The evolutionary landscape of SARS-CoV-2 variant B.1.1.519 and its clinical impact in Mexico City

Alberto Cedro-Tanda[1+], Laura Gómez-Romero[1+], Nicolás Alcaraz[1], Guillermo de Anda-Jauregui[1], Fernando Peñaloza[1], Bernardo Moreno[1], Marco A. Escobar-Arrazola[5], Oscar A. Ramirez-Vega[5], Mireya Cisneros-Villanueva[1], Jose L. Moreno-Camacho[7], Jorge Rodriguez-Gallegos[7,8], Marco A. Luna-Ruiz Esparza[6], Miguel A. Fernández Rojas[6], Alfredo Mendoza-Vargas[1], Juan Pablo Reyes-Grajeda[1], Abraham Campos-Romero[6] Ofelia Angulo[2], Rosaura Ruiz[2], Claudia Sheinbaum[3], José Sifuentes-Osornio[4], David Kershenobich[4], Alfredo Hidalgo-Miranda[1]*, Luis A. Herrera[1,5]*

[1] Instituto Nacional de Medicina Genómica, INMEGEN, Mexico City, Mexico.

[2] Secretaría de Educación, Ciencia, Tecnología e Innovacion, Mexico City, Mexico.

[3] Gobierno de la Ciudad de México.

[4] Instituto Nacional de Ciencias Médicas y Nutrición Salvadro Zubirán, Mexico City, Mexico.

[5] Unidad de Investigación Biomédica en Cáncer, Instituto Nacional de Cancerología-Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico.

[6] Innovation and Research Department, Salud Digna, Culiacan, 80000, Sinaloa, Mexico.

[7] Clinical Laboratory Division, Salud Digna, Culiacan, 80000, Sinaloa, Mexico.

[8] Molecular Biology Laboratory, National Reference Center, Salud Digna, Tlalnepantla de Baz, 54075, Estado de Mexico, Mexico.

[+]Co-first authorship.

*Corresponding authors: LAH: lherrera@inmegen.gob.mx; AHM: ahidalgo@inmegen.gob.mx

Abstract

The SARS-CoV-2 pandemic is one of the most concerning health problems around the globe. We report the emergence of SARS-CoV-2 variant B.1.1.519 in Mexico City. This variant represented up to 90% of sequenced cases in February 2021. It is characterized by three amino acid changes in the spike protein: T478K, P681H, and T732A. We report the effective reproduction number of B.1.1.519 and present evidence of its geographical origin based on phylogenetic analysis. We also studied its evolution via haplotype analysis and identified the most recurrent haplotypes. Finally, we studied the clinical impact of B.1.1.519: patients infected with variant B.1.1.519 showed a highly significant adjusted odds ratio (aOR) increase of 1.85 over non-B.1.1.519 patients for developing a severe/critical outcome (P = 0.000296, 1.33–2.6 95% CI) and a 2.35-fold increase for hospitalization (P = 0.005, 1.32–4.34 95% CI). The continuous monitoring of this and other variants will be required to control the ongoing pandemic as it evolves.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the etiological cause of Coronavirus Disease 19 (COVID-19), and it has caused the largest and most severe pandemic of this century[1].

Since December 2019, scientists around the world have generated 2.32 M whole-genome sequences of SARS-CoV-2 that have been made publicly available in the Global Initiative on Sharing all Influenza Data (GISAID) initiative database[2]. This massive genome sequencing effort has had an impact on public health and the handling of the pandemic, since it has allowed the design and updating of molecular tests for viral detection[3,4] and guided the design of vaccines and antiviral treatments[5,6]. Moreover, it has enabled the study of viral evolution, with in-depth investigation into the emergence and pursuit of variants of concern (VOCs), such as Alpha, Beta, Gamma, and Delta, and variants of interest (VOIs), such as Eta, Lambda, Iota and Kappa[7].

Monitoring the emergence of new variants of SARS-CoV-2 is a worldwide priority, as alterations such as amino acid substitutions in the viral genome could be related to alterations in biological properties, such as the ligand-like affinity receptor, the neutralization efficiency resulting from naturally acquired immunity or vaccination[8–10], or the transmission capacity[11], as well as the impact on the clinical presentation of the disease[12].

In Mexico, SARS-CoV-2 variants have been monitored since March 2020. During the third epidemic peak between February and March 2021, we observed an increase in variant

B.1.1.519, which possesses three substitutions in the spike protein (T478K, P681H, and T732A). This study reports the emergence and spread of the new B.1.1.519 variant in Mexico City and its evolution, transmissibility and association with relevant clinical traits.

Methods

Sample preparation

Participants. Nasopharyngeal swabs (NPSs) were collected from 1,835 patients for SARS-CoV-2 detection. The study was approved by the ethics and research committees of the Instituto Nacional de Medicina Genómica (CEI/1479/20 and CEI 2020/21).

Sample collection. NPSs were collected by a trained clinician with a flexible nylon swab that was inserted into the patient's nostrils to reach the posterior nasopharynx. It was left in place for several seconds and slowly removed while rotating. The swab was then placed in 3 mL of sterile viral transport medium. Swabs from both nostrils were deposited in a single viral transport tube, taken to a clinical laboratory and processed immediately.

SARS-CoV-2 RNA extraction. Total nucleic acid was extracted from 300 µL of viral transport medium from the NPSs or 300 µL of whole saliva using the MagMAX Viral/Pathogen Nucleic Acid Isolation Kit (Thermo Fisher Scientific) and eluted into 75 µL of elution buffer.

RT-qPCR. For SARS-CoV-2 RNA detection, 5 µL of RNA template was tested using the US CDC real-time RT-qPCR primer/probe sets for 2019-nCoV_N1 and 2019-nCoV_N2 and human RNase P (RP) as an extraction control. Samples were classified as positive for SARS-CoV-2 when both the N1 and N2 primer/probe sets were detected with a Ct value lower than 40 (Centers for Disease Control and Prevention, 2020). If only one of these genes was detected, the sample was labeled inconclusive. Also, RT-PCR was performed using the DA-930-Detection Kit for 2019 Novel Coronavirus (2019-nCoV) RNA (PCR-Fluorescence Probing) (DaAn Gene Co., Ltd. Of Sun Yat-sen University, Guangzhou, China) following the manufacturer´s instructions. All tests were run on Thermo Fisher ABI QuantStudio 5 or QuantStudio 7 real-time thermal cyclers. Samples were selected for inclusion in this study based on viral Ct <30.

Sequencing

Oxford Nanopore-Sequencing. We performed PCR tiling of the COVID-19 virus, version PTC_9096_v109_revF_06Feb2020. For nanopore amplicon sequencing of SARS-CoV-2, the ARTIC v3 amplification products of each sample were mixed and purified by using Agencourt AMPure XP beads at a 1:1 ratio and finally diluted in 30 µL of water. One microliter of purified DNA amplicons was used for quantification by Qubit (Qubit™ dsDNA HS Assay Kit). Sequencing library preparation consisted of two steps: native barcode ligation and sequencing adapter ligation. Native barcoding of amplicons was performed in a 20 µL reaction volume (1.5 µL DNA amplicons, 5 ng, 5.5 µL nuclease-free water, 2.5 of Native Barcode EXP-NBD104 and EXP-NBD114, 10 µL NEBNext Ultra II Ligation

Master Mix, 0.5 µL of NEBNext Ligation Enhancer) for 20 min at 20 °C and 10 min at 65 °C. The sequencing adapter was ligated in a 50 µL reaction, with 50 ng of 24 barcoded amplicon pools, 10 µL of 5x NEBNext Quick Ligation Reaction Buffer, 5 µL AMII adapter mix, and 5 µL Quick T4 DNA Ligase, using an SQK-LSK109 kit. The ligation reaction was performed at room temperature for 20 min. The library was purified using AMPure XP beads and quantified using Qubit. Sequencing was performed on the MinION platform, and the final library (15 ng) was loaded onto the flow cell R.9 according to the manufacturer's instructions. ONT MinKNOW software was used to collect raw sequencing data.

Oxford-Nanopore raw data processing and sequencing data quality assessment. Basecalling and barcode demultiplexing were performed with guppy (v.4.4.1). Reads were processed according to the ARTIC Network protocols for COVID-19[13] using a nextflow pipeline (https://github.com/connor-lab/ncov2019-artic-nf). Briefly, for each sample, raw reads were mapped to the Wuhan reference sequence MN908947.3 using primer scheme V3 and Minimap (v.2.17). Post-alignment processing consisted of assigning reads to their derived amplicon and read group based on the primer pool, removing primer sequences, normalizing/reducing the number of read alignments to each amplicon and removing reads with imperfect pairings. Variant calling was performed with medaka (v.1.0.3) on the filtered and trimmed bam files. A final consensus FASTA file was generated by first marking positions not covered by at least 20 reads from either group as low coverage and building a pre-consensus FASTA with BCFtools consensus, which was subsequently aligned against the reference sequence using muscle (v.3.8.1551).

Illumina sequencing. The libraries were prepared using the Illumina COVID-seq protocol following the manufacturer's instructions. First-strand synthesis was carried out on RNA samples. The synthesized cDNA was amplified using ARTIC primers V3 for multiplex PCR, generating 98 amplicons across the SARS-CoV-2 genome. The PCR-amplified product was tagmented and adapted using IDT for Illumina Nextera UD Indices Set A, B, C, D (384 indices). Dual-indexed single-end sequencing with a 36 bp read length was carried out on the NextSeq 550 platform.

Illumina raw data processing and sequencing data quality assessment. The raw data were processed using DRAGEN Lineage v3.3 with standard parameters. Further samples with SARS-CoV-2 and at least 90 targets detected were processed for lineage designation.

Genomic data collection

Most B.1.1.519 sequences were generated at the Instituto Nacional de Medicina Genómica (INMEGEN) by the abovementioned protocol (N = 1,710). For completeness, we also downloaded from GISAID all sequences from Mexico City with their associated metadata (collection date < 2021-05-31, N = 906, not sequenced by INMEGEN). When high-quality sequences were required, we filtered by sequences with at most 1% N and less than 0.05% singletons (high coverage) (N = 1,874) [2]. Only INMEGEN samples had associated clinical information.

Effective reproduction number estimation for variants B.1.1.222 and B.1.1.519.

We grouped all sequenced samples based on the epidemiological week as the date of sample collection. We then calculated the percentage of samples for the variants of interest B.1.1.222 and B.1.1.519 and the percentage of samples that belonged to the ensemble of other variants. Using these percentages, we extrapolated the total number of confirmed cases using the federal database for residents of Mexico City treated in medical units within Mexico City. With this, we calculated an incidence time series for both variables of interest and the ensemble of other variants.

Using this percentage, we considered all confirmed cases in the federal database for residents of Mexico City treated in medical units within Mexico City. We assumed that these samples were divided in the same percentages as the ones observed in the sequenced samples for a given epidemiological week. Then, we calculated an incidence time series for both variables of interest and the ensemble of other variants.

We estimated the effective reproduction number (Rt) using the parametric method of Cori et al. 2013[14] and the parameters reported for the SARS-CoV-2 serial interval by Nishiura et al. 2020[15]. We restricted this analysis to the period beginning with epidemic week 2020-46, corresponding to the first detections of variant B.1.1.519.

Haplotype analysis for variant B.1.1.519.

Only high-quality sequences were considered. SARS-CoV-2 reference genome NC_045512.2 was downloaded from NCBI. SNVs and indels per SARS-CoV-2 sequence were obtained with nucmer[16]. Nucmer was executed with the following parameters: map each position of each query to its best hit in the reference, map each position of each reference to its best hit in the query and exclude alignments with ambiguous mapping. Variable positions in any SARS-CoV-2 sequence were obtained. Only variable positions observed in at least 5 genomes were further considered. Each SARS-CoV-2 sequence was translated into a compressed representation in which only the genotype of the list of variable positions was included. A unique combination of alleles, e.g., a unique compressed representation, was considered a haplotype. Haplotypes were used to infer a haplotype network using the haploNet function from the Population and Evolutionary Genetics Analysis System package (pegas)[17]. Briefly, genetic distances (Hamming distance) between all pairwise combinations of haplotypes were calculated using the dist.dna function of the Analyses of Phylogenetics and Evolution package (ape)[18]; from this distance matrix, the minimum spanning tree and the median-joining network were computed using pegas[19].

Phylogenetic Analysis

The sequences were aligned with MAFFT (version 7.475) using the FFT-NS-2 algorithm[20,21]. A maximum-likelihood phylogeny was calculated with FastTree (version

2.1.11) compiled with the double precision tag using a generalized time-reversible model (GTR)[22,23]. The resulting tree was visualized using the Interactive Tree Of Life (iTOL)[24].

Clinical data collection

To gather and correlate clinical data from our patients, we used the National Epidemiologic Surveillance System for Viral Respiratory Diseases (SISVER). This system gathers information including personal identification data, contact information, comorbidities, date of diagnosis, symptoms, progression, and outcome of all the COVID-19 cases reported in Mexico. After downloading the data collected in this system, we verified and complemented this information with our own variants of interest by applying a telephone survey.

Verbal consent and identification were the first steps when calling each subject included in the final analysis. Questions on our survey covered comorbidities (diabetes, hypertension, cardiovascular diseases, chronic renal failure, COPD, asthma, HIV, cancer, obesity, smoking, pregnancy status and immunosuppression), date of symptom onset, sampling date, COVID-19 symptoms (fever, NSAID-resistant fever, cough, dyspnea, chest pain, oxygen saturation, headache, myalgias, arthralgias, odynophagia, anosmia, ageusia, diarrhea, vomiting, rhinorrhea, polypnea, cyanosis, conjunctivitis and abdominal pain), disease progression (ambulatory or hospitalized). In cases of hospitalized patients, we asked the length of hospital stay, treatment (need for supplementary oxygen or intubation) and outcome (alive, dead or under treatment). For underage or deceased

patients, we tried to reach a close relative who was taking care of the individual and could answer all the questions with certainty.

Statistical analysis

Binary multivariate logistic regression models were fitted to predict the association of symptoms with variant B.1.1.519, as well as the association of hospitalization with the variant. An ordinal multivariate logistic regression model was fitted to predict the association of disease severity with the variant. The severity score was coded as 0 for asymptomatic or mild symptoms, 1 for severe symptoms and 2 for death. Individuals classified with severe disease were those who presented with at least one of the following: dyspnea, polypnea, cyanosis, requiring supplemental oxygen or intubation. All models were adjusted for covariates (age and sex) and comorbidities (immunosuppression, heart disease or hypertension, diabetes, obesity, asthma or smoking).

Results

Identification of variant B.1.1.519 in Mexico City

On November 3, 2020, the first patient carrying variant B.1.1.519 was detected in Mexico City, representing the second case recorded worldwide. The frequency of the B.1.1.519 variant began to increase in Mexico City, from 16% (17/106) to a peak in February 2021 of 90% (496/552). In March 2021, its frequency began to decrease, and in May 2021, it had dropped to 51%.

Variant B.1.1.519 represented 74.3% of the sequences generated in Mexico City (2,296/3,092) from November 2020 to May 2021 and was distributed evenly across all of Mexico City (Figure 1B). B.1.1.519 was detected in 31 countries, predominantly in Mexico at (55%, 6,041/10,922), followed by the USA (2.2%, 11,937/548,492), Canada (0.87%, 456/52,409) and Germany (0.14%, 192/130,634) by May 2021.

According to the phylogenetic analysis, this variant is grouped in an independent clade derived from the clade 20B NextClade classification (Figure 1C). The B.1.1.159 variant is characterized by 9 mutations (C203T, C222T, C3140T, C10954T, A11117G, C12789T, C21306T, C22995A, and C23604A), four ORF1a substitutions (P959S, T3255I, I3618V, and T4175I) and three spike substitutions (T478K, P681H, and T732A) (Figure 1C). The diversity along the SARS-CoV-2 genome for variant B.1.1.159 is presented in Figure 1D.

Rt: Effective Reproduction Number

We studied the effective reproduction number, defined as the average number of secondary cases per primary case at a given calendar time, to characterize the transmissibility of the B.1.1.519 variant. Matching the rapid increase in detection of variant B.1.1.519, we observed an increase in Rt for variant B.1.1.519 during the month of December 2020 up to a value of 2.9 in the second week of December, before stabilizing between 0.5 and 1 in the following months.

The second most frequent variant in Mexico City was B.1.1.222. All remaining variants had small frequencies and were considered one joint group. Variant B.1.1.222 reached a maximum Rt of 1.93 during the second week of December. Its estimated Rt values fluctuated strongly in the following months, which could be influenced by the small number of cases of this variant. In comparison, the estimated Rt for the ensemble of other variants has fluctuated steadily since the winter, with increases in the first week of January 2021, fourth week of February and second week of March, before stabilizing (Figure 2A and 2B) or disappearing (Figure 1A).

Genomic Findings

Phylogenetic analysis

We calculated a maximum-likelihood phylogeny including all SARS-CoV-2 genomes of interest (Methods) to study the geographic origin of the B.1.1.519 variant and its evolutionary relationship with the B.1.1.222 variant (Fig. 3). The phylogenetic tree shows 3 defined clusters, two of which correspond only to B.1.1.222 and B.1.1.519 variants, with clear separation, and a mixed cluster displaying a nonclearly defined separation among lineages. Thus, the detailed evolution of this SARS-CoV-2 lineage is still unclear. The mixed cluster is formed by the B.1.1.519 sequences most closely related to B.1.1.222 sequences. As part of the mixed cluster, we can observe a clade formed by a small subclade of B.1.1.222 and a small subclade of B.1.1.519 sequences. Most B.1.1.519 subclade sequences were sequenced in the United States (4 out of 5) and one in Mexico City. Therefore, the geographic origin of the B.1.1.519 variant remains unclear.

Haplotype analysis

A haplotype network can provide new insights into evolutionary processes when external and internal nodes of a phylogeny are simultaneously studied. The continuous sequencing of SARS-CoV-2 samples throughout the pandemic enables the study of ancestral and child sequences simultaneously. A haplotype network for variant B.1.1.519 was constructed in this study to enable the analysis of how the evolutionary and mutational processes have impacted its dispersion and prevalence (Figure 4). A haplotype was defined based on all variable sites for the B.1.1.519 variant.

The prevalence of any haplotype was defined as the period of time (measured as the number of days) in which a haplotype was observed. The prevalence was defined as zero if a haplotype was observed in only one sample. The month of appearance corresponds to the month in which the first sequence of any specific haplotype was observed.

In Figure 3, the most ancient B.1.1.519 sequence can be observed as a blue-bordered node. This node can be used as an anchor to suggest the temporal direction of the haplotype network. In this representation, the size of a node is proportional to the prevalence of the haplotype. The two most ancient nodes present two contrasting behaviors. One of them (red-bordered small node) was observed in very few samples (9 samples) over a long period of time (180 days), which could suggest a persistent, although not very transmissible, virus variant. The second most ancient sequence (red-bordered large node, haplotype III) was observed in the largest number of samples (106 samples) during the longest period of time (190 days), which suggests a persistent and transmissible virus variant.

The results show that haplotype III diverged in the three next-largest nodes, which implies that these three haplotypes are the next-most commonly observed haplotypes (70, 68 and 62 samples, respectively). All of these secondary haplotypes showed persistent behavior across time (166, 167, and 139 days of prevalence, respectively). Interestingly, all of these haplotypes diverged into a large number of less efficient virus variants.

Additionally, the month of appearance was consistent with the peak in the effective reproductive number described earlier, as most haplotypes were first observed in November and some in December 2020.

Clinical association

Finally, we studied the clinical impact of variant B.1.1.519. We analyzed the associations between variant B.1.1.519 and a number of clinical traits. Only sequences with complete clinical data were considered (N = 600). We found patients infected with variant B.1.1.519 to show a significant increase in the odds of developing symptoms affecting the respiratory tract relative to non-B.1.1.519 variants. In particular, logistic regression models adjusted for covariates (age, sex, viral Ct and number of commorbidities) showed that variant B.1.1.519 was associated with a 1.786-fold increase in dyspnea (P = 0.0028, 0.202–0.964 95% CI), a 1.489-fold increase in chest pain (P = 0.035, 0.029–0.769 95% CI) and a 3.655-fold increase in cyanosis (P=0.0456, 0.159–2.793 95% CI) (Table 1).

To investigate the relationship between variant B.1.1.519 and an increased risk of developing serious illness or death, we stratified patients into four age groups and compared their outcomes. Although we observed an overall trend of increasing disease seriousness with increasing age groups, infection with B.1.1.519 was still associated with a higher fraction of patients with serious illness and/or death than non-B.1.1.519 infection within each group (Figure 5).

We fitted logistic regression models to predict the severity of disease (see Methods section). After adjusting for covariates and various comorbidities, we still found that variant B.1.1.519 had a highly significant adjusted odds ratio (aOR) increase of 1.85-fold over non-B.1.1.519 variants (P = 0.000296, 1.33–2.6 95% CI) for developing a severe/critical outcome. Multivariate analyses adjusted for covariates also showed infections with variant B.1.1.519 to be associated with a 2.35-fold increase in the hospitalization rate (P = 0.005, 1.32–4.34 95% CI).

Discussion

SARS-CoV-2 variant B.1.1.519 has been tagged by an alert for further monitoring by the WHO, which implies that this variant could pose a future threat, but there is no evidence about phenotypic or clinical associations of concern. In this paper, we genomically describe the B.1.1.519 variant and its evolution, transmissibility and clinical impact. The first patient carrying variant B.1.1.519 was detected in Mexico City in November 2020, representing the second case recorded worldwide. Three defined clusters were defined in the phylogenetic tree, two of them corresponding to B.1.1.222 and B.1.1.519 variants with a clear separation, and a mixed cluster. Finally, patients infected with variant B.1.1.519 seemed to show a significant increase in developing symptoms affecting the respiratory tract relative to those with non-B.1.1.519 variants. In addition, logistic regression models showed that variant B.1.1.519 was associated with an increase in dyspnea, chest pain, and cyanosis.

Worldwide, new variants of SARS-CoV-2 classified by the WHO as AFM have emerged. These variants, such as P.2[25], B.1.621[26], and B.1.1.318[27], show spike mutations in receptor binding and S1/S2 cleavage sites and have spread widely within countries. By May 2021, B.1.1.519 had been detected in 31 countries and was predominantly found in Mexico (55%, 6,041/10,922), followed by the USA (2.2%, 11,937/548,492), Canada (0.87%, 456/52,409), and Germany (0.14%, 192/130,634). B.1.1.519 has three substitutions in spike: T478K, P681H, and T732A. The S:T478K substitution is structurally localized in the region of interaction with the human ACE2 receptor. SARS-CoV-2

attaches to this receptor to infect cells, thus spreading the infection more effectively[28,29]. A study of in silico molecular dynamics on the spike has shown that the distribution of charges in S:T478K is most drastically affected at the site of substitution and its immediate vicinity on the surface of the folded protein. This effect may critically change the specific interactions with drugs, antibodies or the ACE2 receptor, increasing infectivity[30]. Accordingly, the Delta variant (B.1.617.2) carries the S:T478K substitution. This substitution could impact B.1.1.519 transmissibility and may suggest why B.1.1.519 had a transmission advantage over other variants without S:T478K in Mexico City.

The S:P681H substitution is located immediately adjacent to amino acids 682–685, which correspond to a furin cleavage site at the S1/S2 binding site, where the more basic the string of amino acids is, the more effectively furin recognizes and cuts it[31]. An in vitro assay with SARS-CoV-2 S:P681H using fluorogenic peptides mimicking the S1/S2 sequence reported an increase in spike cleavage by furin-like proteases[32]. This furin cleavage site is key to SARS-CoV-2 replication and pathogenesis because more furin cuts mean more spike proteins primed to enter human cells[33]. The Alpha (B.1.1.7) and Gamma (P.1) variants (recognized by the WHO as VOCs) carry S:P681H[32,34]. The S:P681H substitution could also be involved in the increased transmissibility of B.1.1.519 in Mexico City.

Genomic surveillance has proven to be an important tool for the identification and characterization of viral spreading potential and the monitoring of novel variants of concern in SARS-CoV-2[11]. Based on genomic surveillance, we observed that variant

B.1.1.519 showed increased transmission during the first and third weeks of December 2020 at the beginning of the second COVID-19 wave in Mexico City. After this increased transmission period, we estimate that B.1.1.519 became the dominant variant in circulation for the remaining period analyzed in this manuscript until late May 2021, completely displacing the previously dominant B.1.1.222. Such behavior is similar to that exhibited by other SARS-CoV-2 variants found in other regions[35].

Although we estimate two peaks of increased transmission for B.1.1.519, it could very well be that a low number of viable samples available for sequencing during epidemiological week 50 artificially split the transmission peak of B.1.1.519, given the behavior exhibited by other SARS-CoV-2 variants[36]. Regardless of this possible artifact, we estimate that the bulk of observed cases in Mexico City during the winter wave of COVID-19 were associated with variant B.1.1.519, with the associated clinical implications described in this manuscript.

Phylogenetic methods can be applied to provide some insight into the evolution and spread of SARS-CoV-2[37]. However, conclusions drawn from phylogenetic and downstream analyses should be considered and interpreted with caution, as the sequences are too closely related. B.1.1.519 geographical origin could be inferred from the monophyletic group with its ancestor the B.1.1.222 lineage. However, the mixed inferred origin suggested fast dispersion due to human movement.

Evolutionary analysis of SARS-CoV-2 has been used to understand the spatiotemporal dynamics of the pandemic. Specifically, haplotype networks have been used to unravel the genetic diversity among monomorphic populations with small genetic distances between individuals, usually at the intraspecific level. Haplotype networks can be used to infer an evolutionary path for a given population. A median-joining network (MJN) is derived from a minimum spanning tree that traces a path between all studied sequences such that the total length is minimal. Additionally, an MJN will infer additional sequence types that minimize the inferred length; such inferred sequences can be considered biologically as unseen or extinct sequences. The distance between two sequence types will equal the number of nucleotide differences observed between them (Hamming distance)[38–40].

Recent studies have investigated the evolution and spatiotemporal distribution of SARS-CoV-2 via haplotype networks. Pereson, MJ, et al. studied the diversification of the spike protein in each SARS-CoV-2 clade, showing that two haplotypes predominated in specific clades (Hap-1 for clades G, GH and GR; and Hap-2 for clades L, O, S and V)[41]. In addition, sequence similarity and network structure were used to infer the import of SARS-CoV-2 from multiple countries in Bangladesh[42]. The edges in a haplotype network represent specific nucleotide substitutions, and the nodes represent specific sequence types or haplotypes. Garvin, MR, et al. inferred a haplotype network from 15,789 SARS-CoV-2 genomes to model their evolutionary success based on their duration, dispersal and frequency in the human population. They identified that the Pro323Leu mutation in the RNA-dependent RNA polymerase led to the rapid spread of the virus instead of the

previously reported Asp614Gly mutation in the spike glycoprotein. Importantly, they also inferred that the Pro323Leu mutation occurred on an Asp614Gly background[43].

The B.1.1.519 haplotype network shows a star-form, characteristic of an ongoing pandemic: ancestral central nodes surrounded by newly mutated peripheral nodes[44]. Continuous monitoring of SARS-CoV-2 genomes by this tool could highlight successful haplotypes with either high frequency, high prevalence or both. It could also highlight specific mutations responsible for increased transmission or prevalence. Indeed, some resources have been created to dynamically visualize haplotype networks of all worldwide SARS-CoV-2 genomes[45].

Finally, we observed that variant B.1.1.519 was significantly associated with severe disease, hospitalization, and death. Particularly with symptoms related to severe disease such as dyspnea, chest pain and cyanosis, which were more prevalent in B.1.1.519 compared with non-B.1.1.519 variant infections. Similarly, the Alpha VOC has been associated with an increased risk of hospitalization and greater disease severity or death[46,47]. Although there has been some contradictory evidence[48,49] concerning this point, more recent reports[50] have noted shortcomings of previous studies and reaffirmed the association of the variant with clinical severity. Some recent studies[51,52] have also shown that the Delta VOC is associated with an increased risk of hospitalization and severe illness/disease compared to infections with non-Delta variants that circulate at the same time. This increase, although smaller, is still significant compared to infections involving the Alpha, Beta and Gamma VOCs. Similarly, the Gamma VOC has also shown

an increased risk of hospitalization[53] and severity in young adults with pre-existing conditions[54]. There is little evidence relating the Beta VOC to more severe disease or death, with only one study[55] comparing differences in the first and second waves in South Africa as a "proxy" for the Beta variant, showing higher in-hospital mortality.

Conclusions

Sustained genomic surveillance plays a decisive role in identifying newly emerging SARS-CoV-2 variants and guiding the decisions of the public health care system in a country. Detailed evolutionary analysis is important to understand the origin and progression of newly evolving variants. Any significant clinical associations could be of interest in pandemic handling and containment.

References

1. Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. & Hsueh, P.-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents* **55**, 105924 (2020).

2. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. *Global Challenges* **1**, 33–46 (2017).

3. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **25**, (2020).

4. Centers for Disease Control and Prevention. Real-time RT-PCR Primers and Probes. (2020).

5. Mirtaleb, M. S. *et al.* Potential therapeutic agents to COVID-19: An update review on antiviral therapy, immunotherapy, and cell therapy. *Biomedicine & Pharmacotherapy* **138**, 111518 (2021).

6. Maurya, V. K., Kumar, S., Bhatt, M. L. B. & Saxena, S. K. Therapeutic Development and Drugs for the Treatment of COVID-19. in *Coronavirus Disease 2019 (COVID-19)* (ed. Saxena, S. K.) 109–126 (Springer Singapore, 2020). doi:10.1007/978-981-15-4814-7_10.

7. World Health Organization. Tracking SARS-CoV-2 variants. (2021).

8. Yadav, P. D. *et al. Neutralization of variant under investigation B.1.617 with sera of BBV152 vaccinees*. http://biorxiv.org/lookup/doi/10.1101/2021.04.23.441101 (2021) doi:10.1101/2021.04.23.441101.

9. Shen, X. *et al. SARS-CoV-2 variant B.1.1.7 is susceptible to neutralizing antibodies elicited by ancestral Spike vaccines*. http://biorxiv.org/lookup/doi/10.1101/2021.01.27.428516 (2021) doi:10.1101/2021.01.27.428516.

10. Wang, P. *et al. Increased Resistance of SARS-CoV-2 Variant P.1 to Antibody Neutralization*. http://biorxiv.org/lookup/doi/10.1101/2021.03.01.433466 (2021) doi:10.1101/2021.03.01.433466.

11. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).

12. Horby, Peter *et al.* NERVTAG paper on COVID-19 variant of concern B.1.1.7. (2021).

13. Nick Loman, Andrew Rambaut, & Will Rowe. nCoV-2019 novel coronavirus bioinformatics protocol. (2020).

14. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology* **178**, 1505–1512 (2013).

15. Nishiura, H., Linton, N. M. & Akhmetzhanov, A. R. Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases* **93**, 284–286 (2020).

16. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

17. Paradis, E. pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419–420 (2010).

18. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

19. Paradis, E. Analysis of haplotype networks: The randomized minimum spanning tree method. *Methods Ecol Evol* **9**, 1308–1317 (2018).

20. Katoh, K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059–3066 (2002).

21. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

22. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26**, 1641–1650 (2009).

23. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).

24. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293–W296 (2021).

25. Resende, P. C. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 P.2 Lineage Associated with Reinfection Case, Brazil, June–October 2020. *Emerg. Infect. Dis.* **27**, 1789–1794 (2021).

26. Laiton-Donato, K. *et al. Characterization of the emerging B.1.621 variant of interest of SARS-CoV-2*. http://medrxiv.org/lookup/doi/10.1101/2021.05.08.21256619 (2021) doi:10.1101/2021.05.08.21256619.

27. Tegally, H. *et al. Genomic epidemiology of SARS-CoV-2 in Mauritius reveals a new wave of infections dominated by the B.1.1.318, a variant under investigation*. http://medrxiv.org/lookup/doi/10.1101/2021.06.16.21259017 (2021) doi:10.1101/2021.06.16.21259017.

28. Di Giacomo, S., Mercatelli, D., Rakhimov, A. & Giorgi, F. M. Preliminary report on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Spike mutation T478K. *J Med Virol* **93**, 5638–5643 (2021).

29. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Molecular Cell* **78**, 779-784.e5 (2020).

30. Rezaei, S., Sefidbakht, Y. & Uskoković, V. Comparative molecular dynamics study of the receptor-binding domains in SARS-CoV-2 and SARS-CoV and the effects of mutations on the binding affinity. *Journal of Biomolecular Structure and Dynamics* 1–20 (2020) doi:10.1080/07391102.2020.1860829.

31. Maison, D. P., Ching, L. L., Shikuma, C. M. & Nerurkar, V. R. Genetic Characteristics and Phylogeny of 969-bp S Gene Sequence of SARS-CoV-2 from Hawai'i Reveals the Worldwide Emerging P681H Mutation. *Hawaii J Health Soc Welf* **80**, 52–61 (2021).

32. Lubinski, B., Tang, T., Daniel, S., Jaimes, J. A. & Whittaker, G. R. *Functional evaluation of proteolytic activation for the SARS-CoV-2 variant B.1.1.7: role of the P681H mutation*. http://biorxiv.org/lookup/doi/10.1101/2021.04.06.438731 (2021) doi:10.1101/2021.04.06.438731.

33. Johnson, B. A. *et al. Furin Cleavage Site Is Key to SARS-CoV-2 Pathogenesis*. http://biorxiv.org/lookup/doi/10.1101/2020.08.26.268854 (2020) doi:10.1101/2020.08.26.268854.

34. Felipe Naveca *et al.* Emergence and spread of SARS-CoV-2 P.1 (Gamma) lineage variants carrying Spike mutations **Δ**141-144, N679K or P681H during persistent viral circulation in Amazonas, Brazil. (2020).

35. Galloway, S. E. *et al.* Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021. *MMWR Morb. Mortal. Wkly. Rep.* **70**, 95–99 (2021).

36. Volz, E. *et al. Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from*

linking epidemiological and genetic data. http://medrxiv.org/lookup/doi/10.1101/2020.12.30.20249034 (2021) doi:10.1101/2020.12.30.20249034.

37. Morel, B. *et al.* Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution* **38**, 1777–1791 (2021).

38. Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**, 37–48 (1999).

39. Dumaidi, K. *et al.* Genetic diversity, haplotype analysis, and risk factor assessment of hepatitis A virus isolates from the West Bank, Palestine during the period between 2014 and 2016. *PLoS ONE* **15**, e0240339 (2020).

40. Leigh, J. W. & Bryant, D. popart: full-feature software for haplotype network construction. *Methods Ecol Evol* **6**, 1110–1116 (2015).

41. Pereson, M. J. *et al.* Evolutionary analysis of SARS-CoV-2 spike protein for its different clades. *J Med Virol* **93**, 3000–3006 (2021).

42. Shishir, T. A., Naser, I. B. & Faruque, S. M. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. *PLoS ONE* **16**, e0245584 (2021).

43. Garvin, M. R. *et al.* Potentially adaptive SARS-CoV-2 mutations discovered with novel spatiotemporal and explainable AI models. *Genome Biol* **21**, 304 (2020).

44. Pandit, B., Bhattacharjee, S. & Bhattacharjee, B. Association of clade-G SARS-CoV-2 viruses and age with increased mortality rates across 57 countries and India. *Infection, Genetics and Evolution* **90**, 104734 (2021).

45. Song, S. *et al.* The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes

in 2019nCoVR. *Genomics, Proteomics & Bioinformatics* S1672022920301315 (2020) doi:10.1016/j.gpb.2020.09.001.

46. Bager, P. *et al.* Risk of hospitalisation associated with infection with SARS-CoV-2 lineage B.1.1.7 in Denmark: an observational cohort study. *The Lancet Infectious Diseases* S1473309921002905 (2021) doi:10.1016/S1473-3099(21)00290-5.

47. Davies, N. G. *et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7*. http://medrxiv.org/lookup/doi/10.1101/2021.02.01.21250959 (2021) doi:10.1101/2021.02.01.21250959.

48. Frampton, D. *et al.* Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *The Lancet Infectious Diseases* S1473309921001705 (2021) doi:10.1016/S1473-3099(21)00170-5.

49. Graham, M. S. *et al.* Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. *The Lancet Public Health* **6**, e335–e345 (2021).

50. Giles, B., Meredith, P., Robson, S., Smith, G. & Chauhan, A. The SARS-CoV-2 B.1.1.7 variant and increased clinical severity—the jury is out. *The Lancet Infectious Diseases* S147330992100356X (2021) doi:10.1016/S1473-3099(21)00356-X.

51. Sheikh, A., McMenamin, J., Taylor, B. & Robertson, C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *The Lancet* **397**, 2461–2462 (2021).

52. Ong, S. W. X. *et al.* Clinical and Virological Features of SARS-CoV-2 Variants of Concern: A Retrospective Cohort Study Comparing B.1.1.7 (Alpha), B.1.315 (Beta), and

B.1.617.2 (Delta). *SSRN Journal* (2021) doi:10.2139/ssrn.3861566.

53. Funk, T. *et al.* Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Eurosurveillance* **26**, (2021).

54. Freitas, A. R. R. *et al. The increase in the risk of severity and fatality rate of covid-19 in southern Brazil after the emergence of the Variant of Concern (VOC) SARS-CoV-2 P.1 was greater among young adults without pre-existing risk conditions*. http://medrxiv.org/lookup/doi/10.1101/2021.04.13.21255281 (2021) doi:10.1101/2021.04.13.21255281.

55. Jassat, W. *et al. Increased mortality among individuals hospitalised with COVID-19 during the second wave in South Africa*. http://medrxiv.org/lookup/doi/10.1101/2021.03.09.21253184 (2021) doi:10.1101/2021.03.09.21253184.

Figures

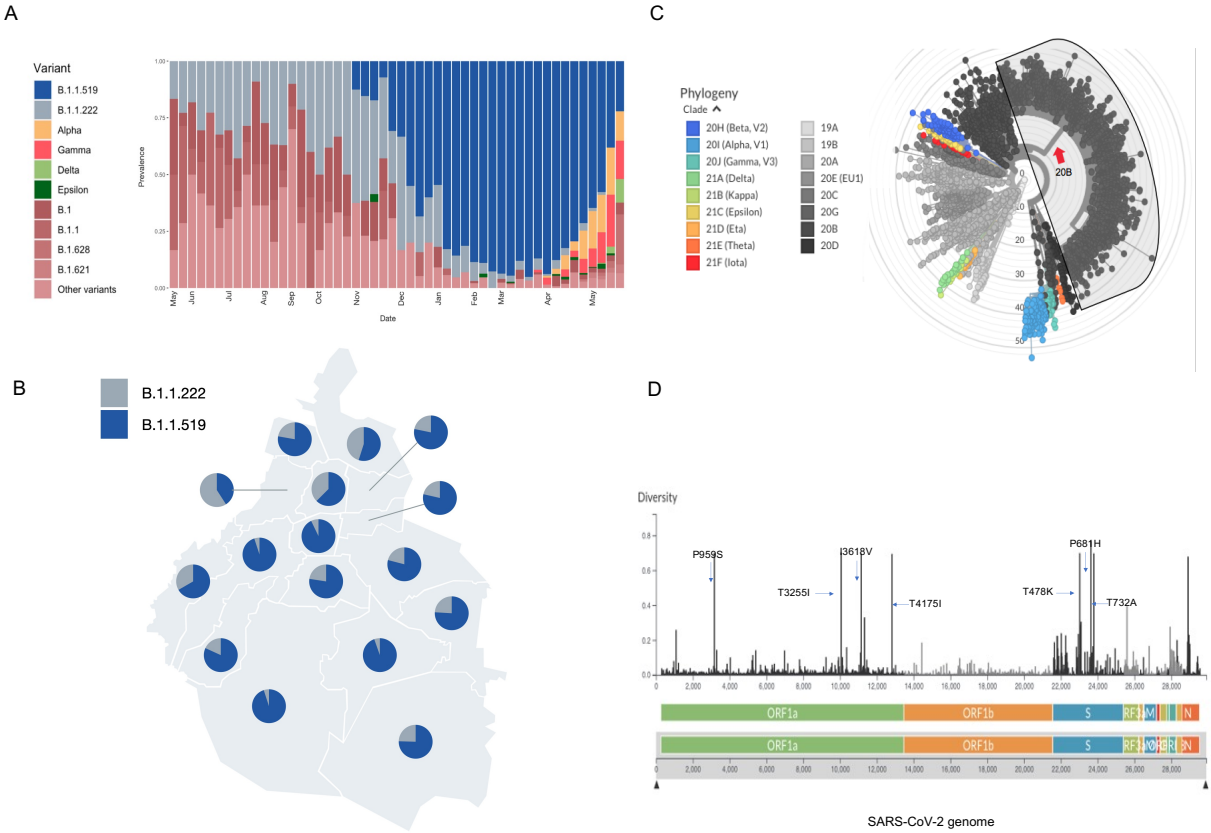

Figure 1. A. Frequencies of the B.1.1.519 variant in Mexico City from November 2020 to May 2021. B. The geographic distribution of B.1.1.519 and B.1.1.222 variants in Mexico City, with dominance of the first variant. C. Phylogenetic tree of SARS-CoV-2 with NextClade clades. The branch indicated with a red arrow represents 1,874 sequences of the B.1.1.519 variant of Mexico City with coverage of >99.5%. D. Genome map of SARS-CoV-2 variant B.1.1.519 with the most representative amino acid substitutions in 1,874 sequences of the B.1.1.519 variant of Mexico City with coverage of >99.5%.

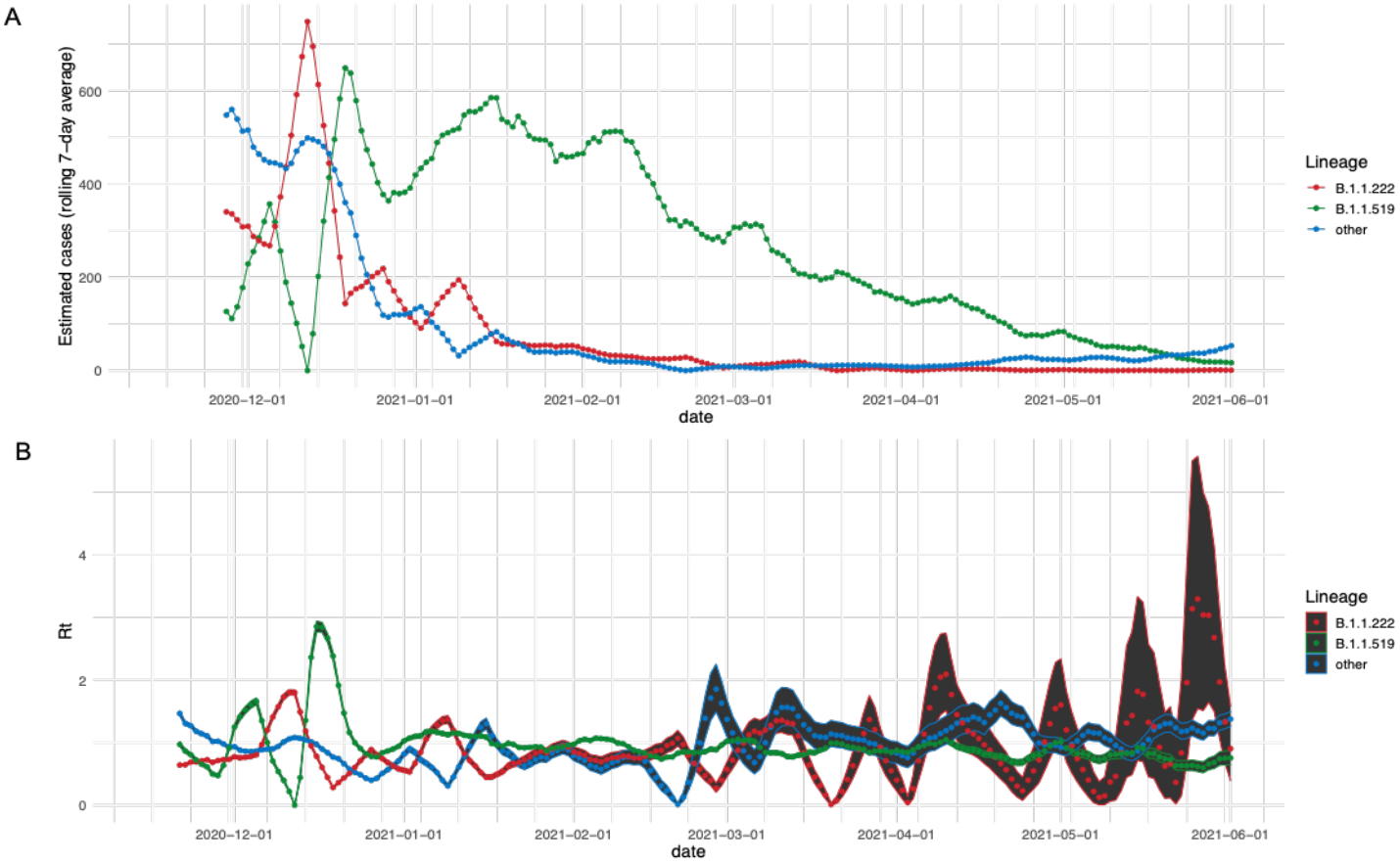

Figure 2. A. Estimated number of cases for each variant based on the frequency observed in sequenced samples at INMG and the daily tally of confirmed cases in SINAVE, 7-day rolling average. B. Time series of estimated Rt. Points represent the mean estimated Rt value per variant. Ribbon boundaries indicate the 5 (lower) and 95 (upper) quantile boundaries of the estimation.

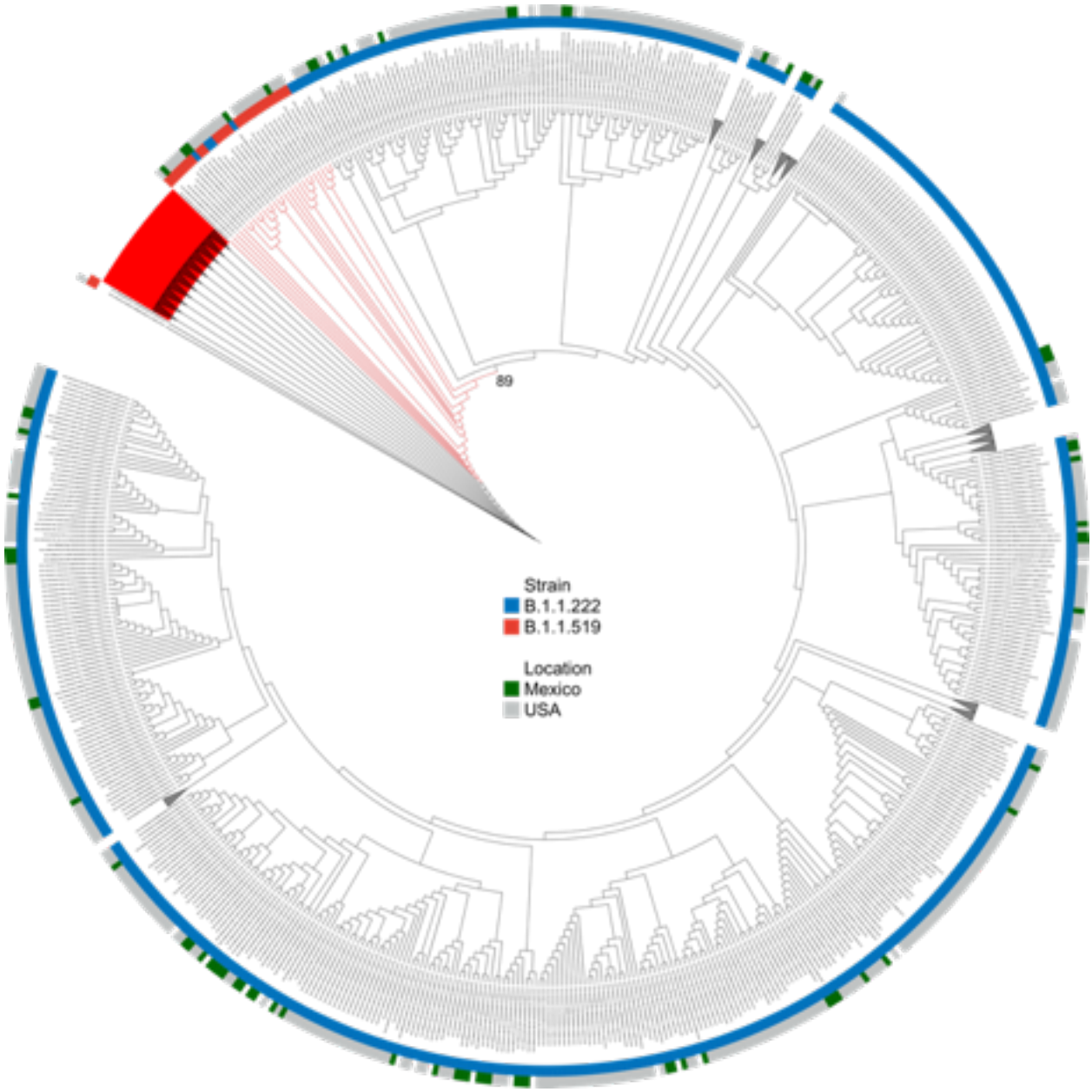

Figure 3. Phylogenetic relationships of SARS-CoV-2 B.1.1.222 and B.1.1.519 lineages. A Maximum-Likelihood phylodynamic inference was done of 84 SARS-CoV-2 sequences from Mexico in a global background of 19312 sequences available in the GISAID EpiCoV database as of 1 May 2020. Leaves are colored according to their Pango lineage: B.1.1.519 (red) and B.1.1.222 (blue) and according to their geographical origin: Mexico

(green) and USA (gray). The bootstrap value of the mixed cluster (described in the main text) is shown.

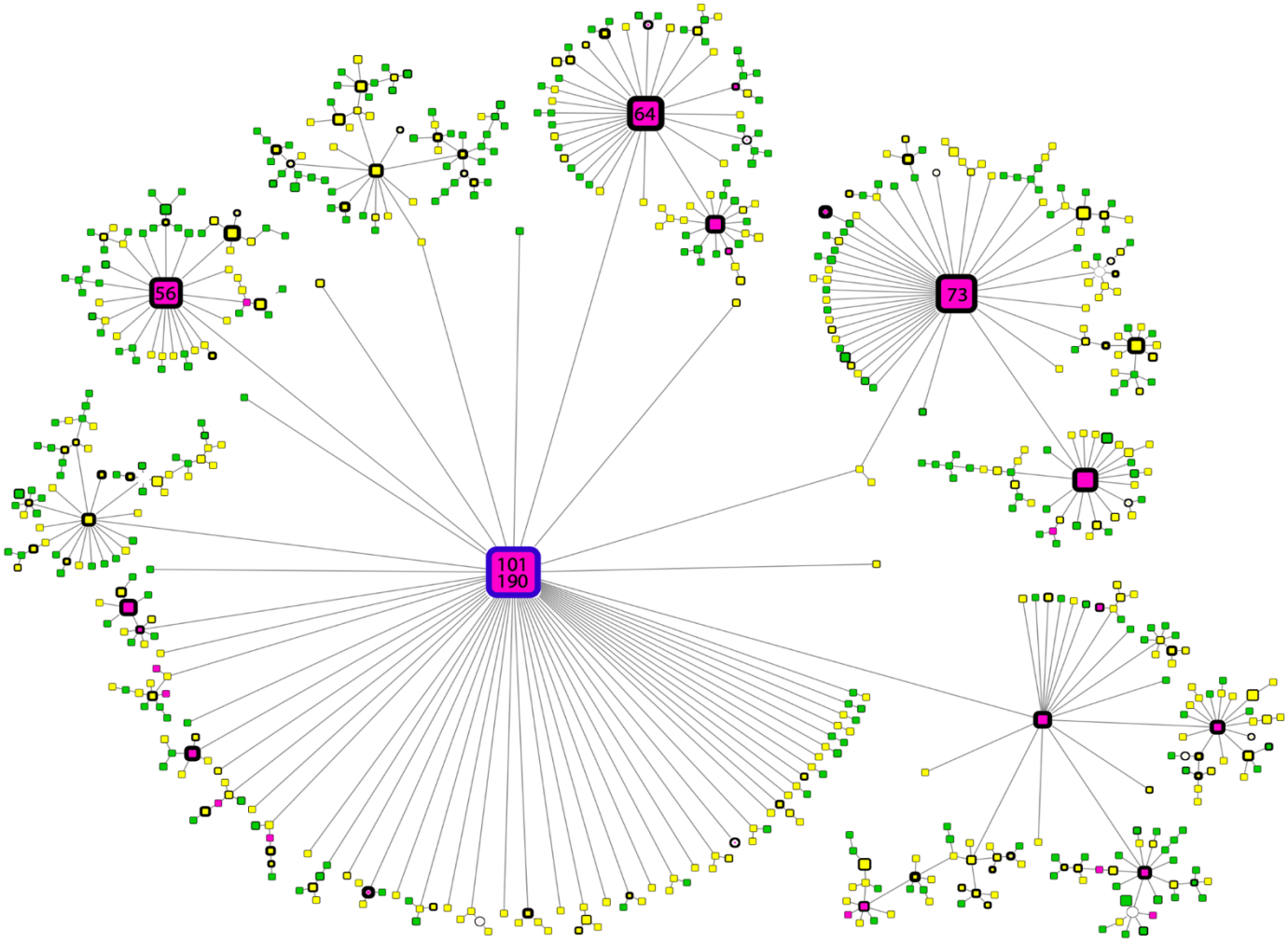

Figure 4. Haplotype network of B.1.1.519 sequences. Node colors represent the month of appearance (pink: November or December, yellow: January, February or March; green: April or May). Node size is proportional to the number of samples for that specific haplotype, and border width is proportional to the prevalence of the haplotype. Numbers correspond to the number of samples for that specific haplotype. The blue-bordered node indicates the haplotype with the most ancient appearance date for lineage B.1.1.519.

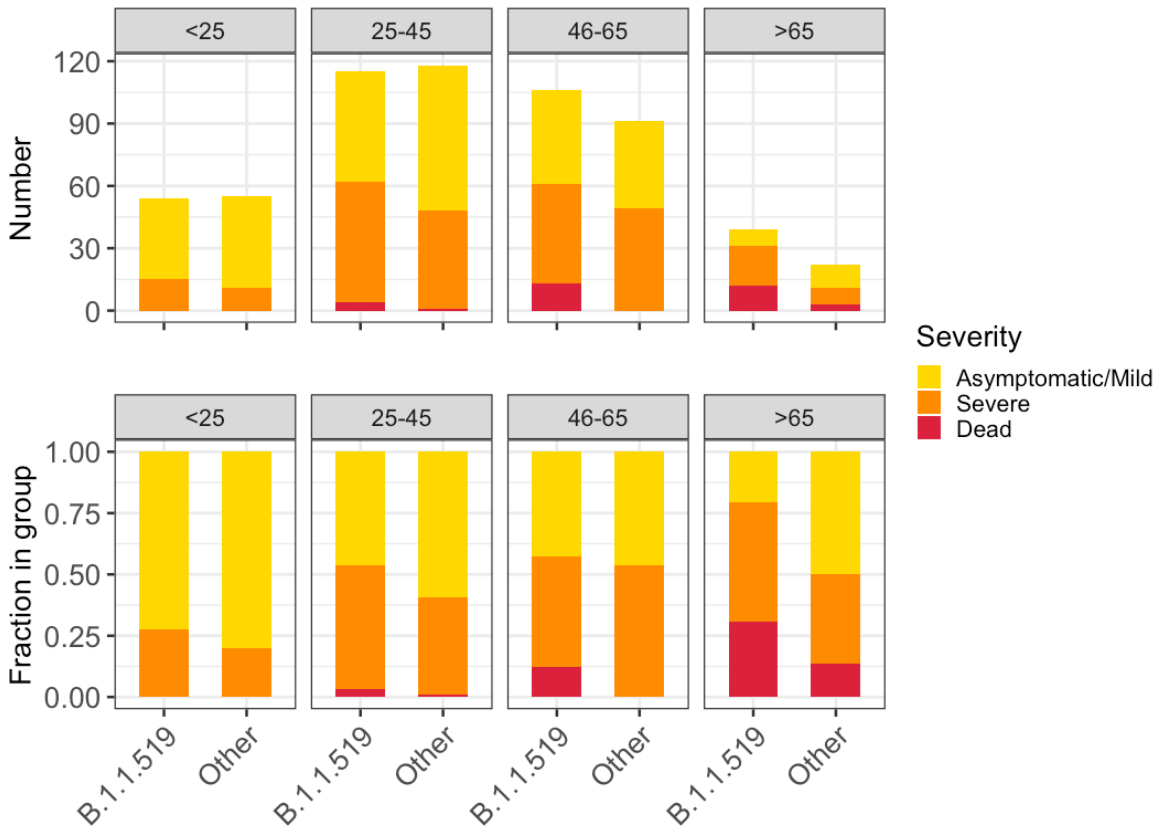

Figure 5. Severity of illness across patient age groups and by presence of B.1.1.519 or non-B.1.1.519 SARS-CoV-2 infections. The figure shows absolute counts (upper) and proportions of patients (lower).

Tables

Table 1. Associations between symptoms and variant B.1.1.519 using multivariate LR adjusted for covariates

Symptom	B.1.1.519 n(%)	Other n(%)	aOR	95% CI	p-value
Dyspnea	154 (49.0%)	103 (36.0%)	1.786	(0.202 - 0.964)	**0.0028**
Chest Pain	162 (51.6%)	117 (40.9%)	1.489	(0.029 - 0.769)	**0.0350**
Cyanosis	20 (6.4%)	9 (3.1%)	3.665	(0.159 - 2.793)	**0.0456**
Diarrhea	113 (36.0%)	91 (31.8%)	1.464	(-0.007 - 0.777)	0.0565
Polipnea	40 (12.7%)	46 (16.1%)	1.721	(-0.067 - 1.200)	0.0909
Myalgia	211 (67.2%)	181 (63.3%)	1.317	(-0.101 - 0.652)	0.1507
Rhinorrhea	98 (31.2%)	103 (36.0%)	0.757	(-0.661 - 0.106)	0.1549
Anosmia	173 (55.1%)	183 (64.0%)	0.769	(-0.636 - 0.107)	0.1647
Conjuntivitis	69 (22.0%)	89 (31.1%)	0.744	(-0.714 - 0.125)	0.1660
Odynophalgia	144 (45.9%)	143 (50.0%)	0.807	(-0.581 - 0.151)	0.2508
Arthralgia	196 (62.4%)	172 (60.1%)	1.235	(-0.159 - 0.581)	0.2628
Cough	204 (65.0%)	166 (58.0%)	1.233	(-0.167 - 0.584)	0.2740
Vomit	31 (9.9%)	26 (9.1%)	1.345	(-0.336 - 0.970)	0.3702
Persistant.Fever	47 (15.0%)	52 (18.2%)	0.800	(-0.719 - 0.279)	0.3780
Cephalea	213 (67.8%)	201 (70.3%)	0.861	(-0.553 - 0.249)	0.4651
Fever	185 (58.9%)	173 (60.5%)	0.925	(-0.456 - 0.298)	0.6868
Abdominal.Pain	31 (9.9%)	37 (12.9%)	1.011	(-0.591 - 0.636)	0.9714

Table 2. Association of the SARS-CoV-2 B.1.1.519 variant with disease severity and hospitalizations. The severity outcomes were coded as 0=Asymptomatic/Mild, 1=Severe, or 2=Dead; an ordinary multivariate LR model was fitted adjusted for covariates. A binary multivariate LR model was fitted for hospitalization.

	Summary	**Ordinal Multivariable LR Model (Severity)**			**Binary Multivariable LR Model (Hospitalization)**		
Characteristic	**N = 600[1]**	**aOR[2]**	**95% CI[2]**	**p-value**	**aOR[2]**	**95% CI[2]**	**p-value**
Severity							
Asymptomatic/Mild	312 (52%)						
Severe	255 (42%)						
Dead	33 (5.5%)						
Hospitalized	69 (12%)						
Age	42 (29, 54)	1.04	1.03, 1.05	**<0.001**	1.06	1.04, 1.08	**<0.001**
Sex							
Female	302 (50%)	—	—		—	—	
Male	298 (50%)	1.21	0.87, 1.70	0.3	1.69	0.95, 3.04	0.075
Ct	19.13 (17.90, 20.40)	0.99	0.93, 1.06	0.8	1.05	0.95, 1.17	0.3
ImmunoSuppressed	18 (3.0%)	2.86	1.12, 7.42	**0.029**	2.43	0.59, 8.23	0.2
HD_Hypertension	107 (18%)	1.18	0.73, 1.90	0.5	1.55	0.82, 2.91	0.2
Diabetes	73 (12%)	0.91	0.53, 1.56	0.7	1.09	0.53, 2.14	0.8
Obesity	236 (39%)	1.42	1.01, 1.99	**0.044**	1.61	0.91, 2.87	0.10
Asthma	20 (3.3%)	1.53	0.62, 3.73	0.3	1.11	0.21, 4.43	0.9
Smoker	164 (27%)	1.21	0.83, 1.75	0.3	0.84	0.43, 1.57	0.6
Variant							
Other	286 (48%)	—	—		—	—	
B.1.1.519	314 (52%)	1.85	1.33, 2.60	**<0.001**	2.35	1.32, 4.34	**0.005**

[1]n (%); Median (IQR)

[2]OR = Odds Ratio, CI = Confidence Interval