SARS-CoV-2 Genomic Surveillance in Costa Rica: Evidence of a Divergent Population and an Increased Detection of a Spike T1117I Mutation

Jose Arturo Molina-Mora[1]*, Estela Cordero-Laurent[2], Adriana Godínez[2], Melany Calderón-Osorno[2], Hebleen Brenes[2], Claudio Soto-Garita[2], Cristian Pérez-Corrales[3], COINGESA-CR Consorcio Interinstitucional de Estudios Genómicos del SARS-CoV-2 Costa Rica[4], Jan Felix Drexler[5], Andres Moreira-Soto[5,1], Eugenia Corrales-Aguilar[1], Francisco Duarte-Martínez[2].

[1] Research Center for Tropical Diseases (CIET), Faculty of Microbiology, University of Costa Rica, San José, Costa Rica.
[2] Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud (INCIENSA), Tres Ríos, Cartago, Costa Rica.
[3] Hospital Nacional De Niños Dr. Carlos Sáenz Herrera, Caja Costarricense de Seguro Social (CCSS), Costa Rica.
[4] See information of members/institutions in Appendix.
[5] Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany.
* Corresponding author; email: jose.molinamora@ucr.ac.cr

Appendix

Members and collaborating institutions of COINGESA-CR Consorcio interinstitucional de estudios genómicos del SARS-CoV-2 - Costa Rica:
Mariel López Moya (Laboratorio Diagnostico de COVID-19, INCIENSA), Grettel Chanto Chacón (CNRB, INCIENSA), Andrei Montero Bonilla (CCSS), Teresita Somogyi (CCSS), Pei Ling Chang (Laboratorios LABIN), Alberto Bonilla Sequeira (Laboratorios LABIN), Ignacio Pacheco Zamora (Laboratorios LABIN), David Rodríguez Masis (Laboratorios LABIN), Roger Soto Palma (Hospital CIMA), Hugo Núñez Navas (Laboratorio Clínico San José), Rodrigo Cruz Jiménez (Hospital Clínica Bíblica), Rodolfo Garbanzo Garvey (Hospital Clínica Bíblica), Karla Gutiérrez González (Hospital Clínica Bíblica), Área De Salud Alajuela Central, Área De Salud Alajuela Norte - Clínica Dr. Marcial Rodríguez, Área De Salud Alajuela Sur, Área De Salud Aserrí, Área De Salud Catedral Noreste, Área De Salud Ciudad Quesada, Área De Salud Corredores, Área De Salud Desamparados 1 - Clínica Dr. Marcial Fallas, Área De Salud Escazú (Coopesana), Área De Salud Fortuna, Área De Salud Goicoechea, Área De Salud La Cruz, Área De Salud La Unión, Área De Salud Los Chiles, Área De Salud Los Santos, Área De Salud Mata Redonda-Hospital - Clínica Dr. Moreno Cañas, Área De Salud Orotina-San Mateo, Área De Salud Pavas (Coopesalud), Área De Salud Tibás (Coopesain) - Clínica Integrada Rodrigo Fournier, Área De Salud Tibas-Uruca-Merced - Clínica Dr. Clorito Picado, Centro Nacional De Rehabilitación Humberto Araya Rojas (CENARE), Clínica Bíblica, Ebais Concepción Norte, Hospital CIMA, Hospital Ciudad Neily, Hospital De Las Mujeres Dr. Adolfo Carit, Hospital De Niños Dr. Carlos Sáenz Herrera, Hospital Dr. Fernando Escalante Pradilla, Hospital Dr. Rafael A. Calderón Guardia, Hospital Dr. Rafael A. Calderón Guardia, Hospital La Anexión, Hospital Metropolitano, Hospital México, Hospital San Juan De Dios, Hospital San Rafael De Alajuela, Hospital San Vicente De Paul, Hospital Upala, Laboratorios LABIN, Laboratorio Clínico San José, y Organismos de Investigación Judicial, Eugenia Corrales-Aguilar (University of Costa Rica), Jose Arturo Molina-Mora (University of Costa Rica), Andres Moreira-Soto (University of Costa Rica), Francisco Duarte-Martínez (INCIENSA), Hebleen Brenes Porras (INCIENSA), Claudio Soto-Garita (INCIENSA), Estela Cordero (INCIENSA), Adriana Godínez (INCIENSA), Melany Calderon (INCIENSA), Cristian Pérez-Corrales (CCSS).

Abstract

Genome sequencing is a key strategy in the surveillance of SARS-CoV-2, the virus responsible for the COVID-19 pandemic. Latin America is the hardest hit region of the world, accumulating almost 20% of COVID-19 cases worldwide. Costa Rica was first exemplary for the region in its pandemic control, declaring a swift state of emergency on March 16th that led to a low quantity of cases, until measures were lifted in early May. From the first detected case in March 6th to December 31st almost 170 000 cases have been reported in Costa Rica, 99.5% of them from May onwards. We analyzed the genomic variability during the SARS-CoV-2 pandemic in Costa Rica using 185 sequences, 52 from the first months of the pandemic, and 133 from the current wave.

Three GISAID clades (G, GH, and GR) and three PANGOLIN lineages (B.1, B.1.1, and B.1.291) are predominant, with phylogenetic relationships that are in line with the results of other Latin American countries, suggesting introduction and multiple re-introductions from other regions of the world. The whole-genome variant calling analysis identified a total of 283 distinct nucleotide variants. These correspond mostly to non-synonymous mutations (51.6%, 146) but 45.6% (129) corresponded to synonymous mutations. The 283 variants showed an expected power-law distribution: 190 single nucleotide mutations were identified in single sequences, only 16 single nucleotide mutations were found in >5% sequences, and only two mutations in >50% genomes. These mutations were distributed through the whole genome. However, 63.6% were present in ORF1ab, 11.7% in Spike gene and 10.6% in the Nucleocapsid gene. Additionally, the prevalence of worldwide-found variant D614G in the Spike (98.9% in Costa Rica), ORF8 L84S (1.1%) is similar to what is found elsewhere. Interestingly, the frequency of mutation T1117I in the Spike has increased during the current pandemic wave beginning in May 2020 in Costa Rica, reaching 29.2% detection in the full genome analyses in November 2020. This variant has been observed in less than 1% of the GISAID reported sequences worldwide in all the 2020. Structural modeling of the Spike protein with the T1117I mutation suggest a potential effect on the viral oligomerization needed for cell infection, but no differences with other genomes on transmissibility, severity nor vaccine effectiveness are predicted. Nevertheless, in-vitro experiments are required to support these *in-silico* findings. In conclusion, genome analyses of the SARS-CoV-2 sequences over the course of COVID-19 pandemic in Costa Rica suggest introduction of lineages from other countries as travel bans and measures were lifted, similar to results found in other studies, as well as an increase in the Spike-T1117I variant that needs to be monitored and studied in further analyses as part of the surveillance program during the pandemic.

INTRODUCTION

COVID-19 (COronaVIrus Disease 2019) is an infectious disease caused by the SARS-CoV-2 virus. It was first described in late December 2019, in an outbreak of atypical pneumonia in the city of Wuhan, Hubei province, China (Lu, Stratton, & Tang, 2020). Until the end of 2020, this virus has resulted in more than 81 million confirmed cases and 1.8 million deaths worldwide (https://coronavirus.jhu.edu/map.html). Latin America is the hardest hit region of the world, accumulating almost 20% of COVID-19 cases worldwide (https://ourworldindata.org/coronavirus). The first confirmed case of COVID-19 in Costa Rica was reported in San José (capital city) on March 6th, 2020. Costa Rica was first exemplary for the region in its pandemic control, declaring a swift state of emergency on March 16th that led to a low quantity of cases, until measures were lifted in early May. 99.5% of the cases have been found from May onwards, thus adding up to form a second bigger wave. Until December 31st, 2020 (Figure 1-A-B), Costa Rica has confirmed 169 321 cases and 2 185 deaths (https://www.ministeriodesalud.go.cr/).

The first complete SARS-CoV-2 genome globally was published on January 10, 2020, revealing that the genome is composed of a single-stranded RNA with 29 903 nucleotides (Wu et al., 2020). Genetic diversity of SARS-CoV-2 genomes has demonstrated a large number of independent introductions of the virus into many countries of the world but China (van Dorp et al., 2020). Although strong relationships between specific variants or clades in SARS-CoV-2 and the clinical outcome have not been identified (Grubaugh, Hanage, & Rasmussen, 2020; Hodcroft et al., 2020; van Dorp et al., 2020), different relevant parameters could be impacted by mutations in the virus genome. These include transmissibility, severity, clinical management, mortality, vaccines as well as the performance of molecular tests for diagnosis (Hodcroft et al., 2020; Koyama, Platt, & Parida, 2020; Teng, Sobitan, Rhoades, Liu, & Tang, 2020). Until now, host's genetic factors and, more importantly, comorbidities seem to have the predominant impact on the fate of COVID-19 patients (LoPresti, Beck, Duggal, Cummings, & Solomon, 2020; Sironi et al., 2020; Toyoshima, Nemoto, Matsumoto, Nakamura, & Kiyotani, 2020). A few studies have suggested possible associations between genetic markers of the virus and clinical fate, which are still under investigation (Hodcroft et al., 2020; Korber et al., 2020; Toyoshima et al., 2020).

Due to new mutations and appearance over time of new genome clusters, whole-genome sequencing is a key step in the surveillance of this pathogen (Castillo et al., 2020). In this regard, diverse classification schemes using specific genetic markers have been proposed to monitor the evolution of the virus in the human population. The GISAID (Global Initiative on Sharing All Influenza Data, https://www.gisaid.org/) classification scheme by clades (GISAID, 2020) and the PANGOLIN (Phylogenetic Assignment of Named Global Outbreak LINeages) classification scheme by lineages (Rambaut et al., 2020) have been widely used for monitoring the current pandemic.

At the end of April 2020, genome sequencing of SARS-CoV-2 virus from Costa Rican cases of COVID-19 was started as part of the national epidemiological surveillance (INCIENSA & Ministerio de Salud, 2020). In this work we present the genome analysis of 185 SARS-CoV-2 sequences which have been obtained from Costa Rican cases of COVID-19 in the period March-November 2020, including 52 genomes from the first months of the pandemic, and 133 from the current second wave (from May onwards). The study aimed to investigate the diversity of the viral population of SARS-CoV-2 genomes circulating in Costa Rica. To this end, we included the genome sequencing and assembly, the analysis of variants or mutations, the construction of a phylogenetic tree, as well as the analysis of the possible association of genomes by clades/lineages with epidemiological data. Also, we report an increase in local detection of a variant with a Spike mutation T1117I not broadly reported elsewhere.

METHODS

Clinical Isolates

Material from nasopharyngeal swabs was obtained from Costa Rican cases of COVID-19, in the period between March and November 2020. All the patients were diagnosed in INCIENSA (Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud) or different public and private clinical laboratories by real-time reverse transcription polymerase chain reaction (RT-PCR), using specific probe and primers according to the guidelines suggested by the Pan American Health Organization and the World Health Organization (PAHO, 2020) and the Ministry of Health of Costa Rica. Only samples with a CT<25 were considered for the genome sequencing. Samples (167) were sent to the sequencing service of INCIENSA for further analyses. Other 18 assembled sequences of Costa Rica were included (see below, ID: CRC-007-024), completing 185 genomes. Gender distribution of the samples was 112 male and 71 female patients (two cases without gender information), and the age ranged from 4 to 92 years old. Ten deaths related to COVID-19 were included in these analyses. More details in Table 1.

Genome sequencing

SARS-CoV-2 RNA was extracted with QIAamp Viral Mini Kit (Qiagen) in a QIACube extractor or with Maxwell® RSC Viral Total Nucleic Acid Purification Kit (Promega) in a Maxwell® RSC instrument following the manufacturer's protocol. cDNA for library preparation was then synthesized and purified using either the ARTIC SARS-CoV-2 amplification protocol as described in https://artic.network/ncov-2019 (for samples 071-138), the SARS-CoV-2 amplification protocol described in (Castillo et al., 2020) (samples 001-006 and 026-070) or the FIOCRUZ protocol (samples 0140-0186) described in (Resende et al., 2020).

The genomics laboratory of INCIENSA prepared the sequencing libraries using the Illumina DNA Prep Kit (Illumina, San Diego, CA, USA) according to the laboratory standard operating procedure for pulsenet nextera DNA flex library preparation (https://www.cdc.gov/pulsenet/pathogens/wgs.html) as recommended by the Centers for Disease Control and Prevention (Atlanta, GA, USA). Paired-end sequencing (250 bp) was performed for each library on a MiSeq instrument using 500 cycles v2 chemistry cartridges (Illumina, San Diego, CA, USA).

Genome assembly

Analyses were performed using custom scripts.

Quality control: Raw sequencing reads were evaluated using FastQC v0.11.7 (Andrews, 2010) and MultiQC (Ewels, Magnusson, Lundin, & Käller, 2016) before and after trimming. Removal of adapters and trimming of low-quality sequences (Q<30) was done Trimmomatic v0.38 (Bolger, Lohse, & Usadel, 2014). FastQ-Screen (Wingett & Andrews, 2018) and BBDuk (http://jgi.doe.gov/data-and-tools/bb-tools/) were used to quantify and remove possible contaminant reads, respectively.

Assembly strategies: Reference-based assembly (consensus sequence) was done by mapping reads to a reference SARS-CoV-2 genome NC_045512.2 using BWA-MEM 0.7.5a-r405 (H. Li & Durbin, 2009) with default parameters. The resulting BAM file was used for the variant calling analyses.
In addition, SARS-CoV-2 *de novo* genome assembly was performed with Megahit v1.1.3 (D. Li, Liu, Luo, Sadakane, & Lam, 2015). Due to the results of both *de novo* and reference-based assemblies were consistent, the consensus sequence was used for subsequent analyses.

Assembly assessment: Assessment of genome assembly was done using the 3C criterion (contiguity, completeness, and correctness) as defined in (J.-A. Molina-Mora, Campos-Sánchez, Rodríguez, Shi, & García, 2020; J. A. Molina-Mora & Garcia, 2020). QUAST (Gurevich, Saveliev, Vyahhi, & Tesler, 2013), Qualimap (Okonechnikov, Conesa, & García-Alcalde, 2016), and BRIG (Alikhan, Petty, Ben Zakour, & Beatson, 2011) results were used for the analysis.

Clade assignment: PANGOLIN Lineages and GISAID clades were assigned based on assembled genomes using Bionumerics software v7.6.3 (https://www.applied-maths.com/) or the Coronavirus typing tool v1.13 (https://www.genomedetective.com/app/typingtool/cov/) after their upload into the GISAID database (Global Initiative on Sharing All Influenza Data, www.gisaid.org). Microreact (Argimón et al., 2016) was used to visualize each case genome according to geographic location and the cluster it belongs to.
In addition, 18 assembled SARS-CoV-2 genomes from Costa Rica (CRC-007 to CRC-024, University of Costa Rica) were retrieved from the GISAID database and included for the comparative analyses.

Comparative genomic analyses
Genome variability: Cd-hit (W. Li & Godzik, 2006) was used to cluster and compare all genome sequences by similarity. We discarded genome sequences belonging to other clusters than the one containing the reference. All the viral sequences passing the quality control step and which resulted clustered with the reference were uploaded into the GISAID database.

Variant calling analysis: BAM file from reads mapping was used to remove duplicates using Picard Markduplicates (http://broadinstitute.github.io/picard). Freebayes v1.3.1 (Garrison & Marth, 2012) was used as a variant caller with the parameters: -p 1 -q 20 -m 60 --min-coverage 10 –V. VCF_filter v3.2 (https://github.com/moskalenko/vcf_filter) was used to remove low-confidence variants. Variant annotation was achieved using SNPeff (Cingolani et al., 2012) with the annotation file of the reference SARS-CoV-2 genome NC_045512.2. To represent relevant variants, structural modeling (3D) of Spike protein was done using PDB (https://www.rcsb.org/).

Phylogenetic analysis: MAFFT v7.471 (Katoh, Misawa, Kuma, & Miyata, 2002) was used to align all genome sequences. Construction of the phylogenetic tree was done using IQ-TREE v1.6.12 (Minh et al., 2020), including ModelFinder (Kalyaanamoorthy, Minh, Wong, Von Haeseler, & Jermiin, 2017) to select the best nucleotide substitution model (using the Bayesian Information Criterion BIC, the best model was TN+F+I). The tree was visualized using iTOL tool v4 (Letunic & Bork, 2019), including information of clades and lineages (GISAID and PANGOLIN classification), key variants, epidemiological week, and geographic location (Costa Rica is administratively composed of seven provinces which are divided into 82 cantons).

RESULTS

To study the SARS-CoV-2 genomes from Costa Rican cases of COVID-19, sequencing, genome assembly, and genomic comparison were performed. Samples were collected from different geographical regions from Costa Rica from March to November 2020 (Figure 1 and Table 1), with a predominance of cases from the Central Valley region (the most highly populated region). No particular features were recognized by geographic origin at the genomic level (see below).

Variant calling analysis identified a range of 3 to 22 variants per genome. In total, we identified 283 distinct genome variants with a frequency of 2120 among all genomes. The total of distinct genome variants includes 146 (51.6%) missense (non-synonymous) mutations and 129 (45.6%) synonymous mutations. More details in Table 1.

The distribution of all the 283 distinct variants among the genomes is shown in Figure 2-A. A power-law pattern was recognized, in which scarce mutations are present in most of the sequences, and so many variants are present in very few sequences (Figure 2-B). This means that only a few variants are present in multiple genomes, and the subsequent analysis can be focused on those variants. For example, only 16 variants were present in at least 5% (9) of the sequences, and were distributed along the genome (Figure 2-C and Table 2). Also, only two variants were found in >50% sequences, and most variants (190 sequences, 67.1%) were found in a single genome. The most frequent substitutions were D614G in the Spike and P4715L in the ORF1ab-RdRp both in 98.9% (183) of the genomes. Details and other cases are presented in Table 2.

Interestingly, the second most detected variant in the Spike gene in our analyses was the T1117I, present in 29.2% of the Costa Rican genomes. As shown in Figure 3, this mutation is very scarcely reported in the world (only present in 213 sequences, i.e., 0.08% according to GISAID, December 31st 2020). In Costa Rica, the frequency during March, April, and May was 0%, in June reached 6.3% (5 out of 79 total sequences), July with 11.1 % (13 out of 117 genomes) and August with 14.5% with 20 sequences in 138 genomes. In September and October the frequency was 22.7 26.6%, respectively. To November, it reached 29.2% with 54 out of 185 genomes (Figure 3 and Supplementary Material). To better visualize the location and possible effects of this variant, a structural modeling of the Spike protein was done, as shown in Figure 4. The T1117I and D614G are located in different domains.

On the other hand, the L84S variant in the ORF8 was only found in two genomes (the same without the D614G in the Spike). These two genomes were clustered separately as part of clade S in the phylogenetic tree (Figure 5, see details below).

In addition, no structural variants in the receptor-binding domain (RBD) of the Spike protein (N439K, T481I, V483A, E484E, N501Y nor G476S) were identified. Neither of the variants has been identified in the genomic regions that are used for the viral diagnosis by RT-PCR in the E and RdRP genes according to the protocol by Corman et al. 2020 (Corman et al., 2020).

Sequence classification showed that three major clades are predominant in Costa Rica according to both GISAID and PANGOLIN classifications (Table 1 and Figure 5). For GISAID clades, G, GH, and GR represent the main groups, with 98.9% of total cases, and two sequences were found as part of the S clade (1.1%). In the PANGOLIN classification, 20 different lineages were recognized, but 71.9% of the cases are represented by B.1, B.1.1, and B.1.291. No clear geographic distribution was found for the clades along the country (Figure 1 and Supplementary Material).

During the first five weeks of the outbreak in Costa Rica (from March to April) lineages B.1 and B.1.5 were mainly identify. The epidemiological information shown that those patients traveled to Europe (France and Spain), North America (USA and Mexico) and South America (Colombia). Later on lineage B.1.1 was documented on epidemiological week 24 and circulated actively in the north region of the country. Since then it was detected in samples of patients from six out of seven provinces from June to August 2020.

Based on the variants, the sequence comparison was done to build the phylogenetic tree, in which the best substitution model was TN+F+I (Figure 5). The clades and lineages are separated in the tree, including a clear cluster for all the 54 sequences carrying the Spike-T1117I variant (as part of the clade GR). Besides, the GR clade was enriched with cases from the 23-48 epidemiological weeks 2020 (June-November), but no other temporal patterns were identified. The only two sequences belonging to the clade S were clustered together and separately from the other Costa Rican cases, as expected. Regarding the ten deaths, four of the respective genomes clustered together (CRC-110, CRC-111, CRC-112 and CRC-140). Patients belonged to a nursing home outbreak. No epidemiological linkage was found between them and CRC-101. In general, fatal cases were distributed among the clades without a clear association with a specific taxonomic group.

DISCUSSION

The analysis of genome sequences from pathogens has been recognized as a key tool in infectious disease epidemiology (van Dorp et al., 2020). Thus, to effectively combat COVID-19, epidemiological surveillance is a requirement to track circulating clades (Koyama et al., 2020).

In order to study the SARS-CoV-2 genome diversity circulating in Costa Rica, we compared 185 viral genomes. Concerning the variant calling analysis, the genome comparison against the reference genome (Wuhan-1, NC_045512.2) revealed between 3 and 22 mutations per genome, with a sequence homologies of 99.92-99.98%. Although SARS-CoV-2 is a RNA virus, the proof-reading activity of a nonstructural gene explains the relatively slow mutation rate of 1-2 mutations per month along the genome sequence (Deng et al., 2020). Interestingly, this observation is in line with our results, in which the genomes circulating in March-April 2020 reached up to 11 mutations, up to 18 in August, and up 22 in November.

In the analysis among all the genomes, 2120 mutations were identified, corresponding to 283 distinct genome variants. The last were represented mainly by missense (51.6%) and synonymous (45.6%) mutations. Similar distributions have been reported in other studies, in which the frequency of missense mutations in SARS-CoV-2 genomes is greater than synonymous mutations (Jaroszewski, Iyer, Alisoltani, Sedova, & Godzik, 2020; Koyama et al., 2020). By gene, most of the variants were identified in ORF1ab (63.6%), Spike (11.7%), and Nucleocapsid (10.6%). This predominance has also been reported in other studies (Vankadari, 2020).

Besides, the frequency of the distinct variants follows a power-law distribution. Only 16 out of the 283 variants were found in >5% of the genomes and most variants (67.1%) were found in a single genome. The most common variants were P4715L in the ORF1ab-RdRp and D614G in the Spike protein (both in the same 183 genomes, 98.9%). Similar findings have been reported in previous studies, as shown below.

Spike protein - D614G

Since the Spike glycoprotein is involved in the interaction with the host's receptor (angiotensin-converting enzyme 2, ACE2) (Zhou et al., 2020), variants in this gene are interesting because of the possible change in the dynamics of the transmission (Toyoshima et al., 2020) or vaccine effectiveness (Koyama et al., 2020). As observed also in Costa Rica (98.9%), D614G of the Spike protein is highly prevalent in the world, including America and Europe (Coppée, Lechien, Declèves, Tafforeau, & Saussez, 2020; GISAID, 2020; Mercatelli & Giorgi, 2020). According to GISAID, D614G already occurred worldwide in >90% of all samples.

Despite the relation between D614G and transmissibility, severity or mortality remains unclear (Grubaugh et al., 2020). Different studies have suggested some implications in transmission and immune response (Teng et al., 2020; Zhang et al., 2020), but analytical and epidemiological factors could explain this observation. In patients, D614G has been associated with higher viral load, but without significant changes of the disease severity (Korber et al., 2020; Plante et al., 2020) or vaccine effectiveness (Plante et al., 2020; Zhang et al., 2020).

Spike protein - T1117I

T1117I in the Spike is the second variant found in frequency in the Spike protein (54 genomes, 29.2%). In our study, this mutation co-occurs with the H145Y and L291L variants in the Nucleocapsid. Interestingly, T1117I has been scarcely reported in other latitudes (GISAID, 2020; Long et al., 2020). According to GISAID (December 31st, 2020), T1117I occurred merely 213 times worldwide (0.08% of all samples with Spike sequence in the database) in only 19 countries, including Colombia and New Zealand. Our 54 sequences represent 25.3% out of all the available sequences

with the T1117I variant worldwide. The group defines a clearly separated monophyletic cluster (see below, Figure 5). The first strain detected with this mutation worldwide was collected in March 2020 in Germany. In Costa Rica, T1117I was first reported in June 2020 in San José, and it increased in number during the June-December period (0%, 6.3%, 11.1%, 14.5%, 22.7%, 26.6% and 29.2% from May to November 2020, respectively), mainly in the Central Valley. Even though we cannot exclude a sampling bias from our data, the increasing detection of these variant through time should be further investigated using more sequences to reach a definite conclusion.

Using resolved structures of proteins from related strains, the Spike position equivalent to this mutation is involved in viral oligomerization interfaces needed for cell infection (Meher, Bhattacharjya, & Chakraborty, 2019; Walls et al., 2017; Wrapp et al., 2020). The radical replacement of threonine (T, a polar amino acid) by Isoleucine (I, non-polar) may have an impact on the Spike function. Due to the location of the T1117I variant in the Spike (Figure 3), it is not expected to induce direct changes in the interaction of the Spike (RBD) with the ACE2, nor the immunodominant region of the B-cell epitope (immune response) (Koyama et al., 2020). However, further studies are required to study the possible implications of this variant in the biology of SARS-CoV-2 and the spread in Costa Rica and the world.

ORF1ab - P4715L (RdRp - P323L)

Other studies have shown that almost all sequences with the D614G (Spike) mutation also have the variant P4715L in the ORF1ab (RdRp, P323L). In Costa Rica, all the sequences with D614G also carry the P4715L, with 98.9% out of all the genomes included here. As part of the RdRp, P4715L might affect the replication speed of the virus (Koyama et al., 2020). The relevance of variants in the non-structural proteins is related to functions related to viral replication and transcription, including genes such as RdRp (Toyoshima et al., 2020). Also, because RdRp is the target of putative anti-viral drugs (such as remdesivir and favipiravir), mutations in these regions could eventually be responsible for the emergence of drug-resistant virus (Koyama et al., 2020). However, as found for other variants, the functional significance of P4715L and other variants in the RdRp is yet unknown (Toyoshima et al., 2020).

ORF8-L84S

Only two genomes (1.1%) in this study have the variant L84S in the ORF8. L84S, one out of the two variants that define the clade S (GISAID classification), has been reported mainly for Asian countries, with a lower frequency in European and American countries (Coppée et al., 2020). This mutation was also implicated in the initial appreciation of the L and S genome types (Tang et al., 2020). However, like the other variants, this condition remains an open problem, in part due to the loss of the patterns as the available genome sequences have increased.

Nucleocapsid, GGG>AAC

In the case of the variation GGG>AAC in the positions 28881-28883, the triplet resulted present in 77 (41.6%) genomes. This complex affects the Nucleocapsid protein with two amino acid changes and it has been reported in multiple latitudes (GISAID, 2020; Maitra et al., 2020; Osório & Correia-Neves, 2020). This complex variation was reported at the beginning of the binding site of the forward primer used in the RT-qPCR assays of the Chinese National Institute for Viral Disease Control and Prevention (Osório & Correia-Neves, 2020), however, this is not part of the primers used in other protocols, including those used in this study or any Costa Rican diagnostic laboratories.

Other variants

Recently, a variant N501Y in the Spike has been reported in the United Kingdom, which is located in the RBD and the impact on the transmission or severity is not known (Kemp et al., 2020; Wise, 2020). From the Costa Rican sequenced cases, none harbors yet this mutation, neither the variant 501Y.V2 described in South Africa (Tegally et al., 2020). Also, a European cluster (called 20A.EU1) has been reported with a significant spread in summer 2020, mainly in Spain (Hodcroft et al., 2020). Genome sequences in this group have at least six variants, including the mutation A222V in the Spike protein and A220V in the nucleoprotein. None of the A222V in the Spike and the A220V in the nucleoprotein were found in our study. Nevertheless, local continued genomic surveillance is in place to monitor these variants.

Regarding the RT-qPCR test for the diagnosis of COVID-19, none of the 283 variants were part of the genomic regions of the binding site of the primers used in the Cormans et al. protocols (Corman et al., 2020) implemented in Costa Rica for detecting E and RdRp genes of SARS-CoV-2.

Phylogeny and Integration

Altogether, the variant calling analysis reveals that the patterns found in Costa Rica, with 283 distinct mutations, are similar to those observed worldwide (such as D614G in the Spike and L84S in the ORF8), except for the T1117I in the Spike. In general, these mutations define the genome groups, in which three GISAID clades and three PANGOLIN lineages were found to predominate in Costa Rica. The 54 sequences with the T1117I variant defines a separated cluster with a monophyletic origin, i.e. they have a common ancestor.

The evolutionary analysis of the SARS-CoV-2 genomes from Costa Rican cases suggests the possibility of multiple introductions into the Costa Rican population, as it has been reported worldwide (Koyama et al., 2020; van Dorp et al., 2020). In addition, the sequences from the first months of the pandemic grouped in lineage B.1 and B.1.5 mainly, suggesting low undetected circulation and re-introductions of new lineages not detected in the country during early stages of the pandemic due to the extreme lockdown measures. Although travel records and some contact tracing data are available for many cases, the reconstruction of the possible routes for the spread of particular clades is difficult in this case due we only have sequenced 185 out of the 169 321 (to December 31st, 2020) Costa Rican cases of COVID-19 (0.1%).

Concerning deaths, genomes analyzed from ten dead patients were distributed among the clades without a clear association pattern for a specific group. This distribution is similar to that found around the world in which, until now, no mutation or clade has been clearly associated with a higher fatality rate (Grubaugh et al., 2020; Vankadari, 2020). In the phylogenetic three, a cluster with four sequences of dead patients can be highlighted. It was related to an outbreak occurred in a nursing home for the elderly.

Finally, limitations of our study include a low number of samples, the non-randomized sampling used for the selection of the samples (sampling bias), incomplete clinical and epidemiological data such as travel records and contact tracing for all sequenced samples, and lack of sequenced genome information from key neighboring countries with a strong immigration practice to Costa Rica. Nevertheless this work adds to the efforts of the global scientific community to produce and publicly share SARS-CoV-2 genome sequences. The analyses reported here contribute to the monitoring of the spread of SARS-CoV-2 as part of the surveillance programs during the pandemic, and suggest that surveillance should be continued and (financially-) supported to keep track of important changes in the SARS-CoV-2 genome.

CONCLUSION

In summary, the phylogenetic and variant patterns are similar to those observed worldwide, such as the high prevalence of variant D614G in the Spike and low frequency of L84S in the ORF8, as well as the arrangement by clades. Interestingly, the variant T1117I in the Spike has increased its frequency along time from March to November 2020. This contrasts with the frequency of 0.08% in the world, in which 54 (25.3%) out of all the only 213 sequences in the world are from Costa Rica. Taken together, at the beginning of the pandemic few lineages with few mutations circulated in our country. Restrictions were lifted and borders were opened therefore diversity of lineages and mutations increased. The Spike variant T1117I must be closely surveyed to see if it will become the predominant circulating variant in our country. Though this variant putative biological and clinical relevance is still unknown.

In addition, none of the variants was found in genomic regions used for diagnosis with the protocols implemented in Costa Rica. Until now, changes in severity, deaths, or transmission have not been recognized for specific mutations or clades of SARS-CoV-2 in the world. Nevertheless, further genome sequencing analyses are required to study the impact of the genomic features on the biology of the virus, transmissibility and spread, clinical outcome, and treatment/vaccine effectiveness.

Acknowledgments

We are grateful to clinicians, microbiologists, and other personnel of public (Caja Costarricense de Seguro Social CCSS) and private clinical laboratories for the samples of confirmed cases of COVID-19. We also thank Carlos Mora Garro for his participation as assistant, and all members of CIET-University of Costa Rica, INCIENSA and CCSS for their logistic and financial support in the activities associated with the project.

Author contributions

J.M.M., F.D.M., H.B. and C.S.G. participated in the conception and design of the study. H.B, C.S.G, C.P.C. and COINGESA-CR members were involved in sample processing for diagnosis of COVID-19. F.D.M., E.C.L., A.G and M.C.O. sequenced the genomes. J.M.M. implemented and standardized the bioinformatics pipelines. J.M.M., F.D.M., A.M.S., E.C.A. and J.F.D. participated in the interpretation of results. J.M.M. drafted the manuscript. All authors reviewed and approved the final manuscript.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Genome sequences were uploaded to the GISAID database. All the IDs and related information are available in the “Supplementary Material”.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Funding
This work was funded by Instituto Costarricense de Investigación y Enseñanza en Nutrición y Salud (INCIENSA) and Vicerrectoría de Investigación – Universidad de Costa Rica, with the Project "C0196 Protocolo bioinformático y de inteligencia artificial para el apoyo de la vigilancia epidemiológica basada en laboratorio del virus SARS-CoV-2 mediante la identificación de patrones genómicos y clínico-demográficos en Costa Rica (2020-2022)".

LEGENDS OF FIGURES AND TABLES

Figure 1. Dynamic, geographic and temporal distribution of SARS-CoV-2 genomes from Costa Rican cases. (A) An exponential increment of COVID-19 cases has been reported in Costa Rica since March 2020, with a similar profile for reported deaths (B). Samples for sequencing were obtained from the whole country, mainly from the Central valley, which harbors the most populated region of the country (C). Image in (C) was obtained from Microreact tool (https://microreact.org/project/r7tcnUYgWMRJ5Fdssvv7VZ).

Figure 2. Variant calling analysis of SARS-CoV-2 genomes from Costa Rican cases of COVID-19. (A) Presence/absence of 283 different variants among 185 genomes. A few variants are widely distributed among genomes (*F= Frequency), and many variants are uniquely present in a single genome. (B) Distribution and accumulative percentage of variant frequency among genomes. Most variants are low frequency mutations and only 16 variants are present in at least 5% (9) genomes. The most frequent variants (Spike D614G and ORF1a P4715L) are present in 183 (98.9%) genomes (arrow). The 16 variants are distributed along the SARS-CoV-2 genome, as shown in (C).

Figure 3. Frequency of T1117I in the spike along time in Costa Rica and the world. A notorious increment of the mutation has been reported in Costa Rica since May 2020, contrasting with the prevalence around the world which keeps relatively constant and low (A-B). Map in (B) was obtained from GISAID database.

Figure 4. Structural modeling of spike protein of SARS-CoV-2. The variant D614G is present in 98.6% of the genomes in this study, which is also predominant worldwide (>90%, GISAID). D614G could affect the interaction with the host, as well as the immune response (vaccines), but real effects remains unclear. The variant T1117I is a variant very scarcely reported in the world (0.08%, GISAID), but the frequency in Costa Rica is 29.2%. The possible effect of this variant on the function of the spike is unknown.

Figure 5. Phylogenetic tree of SARS-CoV-2 genomes circulating in Costa Rica. Three GISAID clades (G, GH and GR) and three Pangolin lineages (B.1, B.1.1 and B.1.291) are dominant in Costa Rican cases. Deaths are similarly distributed along clades. Variant T1117I in the spike is present in 54 genomes, which belong to a clearly separated monophyletic cluster (dark-blue). Other variants are presented with different colors.

Table 1. Demographic information, variants and groups of the genomic analysis of SARS-CoV-2 genomes from Costa Rican cases.

Table 2. Variants of SARS-CoV-2 genomes observed in more than 5% (9) genomes from Costa Rican cases of COVID-19.

Supplementary Material.xlsx

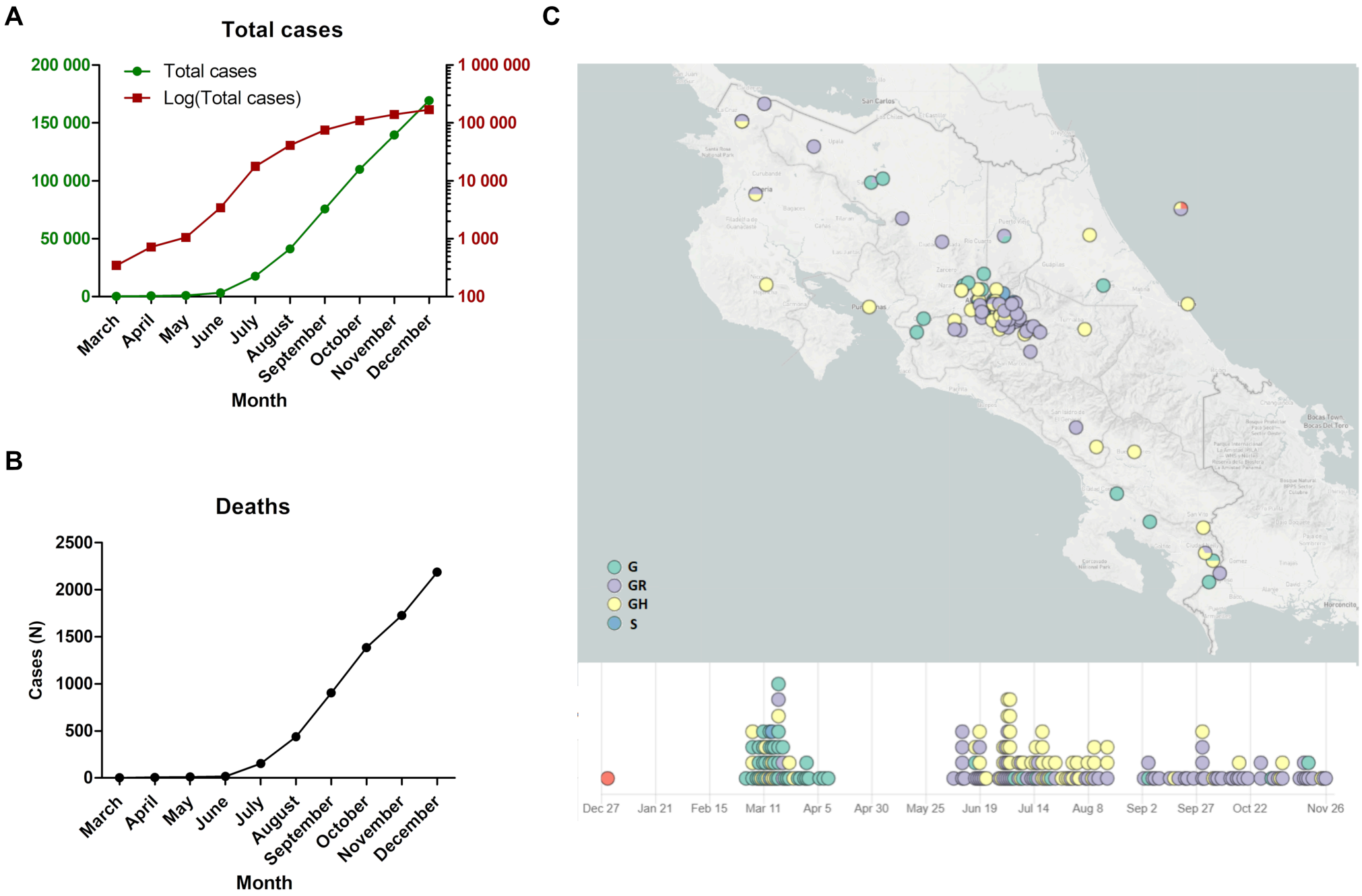
A
Total cases
Total cases
Log(Total cases)
200 000
150 000
100 000
50 000
0
1 000 000
100 000
10 000
1 000
100
March
April
May
June
July
August
September
October
November
December
Month
B
Deaths
Cases (N)
2500
2000
1500
1000
500
0
March
April
May
June
July
August
September
October
November
December
Month
C
G
GR
GH
S
Dec 27
Jan 21
Feb 15
Mar 11
Apr 5
Apr 30
May 25
Jun 19
Jul 14
Aug 8
Sep 2
Sep 27
Oct 22
Nov 26

Figure 1

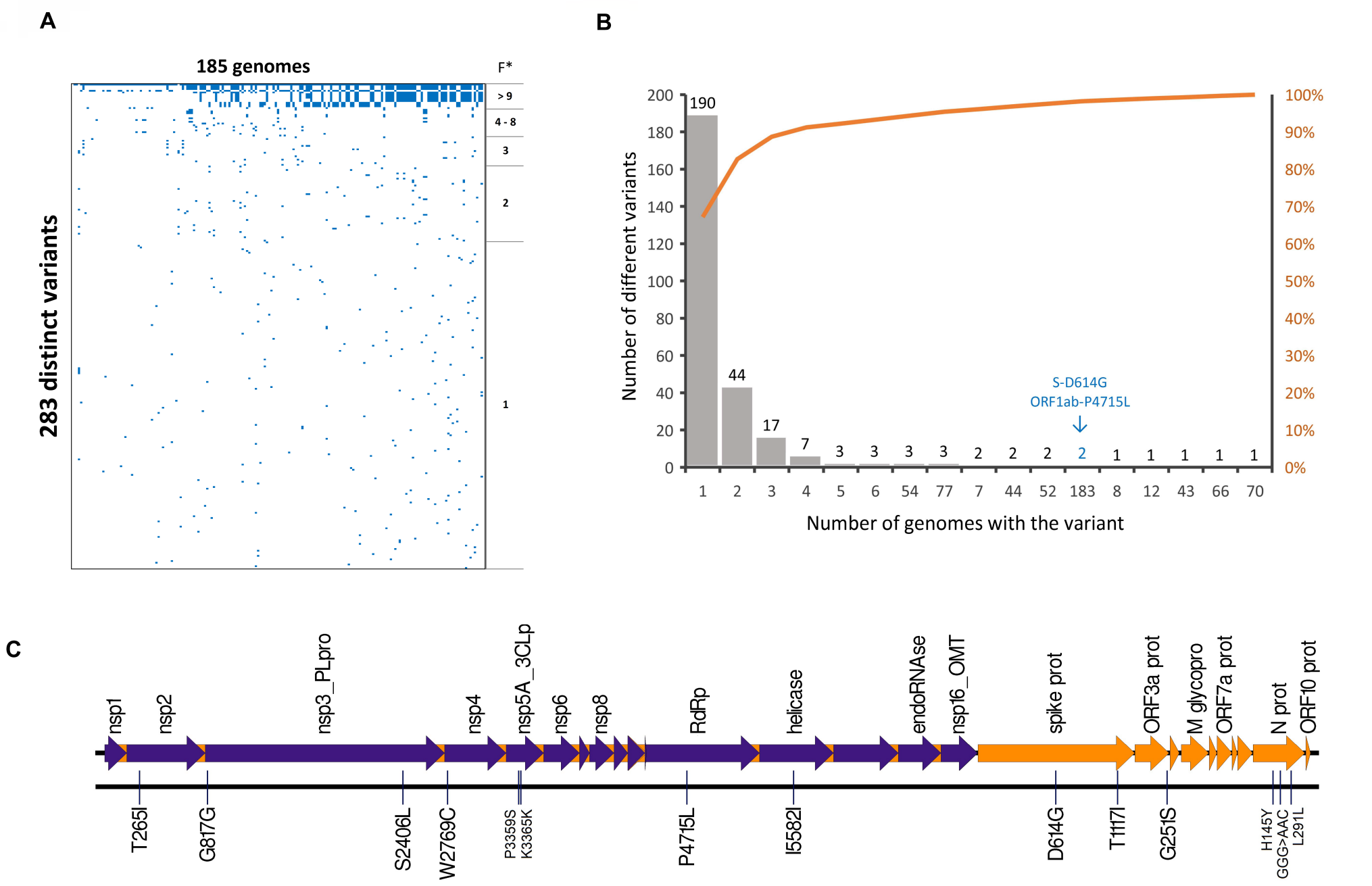
A
185 genomes
F*
> 9
4 - 8
3
2
1
283 distinct variants
B
Number of different variants
Number of genomes with the variant
190
44
17
7
3
3
3
3
2
2
2
2
1
1
1
1
1
1
2
3
4
5
6
54
77
7
44
52
183
8
12
43
66
70
S-D614G
ORF1ab-P4715L
C
nsp1
nsp2
nsp3_PLpro
nsp4
nsp5A_3CLp
nsp6
nsp8
RdRp
helicase
endoRNAse
nsp16_OMT
spike prot
ORF3a prot
M glycopro
ORF7a prot
N prot
ORF10 prot
T265I
G817G
S2406L
W2769C
P3359S
K3365K
P4715L
I5582I
D614G
T1117I
G251S
H145Y
GGG>AAC
L291L

Figure 2.

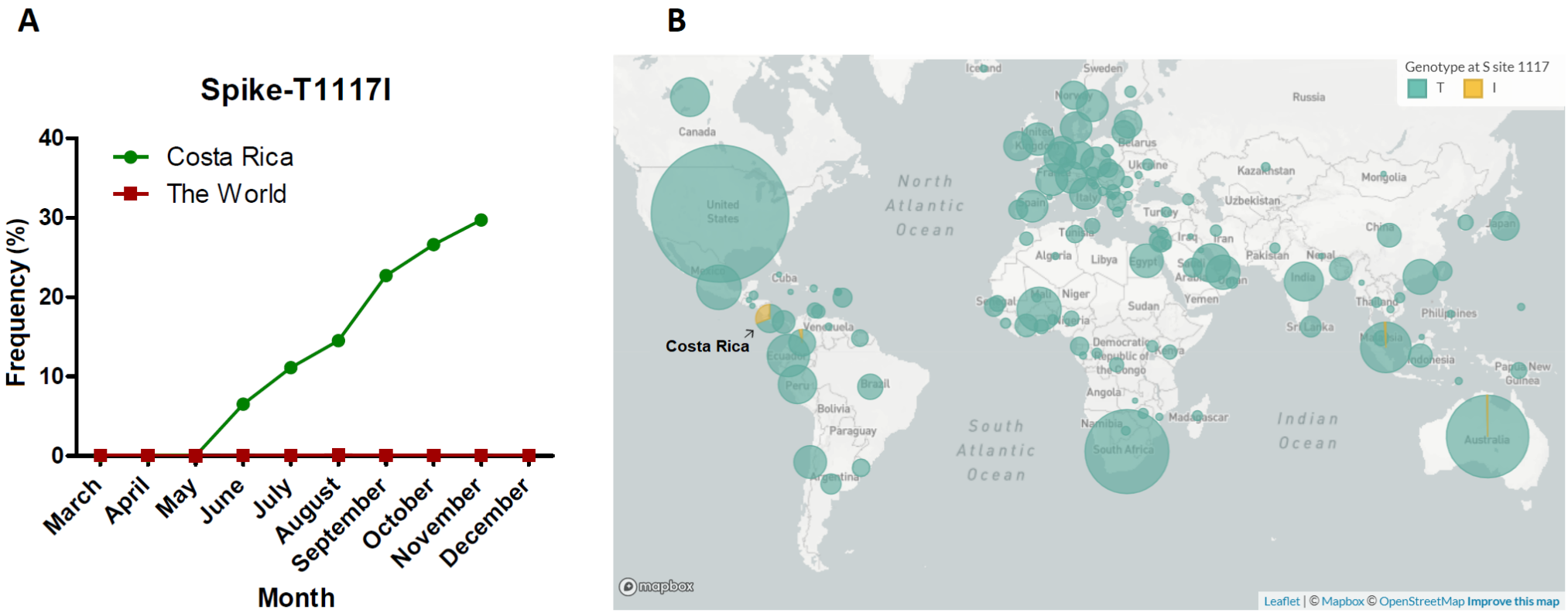
A
Spike-T1117I
Costa Rica
The World
Frequency (%)
40
30
20
10
0
March
April
May
June
July
August
September
October
November
December
Month
B
Genotype at S site 1117
T
I
Costa Rica
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Canada
United States
Mexico
Cuba
Venezuela
Ecuador
Peru
Brazil
Bolivia
Paraguay
Argentina
Iceland
Sweden
Norway
Russia
Belarus
Ukraine
Kazakhstan
Mongolia
Uzbekistan
Turkey
Iran
Iraq
China
Japan
Pakistan
Nepal
India
Sri Lanka
Thailand
Philippines
Indonesia
Papua New Guinea
Australia
Madagascar
Angola
Namibia
South Africa
Democratic Republic of the Congo
Kenya
Sudan
Yemen
Niger
Mali
Algeria
Libya
Egypt
Tunisia
Spain
Italy
France
United Kingdom
mapbox
Leaflet | © Mapbox © OpenStreetMap Improve this map

Figure 3.

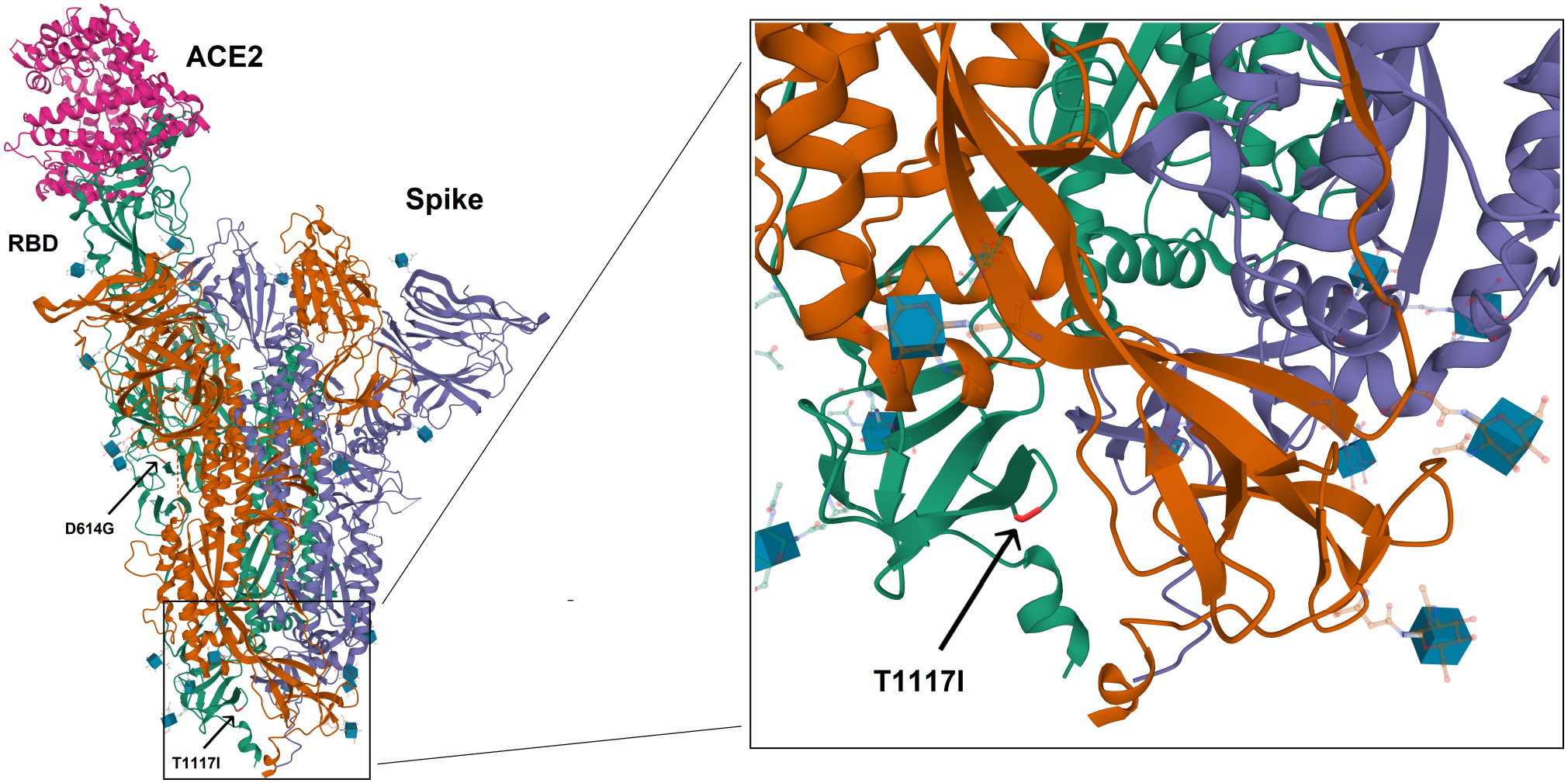

Figure 4.

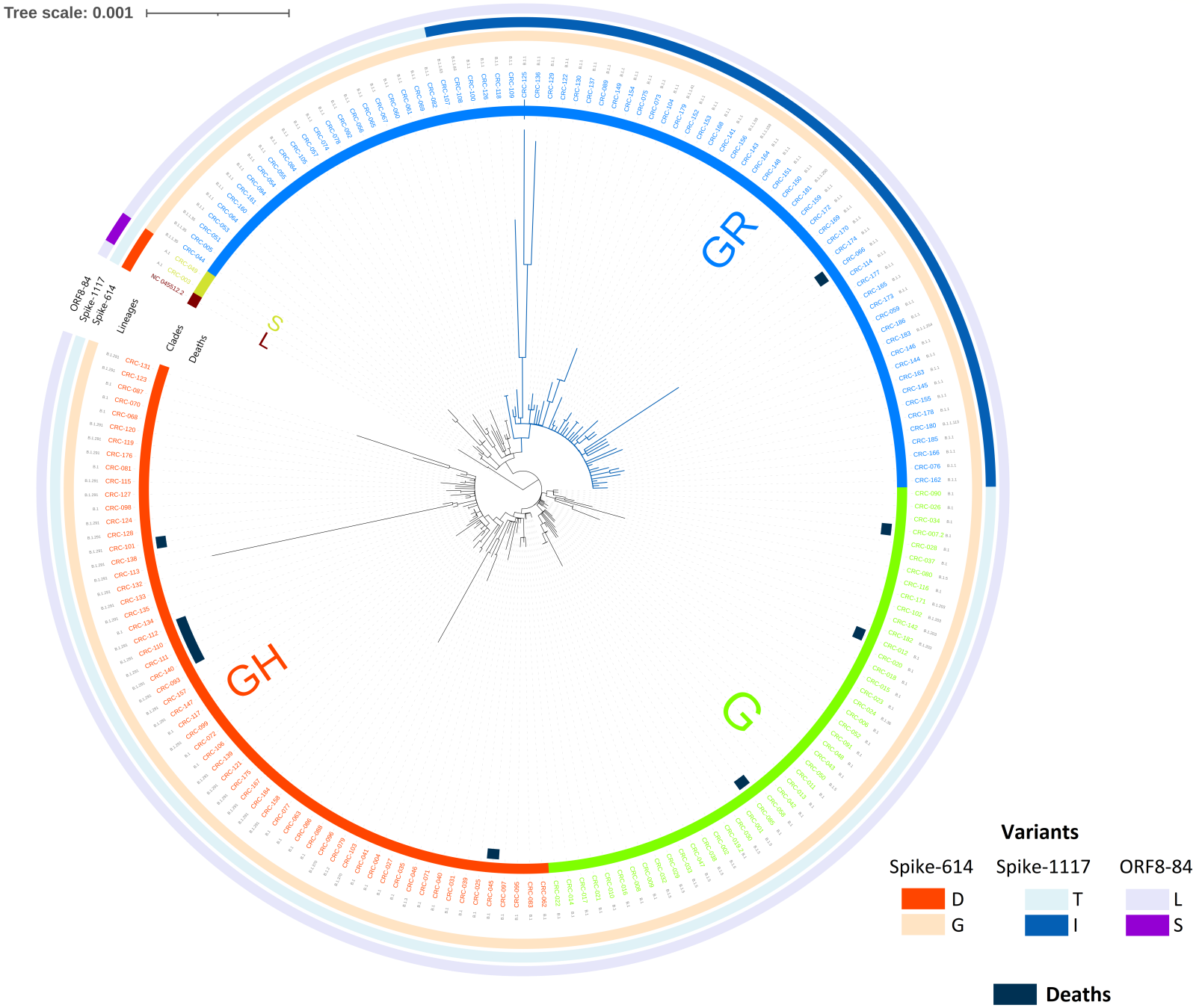
Tree scale: 0.001
ORF8-84
Spike-1117
Spike-614
Lineages
Clades
Deaths
NC 045512.2
GR
G
GH
L
S
Variants
Spike-614
D
G
Spike-1117
T
I
ORF8-84
L
S
Deaths

Figure 5.

Table 1. Demographic information and basic results of the genome analysis of SARS-CoV-2 from Costa Rican cases.

Parameter	Counts	
Demographic data	***Patients*** Total cases: 185 ***Sex*** Male: 112 Female: 71 Unknown: 2 Mortality (known cases): 10	***Location (provinces)*** San José: 82 Alajuela: 36 Heredia: 16 Cartago: 20 Puntarenas: 13 Guanacaste: 9 Limón: 6 Other: 3
Genome sequencing	***Genome sequencing*** Total genomes: 185	***Genome coverage depth*** 900-7000X
Total variants for all genomes	2120, including repeated mutations among genomes	
Total distinct variants	283 variants, regardless frequency among genomes	
Variability	Reference genome size: 29 903 bp Variants (min-max) 3-22 per genome Identity: 99.92 - 99.98%	
Classification of variants	***By gene*** ORF1ab: 180 (63.6%) ORF3a: 11 (3.9%) ORF6: 1 (0.4%) ORF7a: 1 (0.4%) ORF8: 6 (2.1%) S-Spike: 33 (11.7%) E: 5 (1.8%) M: 9 (3.2%) N: 30 (10.6%) ORF10: 7 (2.5%)	***By effect*** Missense: 146 (51.6%) Synonymous: 129 (45.6%) Splice region: 1 (0.3%) STOP gained: 3 (1.1%) Upstream: 4 (1.4%) ***By frequency*** In only 1 genome: 190 (67.1%) At least 5% (9) genomes: 16 (5.7%) At least 25% (46) genomes: 12 (4.2%) At least 50% (92) genomes: 2 (0.7%)
Genome sequence groups: clades and lineages	***GISAID clades*** G: 46 (24.9%) GH: 60 (32.4%) GR: 77 (41.6%) S: 2 (1.1%) V: 0 L: 0	***PANGOLIN lineages*** A.1: 2 (1.1%) B.1: 28 (15.1%) B.1.1: 64 (34.6%) B.1.203: 10 (5.4%) B.1.291: 41 (22.2%) B.1.5: 21 (11.3%) Others: 19 (10.3%)

Table 2. Variants of SARS-CoV-2 genomes observed in more than 5% (9) genomes from Costa Rican cases of COVID-19.

Position genome	Ref	Alt	Gene	Position - cDNA	Position - protein	Type	Frequency (N)	Frequency (%)
14408	C	T	ORF1ab - RdRp	c.14144C>T	p.Pro4715Leu (P4715L)	missense	183	98.9
23403	A	G	Spike - Surface glycoprotein	c.1841A>G	p.Asp614Gly (D614G)	missense	183	98.9
28881	G	A	N - Nucleocapsid phosphoprotein	c.608G>A	p.Arg203Lys (R203K)	missense	77	41.6
28882	G	A	N - Nucleocapsid phosphoprotein	c.609G>A	p.Arg203Arg (R203R)	synonymous	77	41.6
28883	G	C	N - Nucleocapsid phosphoprotein	c.610G>C	p.Gly204Arg (G204R)	missense	77	41.6
1059	C	T	ORF1ab - NSP2	c.794C>T	p.Thr265Ile (T265I)	missense	70	37.8
10360	G	A	ORF1ab - 3C-like proteinase	c.10095G>A	p.Lys3365Lys (K3365K)	synonymous	66	35.7
24912	**C**	**T**	**Spike - Surface glycoprotein**	**c.3350C>T**	**p.Thr1117Ile (T1117I)**	**missense**	**54**	**29.2**
28706	C	T	N - Nucleocapsid phosphoprotein	c.433C>T	p.His145Tyr (H145Y)	missense	54	29.2
29144	C	T	N - Nucleocapsid phosphoprotein	c.871C>T	p.Leu291Leu (L291L)	synonymous	54	29.2
8572	G	T	ORF1ab - NSP4	c.8307G>T	p.Trp2769Cys (W2769C)	missense	52	28.1
10340	C	T	ORF1ab - 3C-like proteinase	c.10075C>T	p.Pro3359Ser (P3359S)	missense	52	28.1
2716	C	T	ORF1ab - NSP2	c.2451C>T	p.Gly817Gly (G817G)	synonymous	44	23.8
17010	C	T	ORF1ab - helicase	c.16746C>T	p.Ile5582Ile (I5582I)	synonymous	44	23.8
7482	C	T	ORF1ab - NSP3	c.7217C>T	p.Ser2406Leu (S2406L)	missense	43	23.2
26143	G	A	ORF3a	c.751G>A	p.Gly251Ser (G251S)	missense	12	6.5

REFERENCES

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, *12*(1), 402. https://doi.org/10.1186/1471-2164-12-402

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved April 10, 2018, from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., … Aanensen, D. M. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics*, *2*(11), e000093. https://doi.org/10.1099/mgen.0.000093

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Castillo, A. E., Parra, B., Tapia, P., Acevedo, A., Lagos, J., Andrade, W., … Fernández, J. (2020). Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *Journal of Medical Virology*, *92*(9), 1562–1566. https://doi.org/10.1002/jmv.25797

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Coppée, F., Lechien, J. R., Declèves, A. E., Tafforeau, L., & Saussez, S. (2020). Severe acute respiratory syndrome coronavirus 2: virus mutations in specific European populations. *New Microbes and New Infections*, *36*, 100696. https://doi.org/10.1016/j.nmni.2020.100696

Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., … Drosten, C. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance*, *25*(3), 2000045. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045

Deng, X., Gu, W., Federman, S., du Plessis, L., Pybus, O. G., Faria, N. R., … Chiu, C. Y. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science*, *369*(6503), 582–587. https://doi.org/10.1126/science.abb9263

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv:1207.3907 [q-Bio.GN]*. Retrieved from http://arxiv.org/abs/1207.3907

GISAID. (2020). GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses. Retrieved November 18, 2020, from https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/

Grubaugh, N. D., Hanage, W. P., & Rasmussen, A. L. (2020). Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear. *Cell*, *182*(4), 794–795. https://doi.org/10.1016/j.cell.2020.06.040

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hodcroft, E. B., Zuber, M., Nadeau, S., Comas, I., González Candelas, F., consortium, S.-S., … Neher, R. A. (2020). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *MedRxiv*, *2020*(October), 2020.10.25.20219063. Retrieved from

https://doi.org/10.1101/2020.10.25.20219063
INCIENSA, & Ministerio de Salud. (2020). Inciensa logra secuenciar el genoma completo del nuevo coronavirus SARS-CoV-2 (COVID-19). Retrieved May 21, 2020, from https://www.ministeriodesalud.go.cr/index.php/centro-de-prensa/noticias/741-noticias-2020/1642-inciensa-logra-secuenciar-el-genoma-completo-del-nuevo-coronavirus-sars-cov-2-covid-19
Jaroszewski, L., Iyer, M., Alisoltani, A., Sedova, M., & Godzik, A. (2020). The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. *BioRxiv*. https://doi.org/10.1101/2020.08.10.244756
Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285
Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436
Kemp, S., Datir, R., Collier, D., Ferreira, I., Carabelli, A., Harvey, W., … Gupta, R. (2020). Recurrent emergence and transmission of a SARS-CoV-2 Spike deletion ΔH69/ΔV70. *BioRxiv*, 2020.12.14.422555. https://doi.org/10.1101/2020.12.14.422555
Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., … Montefiori, D. C. (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*, *182*(4), 812-827.e19. https://doi.org/10.1016/j.cell.2020.06.043
Koyama, T., Platt, D., & Parida, L. (2020). Variant analysis of SARS-cov-2 genomes. *Bulletin of the World Health Organization*, *98*(7), 495–504. https://doi.org/10.2471/BLT.20.253591
Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239
Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/bioinformatics/btv033
Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158
Long, S. W., Olsen, R. J., Christensen, P. A., Bernard, D. W., Davis, J. J., Shukla, M., … Musser, J. M. (2020). Molecular architecture of early dissemination and massive second wave of the SARS-CoV-2 virus in a major metropolitan area. *MBio*, *11*(6), 1–30. https://doi.org/10.1128/mBio.02707-20
LoPresti, M., Beck, D. B., Duggal, P., Cummings, D. A. T., & Solomon, B. D. (2020, September 3). The Role of Host Genetic Factors in Coronavirus Susceptibility: Review of Animal and Systematic Review of Human Literature. *American Journal of Human Genetics*. Cell Press. https://doi.org/10.1016/j.ajhg.2020.08.007
Lu, H., Stratton, C. W., & Tang, Y. W. (2020, April 1). Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology*. John Wiley and Sons Inc. https://doi.org/10.1002/jmv.25678
Maitra, A., Sarkar, M. C., Raheja, H., Biswas, N. K., Chakraborti, S., Singh, A. K., … Das, S. (2020). Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the

ongoing outbreak in India and impact on viral structure and host susceptibility. *Journal of Biosciences*, *45*(1), 76. https://doi.org/10.1007/s12038-020-00046-1

Meher, G., Bhattacharjya, S., & Chakraborty, H. (2019). Membrane Cholesterol Modulates Oligomeric Status and Peptide-Membrane Interaction of Severe Acute Respiratory Syndrome Coronavirus Fusion Peptide. *Journal of Physical Chemistry B*, *123*(50), 10654–10662. https://doi.org/10.1021/acs.jpcb.9b08455

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Frontiers in Microbiology*, *11*, 1800. https://doi.org/10.3389/fmicb.2020.01800

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., … Teeling, E. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Molina-Mora, J.-A., Campos-Sánchez, R., Rodríguez, C., Shi, L., & García, F. (2020). High quality 3C de novo assembly and annotation of a multidrug resistant ST-111 Pseudomonas aeruginosa genome: Benchmark of hybrid and non-hybrid assemblers. *Scientific Reports*, *10*(1), 1392. https://doi.org/10.1038/s41598-020-58319-6

Molina-Mora, J. A., & Garcia, F. (2020). The 3C criterion: Contiguity, Completeness and Correctness to assess de novo genome assemblies. *BMC Bioinformatics, Bioinformatics: From Algorithms to Applications*, *21*(S20: O7), 5. https://doi.org/10.1186/s12859-020-03838-2

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *32*(2), 292–294. https://doi.org/10.1093/bioinformatics/btv566

Osório, N. S., & Correia-Neves, M. (2020). Implication of SARS-CoV-2 evolution in the sensitivity of RT-qPCR diagnostic assays. *The Lancet Infectious Diseases*. Lancet Publishing Group. https://doi.org/10.1016/S1473-3099(20)30435-7

PAHO, P. A. H. O. (2020). *Laboratory Guidelines for the Detection and Diagnosis of COVID-19 Virus Infection*. PAHO. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-

Plante, J. A., Liu, Y., Liu, J., Xia, H., Johnson, B. A., Lokugamage, K. G., … Shi, P. Y. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*. https://doi.org/10.1038/s41586-020-2895-3

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., … Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Resende, P. C., Motta, F. C., Roy, S., Appolinario, L., Fabri, A., Xavier, J., … Siqueira, M. (2020). SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. *BioRxiv*, 2020.04.30.069039. https://doi.org/10.1101/2020.04.30.069039

Sironi, M., Hasnain, S. E., Rosenthal, B., Phan, T., Luciani, F., Shaw, M. A., … González-Candelas, F. (2020). SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective. *Infection, Genetics and Evolution*, *84*(May), 104384. https://doi.org/10.1016/j.meegid.2020.104384

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., … Lu, J. (2020). On the origin and continuing evolution of SARS-CoV-2. *National Science Review*. https://doi.org/10.1093/nsr/nwaa036

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., … De Oliveira, T. (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *Arghavan Alisoltani-Dehkordi*, *10*, 2020.12.21.20248640.

https://doi.org/10.1101/2020.12.21.20248640
Teng, S., Sobitan, A., Rhoades, R., Liu, D., & Tang, Q. (2020). Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. *Briefings in Bioinformatics*, *2020*(00), 1–15. https://doi.org/10.1093/bib/bbaa233
Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., & Kiyotani, K. (2020). SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *Journal of Human Genetics*, *65*(12), 1075–1082. https://doi.org/10.1038/s10038-020-0808-9
van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., … Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, *83*(April), 104351. https://doi.org/10.1016/j.meegid.2020.104351
Vankadari, N. (2020). Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. *Gene*, *752*(April), 1–3. https://doi.org/10.1016/j.gene.2020.144792
Walls, A. C., Tortorici, M. A., Snijder, J., Xiong, X., Bosch, B. J., Rey, F. A., & Veesler, D. (2017). Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), 11157–11162. https://doi.org/10.1073/pnas.1708727114
Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*, 1338. https://doi.org/10.12688/f1000research.15931.2
Wise, J. (2020). COVID-19: New coronavirus variant is identified in UK. *BMJ*, *371*, m4857. https://doi.org/10.1136/bmj.m4857
Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C. L., Abiona, O., … McLellan, J. S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, *367*(6483), 1260–1263. https://doi.org/10.1126/science.aax0902
Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., … Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3
Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., … Choe, H. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2020.06.12.148726
Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., … Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7