and North America, but cases were still appearing in Asia, OTU_2, 3, and 5 increased is prevalence but OTU_1 remained as the most prevalent during this month (Fig. 3.C). Interestingly OTU_5 remained in relatively low frequencies (Fig. 3.C). Apparently, this month contains 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_3 and 5 increased its presence (Fig. 3.D) probably due to its higher representation (compared to March) in several continents such as South America, North America, and Europe (Fig. S4). During this month, Africa showed a high prevalence of OTU_2 (Fig. S4). We also witnessed the apparent establishment of OTU_3 in South America and Europe and OTU_5 in North America (Fig. S4).

May showed the current tendency of OTU_1 declining and OTU_3 increasing; OTU_2 and 5 were presented in similar frequencies between OTU_1 and 3. OTU_4 maintains its relatively low frequency (Fig. 3.E). From this month, South America reported very few isolated genomes and we cannot consider this continent to the analysis of this and follow months (Fig. S4).

The last months analyzed (June and July) presented frequencies distributions very similar to May, showing OTU_3 as the more frequent currently, but with OTU_2 maintaining its frequency, unlike OTU_5 that showed lower frequencies in June when compared with May and July. The current high prevalence of OTU_3 in the world and the observation that also from June in North America, its frequency is rapidly increasing highlights the importance of tracking and study mutations in Nucleocapsid that characterize this OTU.

Age, Gender and Patient Status distribution of OTUs

Relating the distribution of haplotypes according to patient information can help determine the preference of some OTUs for some characteristics of the patients. Thus, we analyze OTUs distribution according to age, gender, and patient status information available as metadata in the GISAID database.

Unfortunately, just 33.65 % of the 48 776 genomes analyzed have age and gender information (Fig. S6) and 4 108 genomes contain some information about the patient status (Fig. S7.B). Distribution of OTUs between age or gender categories did not show any well-defined preference. The distribution of OTUs in different categories was very similar (Fig. 4.A, B, and C).

In the case of patient status analysis, we noted that GISAID categories are not well organized and we had to reclassify the information into four categories, Not Informative, Asymptomatic, Mild, and Severe (Fig. S7.A). Using this classification scheme, we found that 55.82 % (2 293 genomes) falls in the Not Informative category, 37.22 % (1529 genomes) in the Mild category and just 2.31 % (95 genomes) and 4.65 % (191 genomes) could be classified as Asymptomatic and Severe, respectively (Fig. S7.B).

We analyzed the group distribution in the three informative categories (Asymptomatic, Mild and Severe) and found that isolates from patients with mild symptoms presented a relatively homogeneous distribution, with percentages between 27.7 % and 12.1 % from all five OTUs. The severe category was also relatively homogeneous with OTU_1 being the least prevalent (7.9 %).

Conversely, 75.8 % (72 of the 95) of the genomes classified as Asymptomatic belong to OTU_1 (Fig. 4.D).

However, we have to interpret these observations with extreme caution since most of the genomes from asymptomatic patients that belong to OTU_1 was isolated in Asia in February (Fig. S8.A) during a short period of three days (Fig. S8.B). Other genomes in the asymptomatic category belong to other OTUs and were isolated in different months and different continents (Fig. S8.B). Thus, we currently require more robust information to obtain a better-defined distribution of asymptomatic cases, as well as more and better-organized information related to patient status and characteristics to improve our analyses in OTUs distribution related to this data.

For this reason, we have created a web page that, in addition to assigning haplotypes to genomes that users can freely upload and make openly available information on the global geographic and temporal distribution of SARS-CoV-2 haplotype in an interactive way, allows researchers from all over the world to contribute voluntarily by offering correctly organized information on the characteristic of the patient (age, gender, condition (symptoms), comorbidities) to improve the analysis and monitor the possible appearance of haplotypes with certain preferences that can help in improving treatments for patients.

Description of the most frequent mutations

C241T

The C241T mutation is present in the 5` UTR region. In coronaviruses, the 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 recognizes 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 maybe 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

Mutation C3037T is a synonymous mutation in Nsp3, therefore, is more difficult to associate this change to an evolutionary advantage for the virus. This mutation occurred in the third position of a codon, one possibility is that this, change 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. Another option is simply genetic drift.

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 the 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 the absence of P322, the mutation P323L would probably be disfavored due to the flexibilization of the turn at 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 a 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 produces a Q57H variant of Orf3a (Fig. 7.C) that did not show significant differences in expression, stability, conductance, selectivity, or gating behavior[8]. We modeled Q57H mutation and we did not observe differences in the radius of constriction (Fig. 7.C) formed by aminoacid 57 but we observed slight differences in the electrostatic surface due to the ionizability of the histidine side chain (Fig. 7.D).

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 a low impact on 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. S9)

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 9 high-frequency mutations. These important variations (asserted mainly by their frequencies) need to be tracked during the pandemic.

Our haplotypes description 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 to the unprecedented worldwide efforts in the genome sequencing of SARS-CoV-2 and the public databases that rapidly share the information.

Our geographical and temporal analysis showed that OTU_3 is currently the more frequent haplotype circulating in the world. Even in North America that seems to be not the most frequent in the overall analysis, seems to be that in June it is the most frequent. These results highlight the importance to study mutations that characterize this haplotype, those in the nucleocapsid protein R203K and G204R.

Although OTU_1 was the only and the most abundant haplotype at the beginning of the pandemic, now is isolation is rare. This result shows an adaptation process of SARS-CoV-2 that is expected. This enunciate does not mean, that SARS-CoV-2 is now more infectious.

In the next months, these haplotypes description will need to be updated, identification of new haplotypes could be performed by combining the identification of new frequent mutations and phylogenetic analysis. We will continue monitoring the emergence of mutations that exceed our proposed cut-off of 0.1-0.9 NRF_p and this information will be rapidly shared with the scientific community through our web. This will also be accompanied by a continuous update of haplotypes information.

Our weak conclusion related to age, gender, and patient status information is due to the poorly organized metadata publicly available. Thus, we highlight the importance of correct management and organization of genomic metadata. Regarding this, we are setting up a web system where the scientific community can voluntarily share patient information associated with genomic data in an organized manner. This will allow filling current gaps in the public data on the correlation of haplotypes (or variants in general) with the severity of the disease, specific symptoms, comorbidities, among others.

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 test creation, among others.

MATERIAL AND METHODS:

Normalized frequency analysis of each base or gap by genomic position:

To perform the mutation frequency analysis, we first downloaded a total of 48 776 complete and high coverage genomes from the GISAID database (as of July 30th, 2020). This set of genomes was aligned using MAFFT with FFT-NS-2 strategy and default parameter settings[39]. Then, we removed columns that do not correspond to the region from nt 203 to nt 29674, and insertions respect to the genome EPI_ISL_402125. After that, these regions were aligned using MAFFT with FFT-NS-2 strategy and default parameter settings[39]. Subalignments corresponding to genomes divided by continent-month combinations was extracted and relative frequencies of each base or gap in each genomic position were calculated ($RF_{p,m-c}$) using a python script. These relative frequencies were multiplied by the number of cases reported in the respective continent-month combination (CN_{m-c}) obtaining an estimation of the number of cases that present a virus with a specific base or gap in a specific genomic position (RF_pCN_{m-c}). Finally, we added the RF_pCN_{m-c} of each subalignment and divided it by the total number of cases in the world ($\sum_{m-c} RF_oCN_{m-c1})/TCN_w$. This procedure allows us to obtain a relative frequency normalized by cases of each base or gap in each genomic position (NRF_p). The number of cases of each country was obtained from the European Centre for Disease Prevention and Control: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. We used the number of cases of countries with at least one genome sequenced and deposited in GISAID database. Also, we just consider in the analysis month-continent combinations with at least 90 genomes sequenced.

Phylogenetic tree construction:

Using an alignment of the 19 486 complete, high coverage genomes without ambiguities, 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 genome EPI_ISL_408601 as the outgroup. To generate tree figures with continent or date information by tip we used the maximum likelihood tree and ggtree package in R[45,46].

OTUs determination:

Positions with between 0.1 and 0.9 NRF_p were extracted from the alignment of the non-ambiguities data set of 19486 and were associated with the whole-genome rooted tree using the MSA function from the ggtree package[45,46] in R. Then, we visually examined to identify the major haplotypes based in these positions, designated as OTUs (Operational Taxonomic Units). Haplotypes identification based in our NRFp calculation reduced the bias of the different number of genomes sequenced in each continent and each month by integrating the less biased information of the number of cases. Although, other biases are more difficult, if possible, to reduce or eliminate.

Analysis of OTUs geographical distribution:

In this analysis, we randomly separate the genomes into 6 groups of 8 129 genomes each and we analyzed them independently. After that, genomes in each sample was divided by continents and by months. In these divisions, OTUs relative frequencies were calculated for each OTU in each month-continent combination (O_nF_{m-c}). Then, we multiplied these (O_nF_{m-c}) frequencies by the number of cases corresponding to the respective month-continent (CN_{m-c}) to obtain an estimation of the number of cases caused by a specific OTU in a respective month-continent (O_nCN_{m-c}). After, these products were grouped by continents, and those from the same continent were added and then divided by the total number of cases in the continent analyzed

$(\sum_{m-c1} O_n CN_{m-c1})/TCN_{c1}$. Thus, obtaining a frequency normalized by cases for each OTU in each continent. Finally, following this procedure in each sample, we statistically compared the mean of those six samples 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. The number of cases of each country was obtained from the European Centre for Disease Prevention and Control: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. We used the number of cases of countries with at least one genome sequenced and deposited in GISAID database. Also, we just consider in the analysis month-continent combinations with at least 90 genomes sequenced.

Analysis of OTUs temporal distribution:

Following a similar procedure used in the geographical analysis, we now grouped the products $O_n CN_{m-c}$ by months, added them, and then divided by the total number of cases in the analyzed month $(\sum_{m1-c} O_n CN_{m1-c})/TCN_{m1}$. As in the geographical analysis, the mean of the six samples was statistically compared using the same procedures and with exactly the same consideration of month-continent combinations.

Analysis of age, gender, and patient status with OTUs distribution:

4108 complete and high coverage genomes with patient status information were downloaded from the GISAID database (as of 30th July) and classified in OTUs using python scripts. 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 Figure S6.A. We calculate the relative and absolute frequency of OTUs in each patient status category. Also, using all the available information on gender and age in the 48 776 genomes, we calculated the relative and absolute frequency of OTUs by age and gender.

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 25 August 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, Hill V, McCrone J, Ruis C, du Plessis L, Pybus O. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. 2020. Nature Microbiology. https://doi.org/10.1038/s41564-020-0770-5
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.

524 36. Luo H, Chen Q, Chen J, Chen K, Shen X, Jiang H. 2005. The nucleocapsid protein of SARS
525 coronavirus has a high binding affinity to the human cellular heterogeneous nuclear
526 ribonucleoprotein A1. Vol. 579, 2623-2628. FEBS letters.
527 37. Chang C, Chen C, Chiang M, Hsu Y, Huang T. 2013. Transient oligomerization of the SARS-
528 CoV N protein – Implication for virus ribonucleoprotein packaging. Vol. 8(5), e65045.
529 PlosONE.
530 38. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. 2004. Prediction of post-
531 translational glycosylation and phosphorylation of proteins from the aminoacid
532 sequence. Vol. 4, 1633-1649. Proteomics.
533 39. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
534 sequence alignment based on fast Fourier transform. Vol. 30(14), 3059-3066. Nucleic
535 Acids Research.
536 40. Minh B, Schmidt H, Chernomor O, Schrempf D, Woodhams M, von Haeseler A, Lanfear
537 R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the
538 genomic era. Vol. 37, 1530-1534. Molecular Biology and Evolution.
539 41. Tavaré S. 1986. Some mathematical questions in biology: DNA sequence analysis.
540 Lectures on mathematics in the life sciences. Vol. 17, 57-86. American Mathematical
541 society.
542 42. Soubrier J, Steel M, Lee M, Sarkissian C, Guindon S, Ho S, Cooper A. 2012. The influence
543 of rate heterogeneity among sites on the time dependence of molecular rates. Vol.
544 29(11), 3345-3358. Molecular Biology and Evolution.
545 43. Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Vol. 139,
546 993-1005. Genetics
547 44. Hoang D, Chernomor O, Haeseler A, Minh B, Vinh L. 2017. UFBoot2: Improving the
548 ultrafast bootstrap approximation. Vol. 35(2), 518-522. Molecular Biology and
549 Evolution.
550 45. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Vol. 69, 1-18- Current
551 Protocols in Bioinformatics.
552 46. Yu G, Smith D, Zhu H, Guan Y, Lam T. 2017. GGTREE: an R package for visualization and
553 annotation of phylogenetic trees with their covariates and other associated data. Vol. 8,
554 28-36. Methods in Ecology and Evolution.

555 **Competing interests:**

556 The authors declare no competing interests

557 **Acknowledgements:**

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

566

567

Figure 1. Five haplotypes (or OTUs) based in nine positions can classify 96.5 % of the genomes. A) Table showing haplotype of each OTU, regions, and aminoacids changes caused by these mutations. B) Rooted tree of 19486 SARS-CoV-2 complete and non-ambiguous genomes associated with an alignment of nine genomic positions (241, 1059, 3037, 14408, 23403, 25563, 28881, 28882, 28883) showing a good correlation between haplotypes (OTUs) based in these nine positions. Tips of the tree where colored based in the OTU. C) Bar diagram showing OTUs distribution of the genomes (0 correspond to unclassified genomes).

Figure 2. By cases normalized continent distribution of OTUs during seven months of the pandemic. A) Unrooted tree of complete non-ambiguous genomes, tips were colored according to OTUs, and points in each tip were colored according to the continent. B-G) Boxplots of normalized relative frequencies of OTUs in each continent (B, North America; C, South America; D, Europe; E, Asia; F, Oceania; G, Africa).

Figure 3. By cases normalized temporal distribution of OTUs showed OTU_3 as the currently most prevalent. A) Rooted tree of complete non-ambiguous genomes showing temporal distribution. Tips were colored by OTUs and points in each tip were colored according to the isolation date. B-E) Boxplot of OTUs distribution in each month (B, February; C, March; D, April; E, May; F, June; G, July).

Figure 4. Age, gender, and patient status distribution by OTUs do not show preferences for patient characteristics. A) Relative and absolute frequencies of OTUs distribution by age. B) Age distribution was grouped by ranges and relative and absolutes frequency by OTUs is showed. C) OTUs distribution by gender. D) Relative and absolute frequencies of OTUs by patient status categories.

Figure 5. P323L could impact the 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 the methyl group of leucine are shorter with L323.

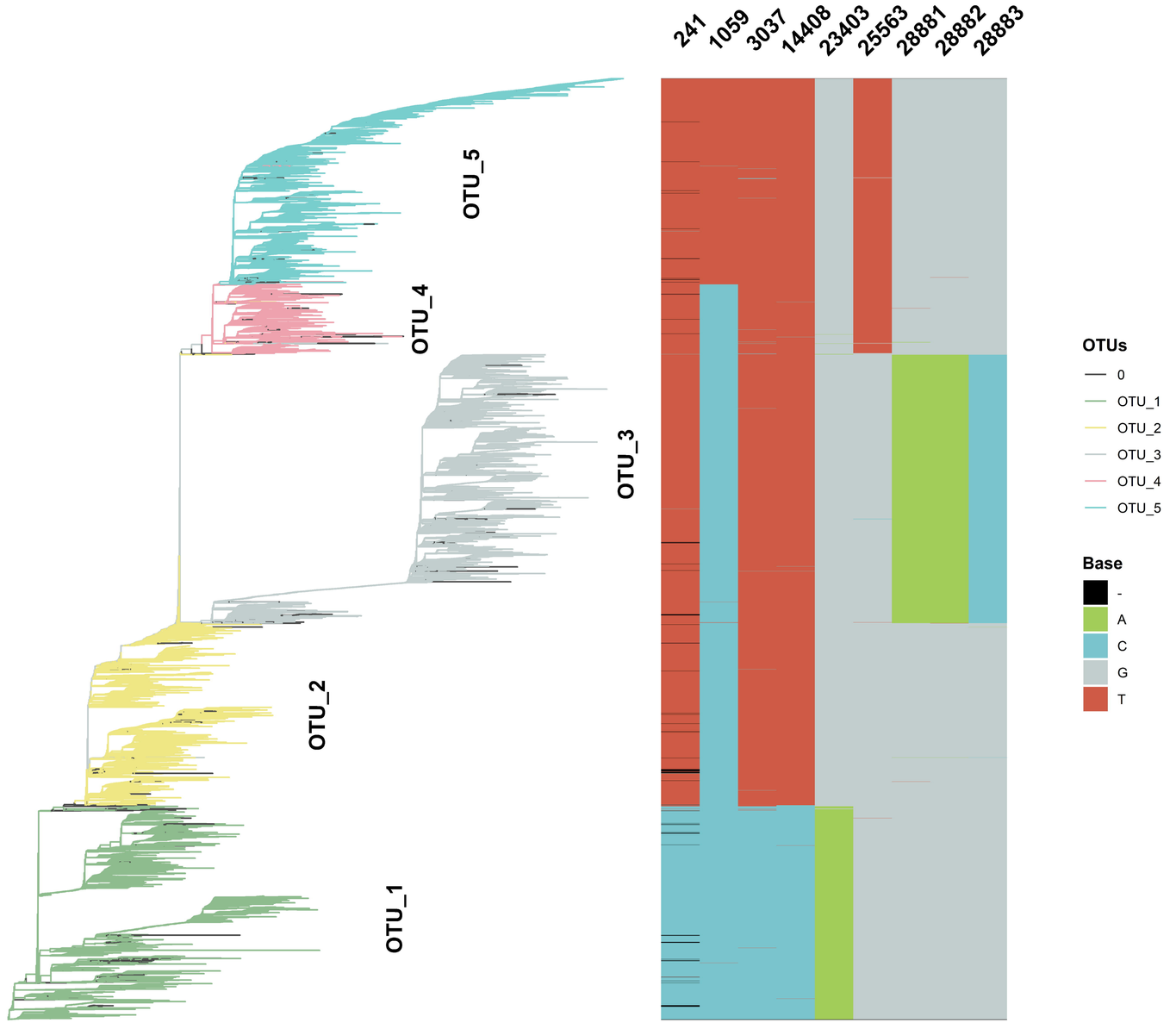
Figure 6. Structural hypotheses about D614G mutation in Spike protein. A) Structure of the open state of Spike trimer (PDB ID: 6YVB) colored by domains. B) Distances between side chains of two possible rotamers of D614 (1`-D614 and 2`-D614) and T859. Except for 1`-D614 and carbonyl group of T859, the 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 points to the opposite side of D614 showing that there is no salt bridge. B, C, and D show electron density maps of the side chains of the labeled residues.

Figure 7. Orf3a Q57H does not modify pore constriction distances but electrostatics distribution. A) Structure of the Orf3a dimer (PDB ID: 6XDC) colored by domains. The right of A shows the same structure but in an upper view. B) Orf3a showing the central pore, in the 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 superposed. Below we show a transversal cut of the pore near to the fifth. The pore radius in two variants is 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 slightly more positive region at the height of the fifth constriction.

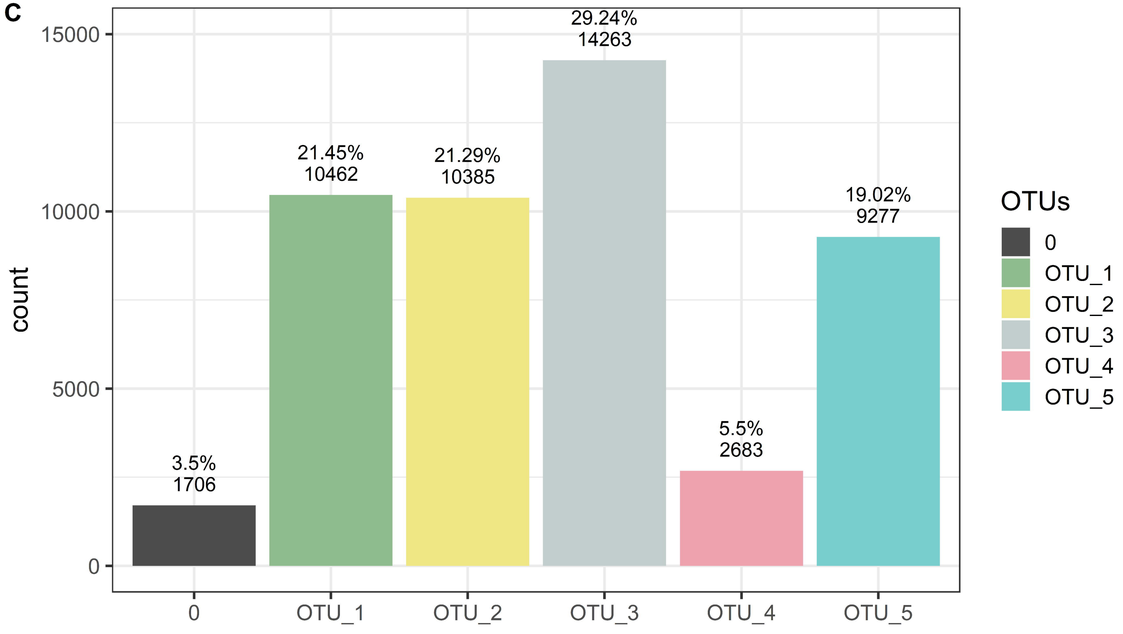
A

Position	OTU_1	OTU_2	OTU_3	OTU_4	OTU_5	Region	AA_change
241	C	T	T	T	T	5'UTR	'''
1059	C	C	C	C	T	Nsp2	T85I
3037	C	T	T	T	T	Nsp3	Syn
14408	C	T	T	T	T	Nsp12	P323L
23403	A	G	G	G	G	S	D614G
25563	T	T	T	T	T	Orf3a	Q57H
28881	G	G	A	G	G	N	R203K
28882	G	G	A	G	G	N	R203K
28883	G	G	C	G	G	N	G204R

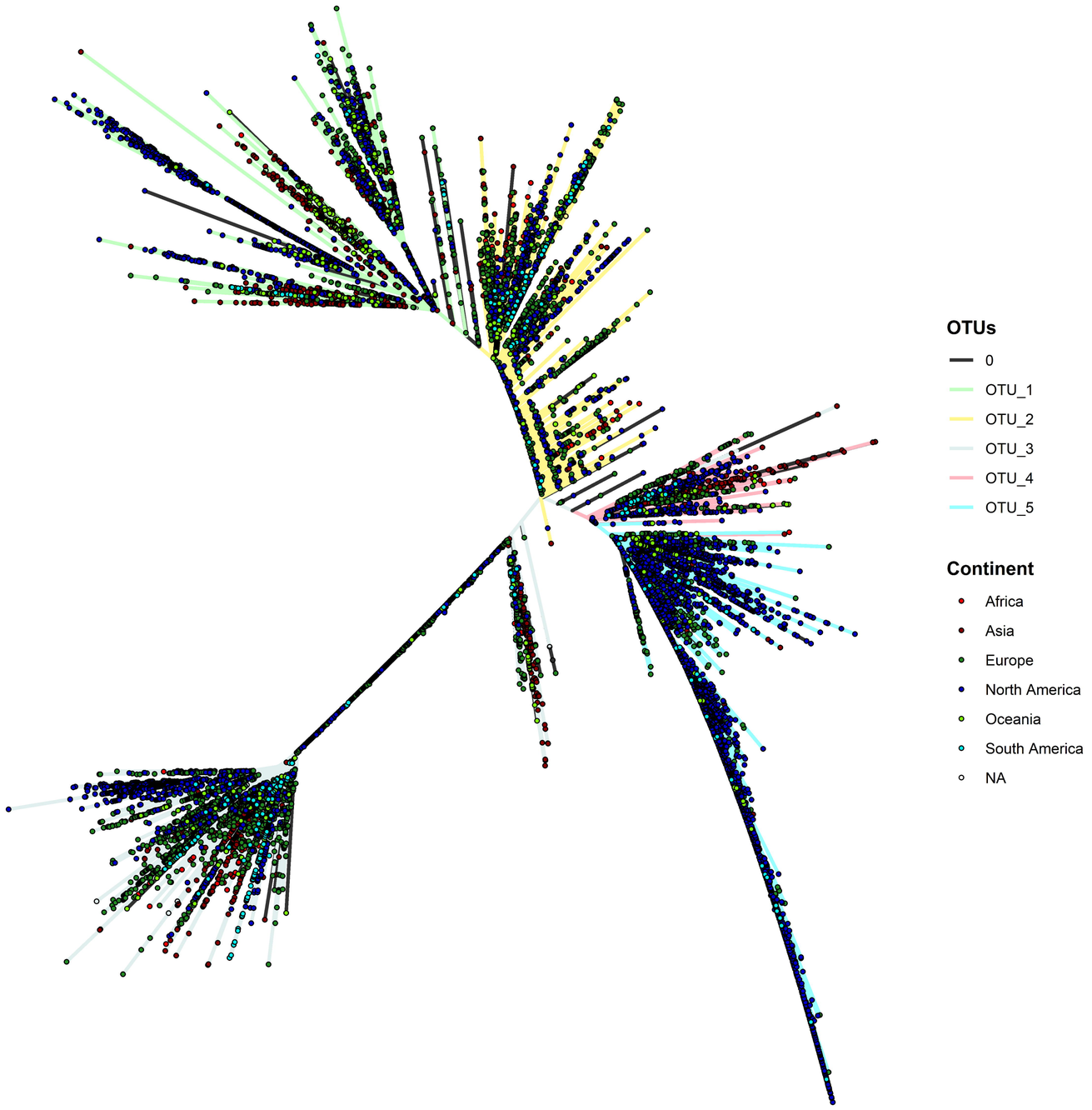
B



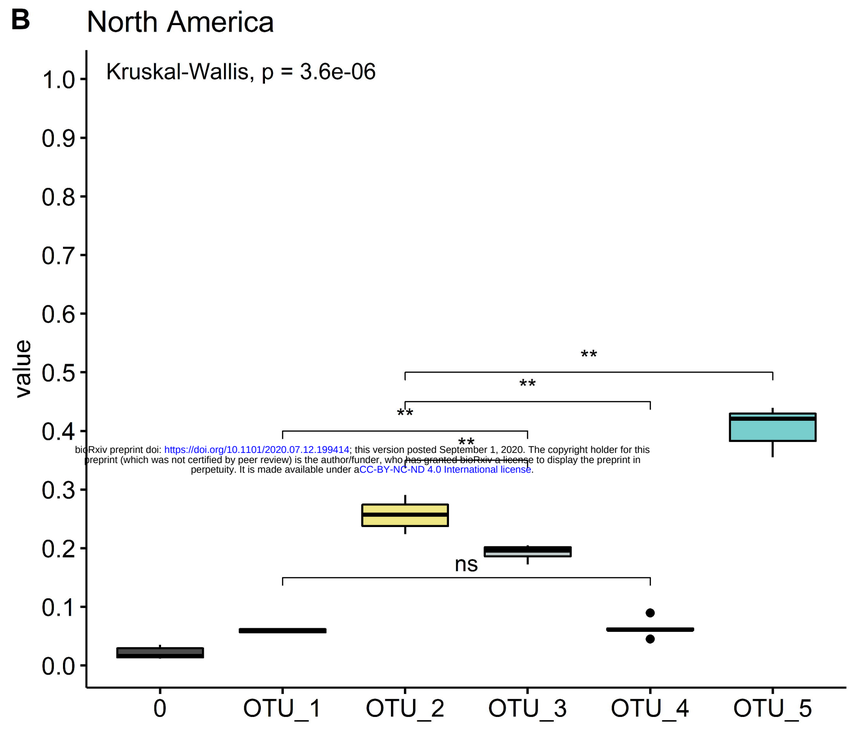
C



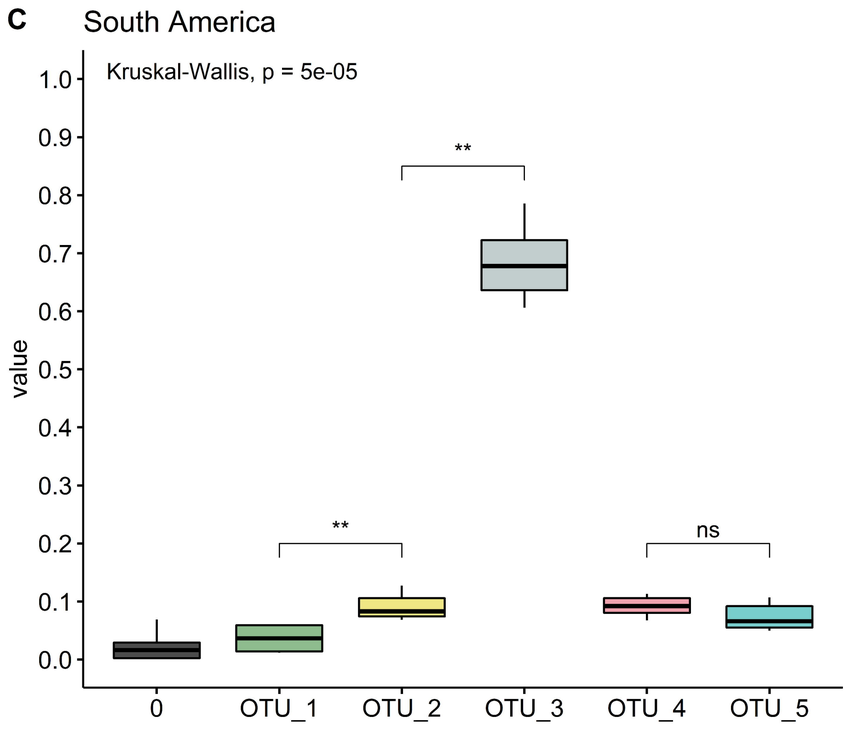
A



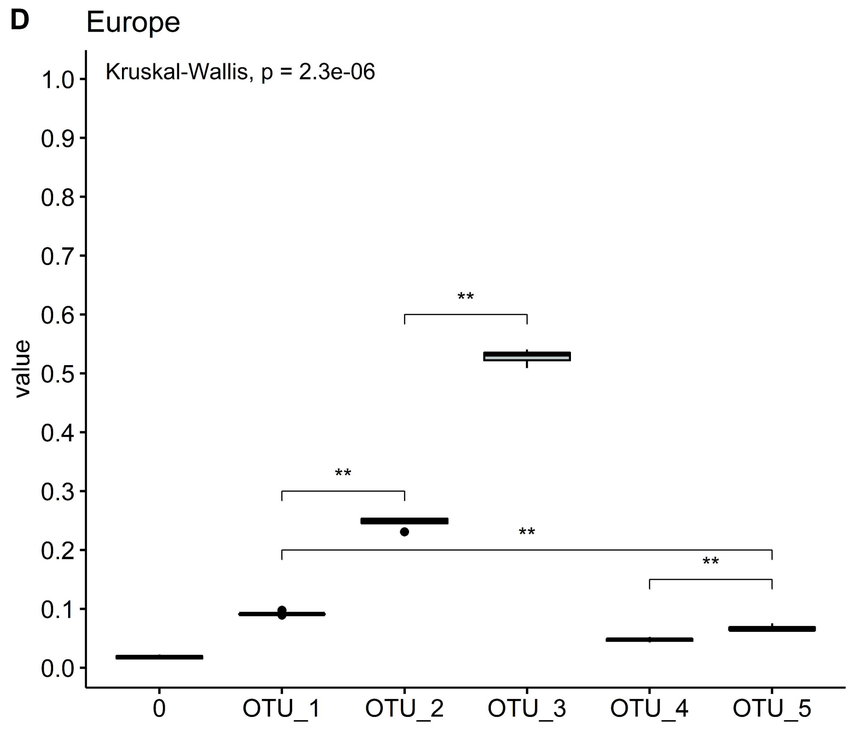
B



C



D

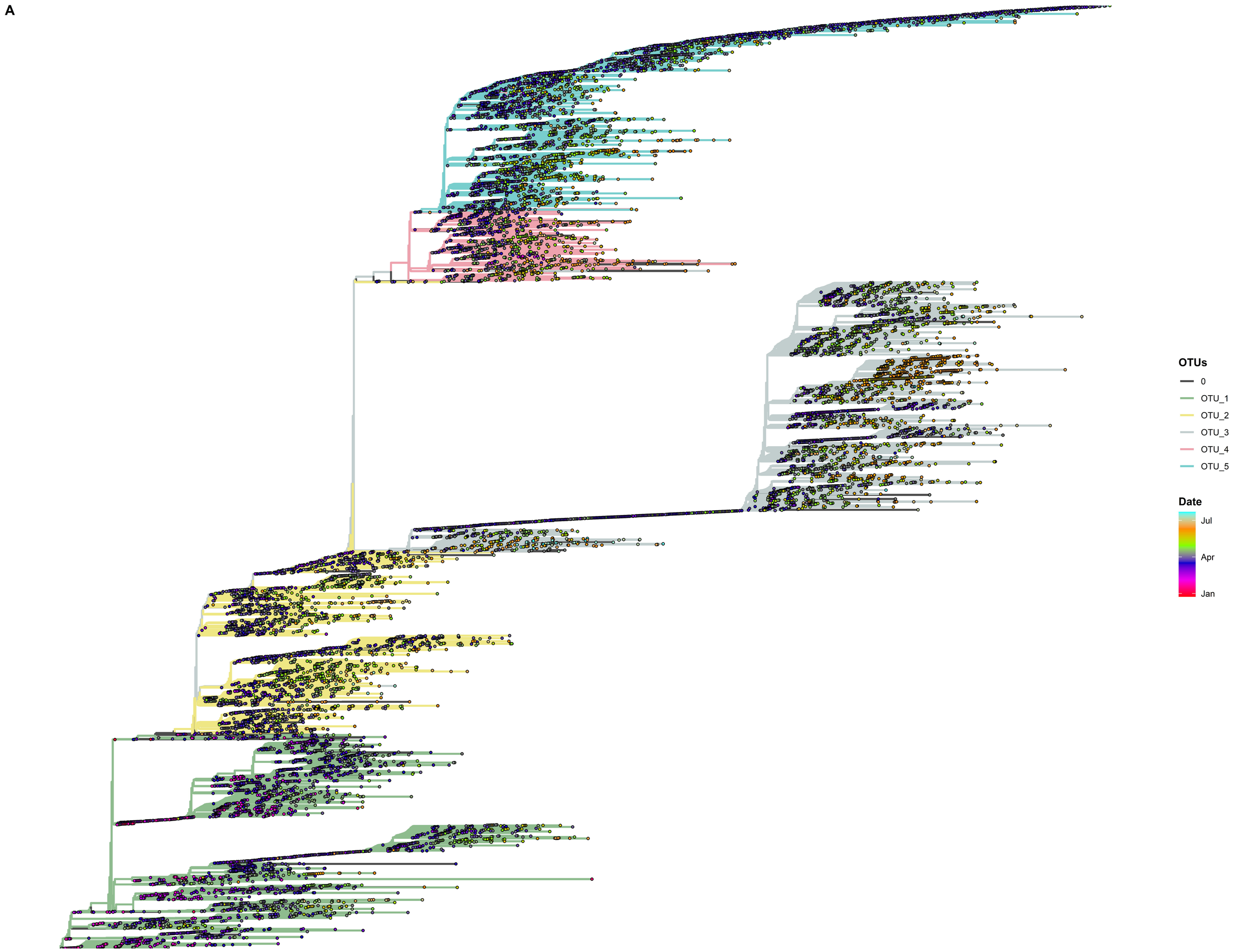


E

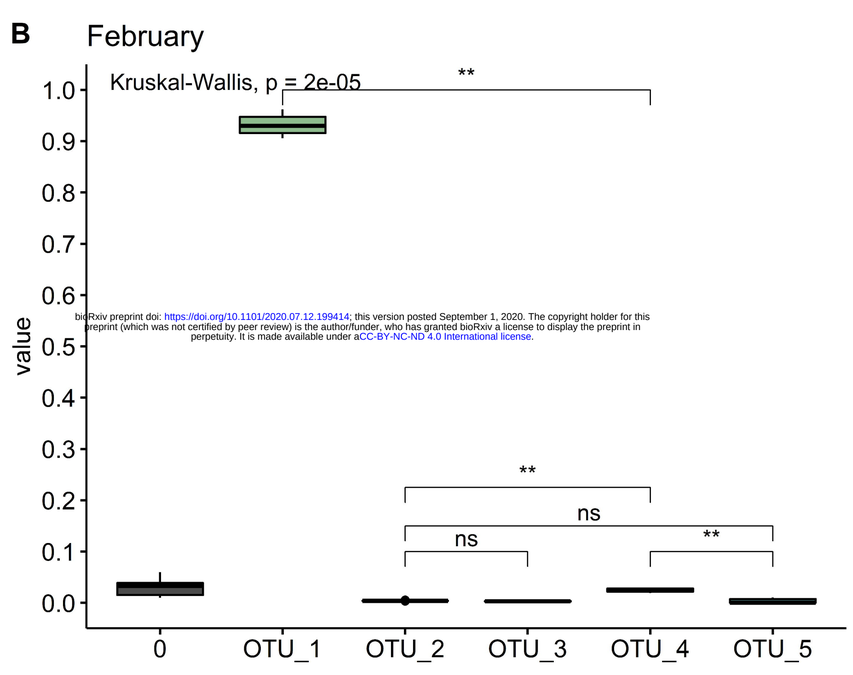
F

G

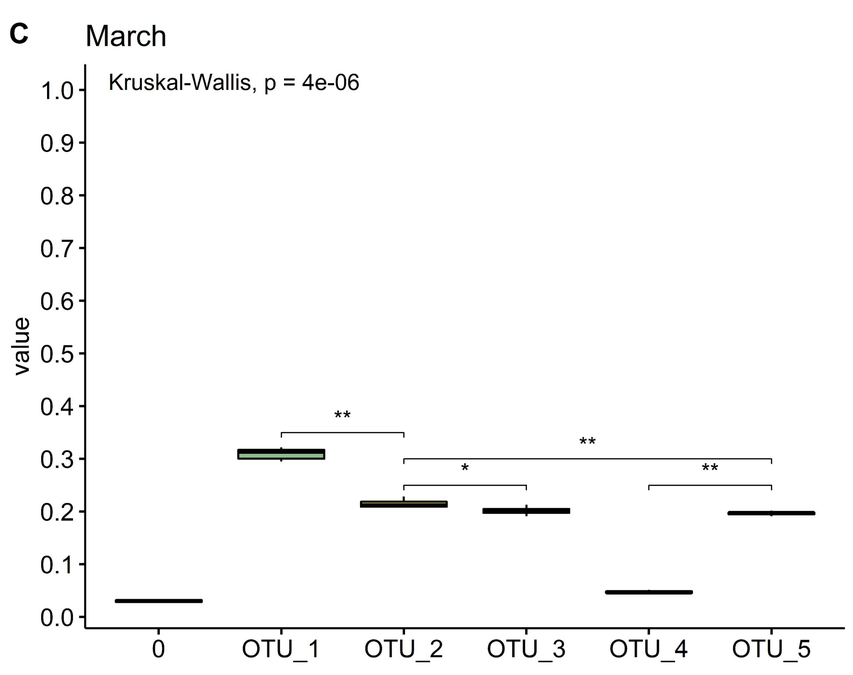
A



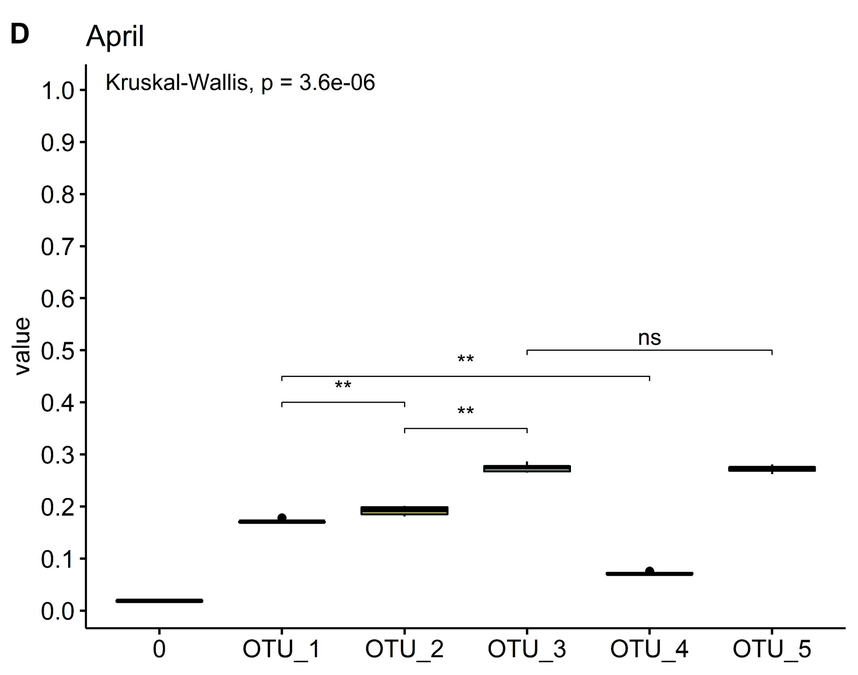
B



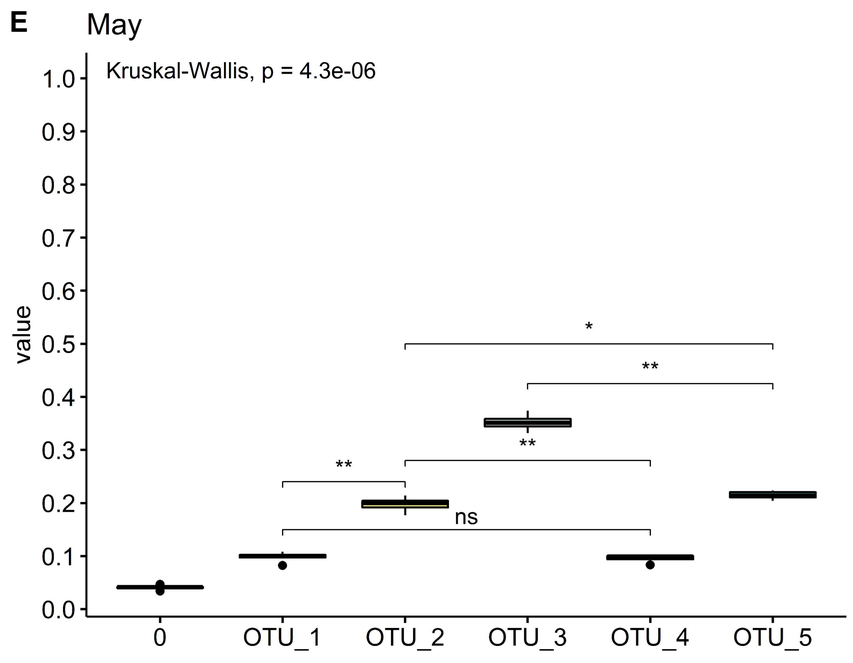
C



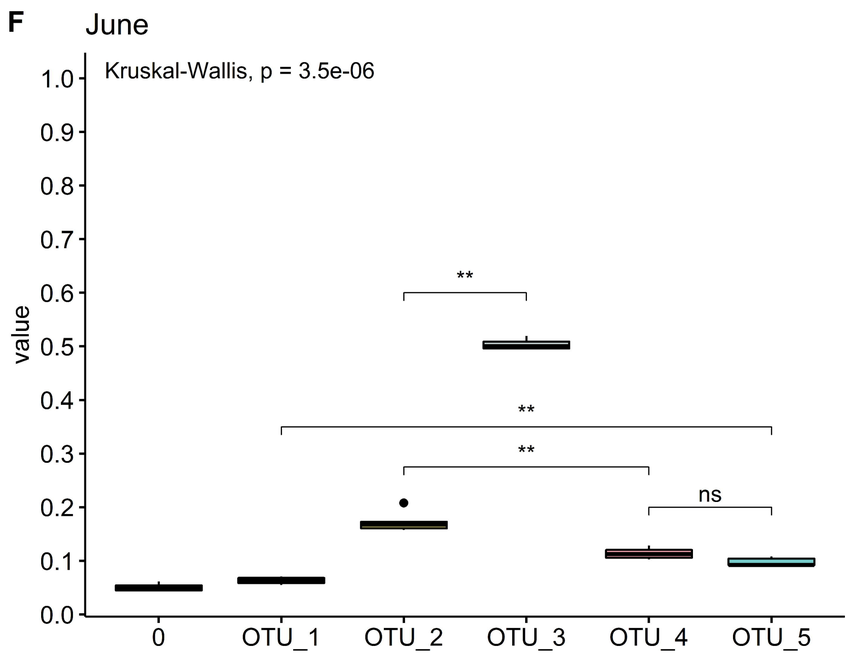
D



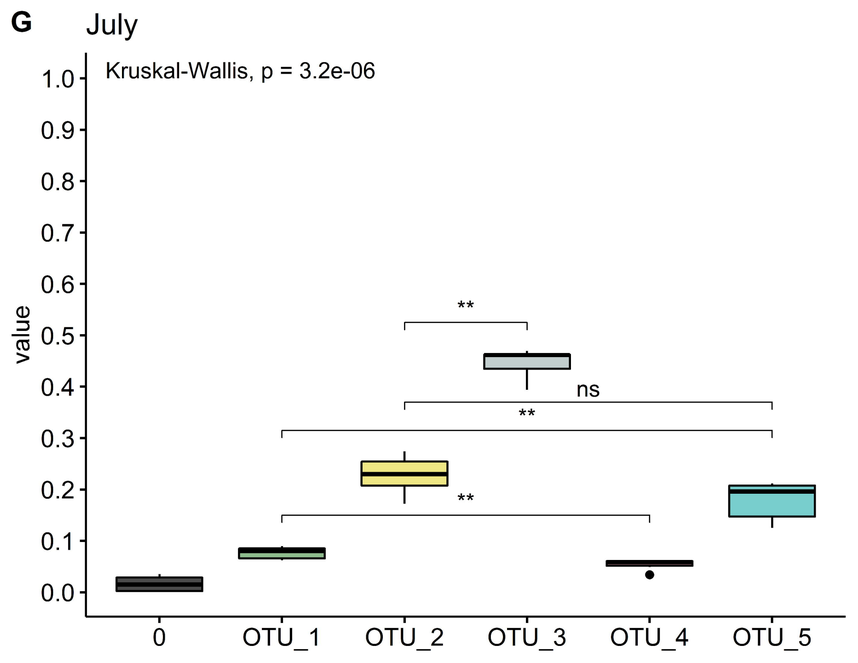
E

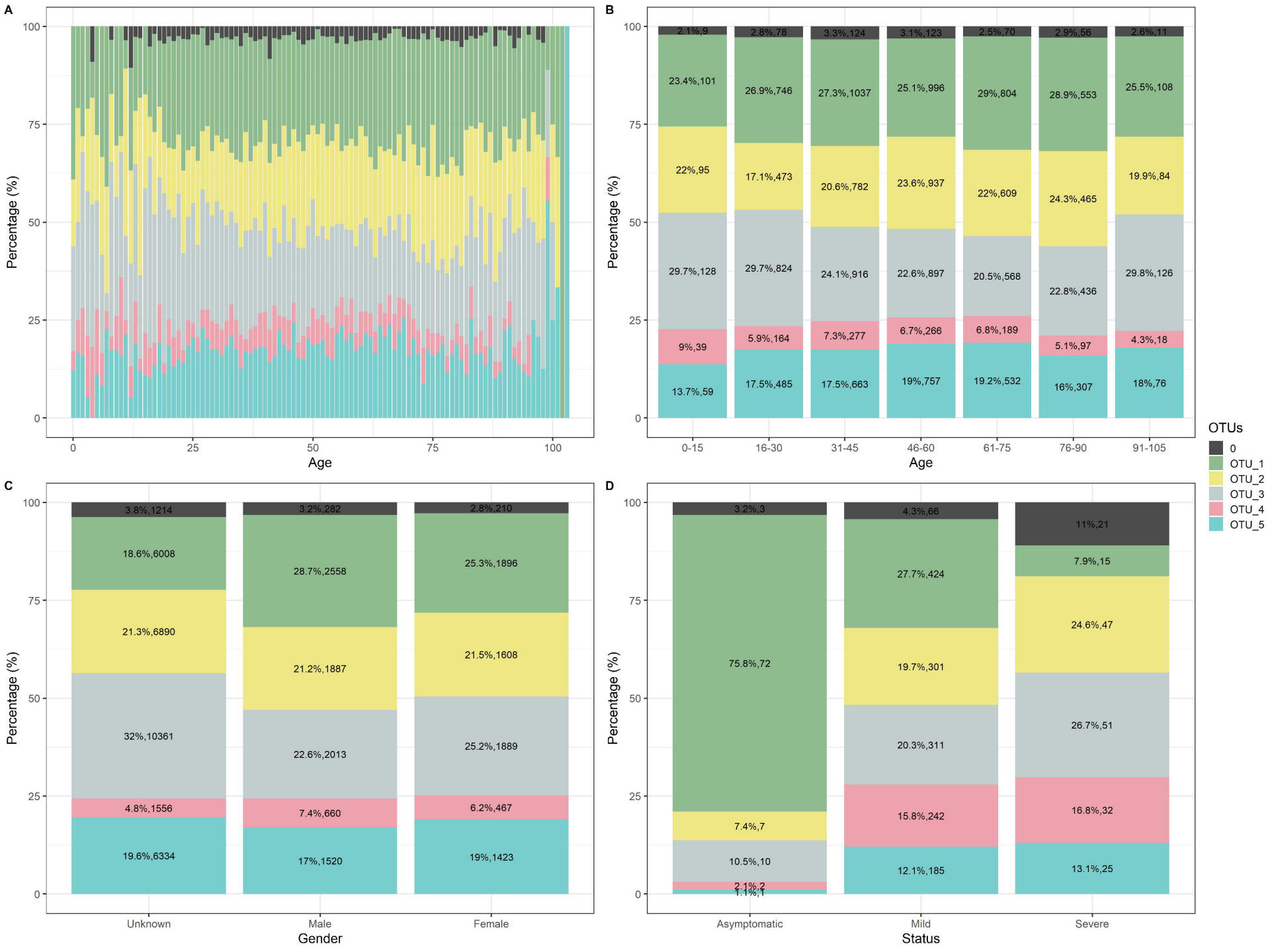


F



G




A
B
C
D
Percentage (%)
Age
Gender
Status
OTUs
0
OTU_1
OTU_2
OTU_3
OTU_4
OTU_5
0-15
16-30
31-45
46-60
61-75
76-90
91-105
Unknown
Male
Female
Asymptomatic
Mild
Severe
2.1%,9
23.4%,101
22%,95
29.7%,128
9%,39
13.7%,59
2.8%,78
26.9%,746
17.1%,473
29.7%,824
5.9%,164
17.5%,485
3.3%,124
27.3%,1037
20.6%,782
24.1%,916
7.3%,277
17.5%,663
3.1%,123
25.1%,996
23.6%,937
22.6%,897
6.7%,266
19%,757
2.5%,70
29%,804
22%,609
20.5%,568
6.8%,189
19.2%,532
2.9%,56
28.9%,553
24.3%,465
22.8%,436
5.1%,97
16%,307
2.6%,11
25.5%,108
19.9%,84
29.8%,126
4.3%,18
18%,76
3.8%,1214
18.6%,6008
21.3%,6890
32%,10361
4.8%,1556
19.6%,6334
3.2%,282
28.7%,2558
21.2%,1887
22.6%,2013
7.4%,660
17%,1520
2.8%,210
25.3%,1896
21.5%,1608
25.2%,1889
6.2%,467
19%,1423
3.2%,3
75.8%,72
7.4%,7
10.5%,10
2.1%,2
1.1%,1
4.3%,66
27.7%,424
19.7%,301
20.3%,311
15.8%,242
12.1%,185
11%,21
7.9%,15
24.6%,47
26.7%,51
16.8%,32
13.1%,25