PROPOSED TITLE: "Comparing the evolutionary dynamics of predominant SARS-CoV-2 virus lineages co-circulating in Mexico"

AUTHOR LIST AND AFFILIATIONS

Hugo G. Castelán-Sánchez [1,2,†], Luis Delaye [1,3,†], Rhys P. D. Inward [4,†], Simon Dellicour [5,6,†], Bernardo Gutierrez [1,4,†], Natalia Martinez de la Vina [8], Oliver G Pybus [4,7], Guillermo de Anda Jáuregui [1,2,8], Plinio Guzmán[9], Marisol Flores Garrido[10], Óscar Fontanelli[1,3], Maribel Hernández Rosales[1,3], Amilcar Meneses[11], Gabriela Olmedo-Alvarez[1,3], Alfredo Herrera-Estrella[1,12,], Celia Boukadida[1,13], Alejandro Sanchez-Flores[1,14], José Esteban Muñoz-Medina [1,15], Andreu Comas-García [1,16], Bruno Gómez-Gil [1,17], Selene Zárate [1,18], Blanca Taboada [1,19], Susana López [1,19], Carlos F. Arias [1,19], Moritz U.G. Kraemer [1,4], Antonio Lazcano [1,20], Marina Escalera-Zamudio [1,4, *]

[1] Consorcio Mexicano de Vigilancia Genómica (CoViGen-Mex)
[2] Programa de Investigadoras e Investigadores por México, Consejo Nacional de Ciencia y Tecnología. Av. Insurgentes Sur 1582, Crédito Constructor, Benito Juárez, CP 03940, Ciudad de México, México
[3] Departamento de Ingeniería Genética, CINVESTAV-Unidad Irapuato, Km. 9.6 Libramiento Norte Carretera Irapuato-León CP 36824 Irapuato, Guanajuato, México
[4] Department of Zoology, University of Oxford, Parks Rd OX1 3PS, Oxford, United Kingdom
[5] Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Av. Franklin Roosevelt 50, 1050 Bruxelles, Belgium
[6] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium
[7] Department of Pathobiology, Royal Veterinary College, 4 Royal College St, NW1 0TU, London, United Kingdom
[8] Instituto Nacional de Medicina Genómica, Periférico Sur 4809, Arenal Tepepan, Tlalpan, CP 14610, Ciudad de México, México
[9] Astronomer LTD, Ciudad de México, México
[10] Escuela Nacional de Estudios Superiores, Unidad Morelia, Universidad Nacional Autónoma de México, Antigua Carretera a Pátzcuaro, No. 8701 Col. Ex-Hacienda de San José de La Huerta, CP 58190 Morelia, Michoacán, México.
[11] Departamento de Ciencias de la Computación, CINVESTAV-IPN, Av. IPN 2508, Gustavo A. Madero, San Pedro Zacatenco, CP 07360, Ciudad de México, México
[12] Laboratorio Nacional de Genómica para la Biodiversidad-Unidad de Genómica Avanzada, CINVESTAV-Unidad Irapuato, Km 9.6 Libramiento Norte Carretera Irapuato-León, CP 36824, Irapuato, Guanajuato, México
[13] Centro de Investigación en Enfermedades Infecciosas, Instituto Nacional de Enfermedades Respiratorias, Calz. de Tlalpan 4502, Belisario Domínguez Secc 16, CP 14080, Ciudad de México, México
[14] Unidad Universitaria de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, Chamilpa, CP 62210, Cuernavaca, Morelos, México.
[15] Coordinación de Calidad de Insumos y Laboratorios Especializados, Instituto Mexicano del Seguro Social, José Urbano Fonseca Núm. 6 Magdalena de las Salinas, CP 07760, Ciudad de México, México
[16] Facultad de Medicina y Centro de Investigación en Ciencias de la Salud y Biomedicina, Universidad Autónoma de San Luis Potosí, Av. Sierra Leona 550, lomas 2a Sección, Lomas de San Luis, CP 78120, San Luis Potosí, México
[17] Centro de Investigación en Alimentación y Desarrollo-CIAD, Unidad Regional Mazatlán en Acuicultura y Manejo Ambiental, Av. Sábalo Cerritos S/N, Cerritos CP 82112, Mazatlán, Sinaloa, México.
[18] Posgrado en Ciencias Genómicas, Universidad Autónoma de la Ciudad de México, Calle Dr. García Diego 168, Doctores, Cuauhtémoc, CP 06720 Ciudad de México, México
[19] Departamento de Genética del Desarrollo y Fisiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, CP 62210 Cuernavaca, Morelos, México
[20] Facultad de Ciencias, Universidad Nacional Autónoma de México, Av. Universidad 3000, Circuito Exterior s/n, Coyoacán, CP 04510, Ciudad Universitaria, Ciudad de México, México

ORCID 0000-0002-4763-0267 Hugo G Castelán-Sánchez hugo.castelan@conacyt.mx
ORCID 0000-0003-4193-2720 Luis Delaye luis.delaye@cinvestav.mx
ORCID 0000-0003-0016-661X Rhys Inward rhys.inward@zoo.ox.ac.uk
ORCID 0000-0001-9558-1052 Simon Dellicour simon.dellicour@ulb.be
ORCID 0000-0002-9220-2739 Bernardo Gutierrez bernardo.gutierrez@zoo.ox.ac.uk
Natalia Martinez de la Vina nataliamv@gmail.com
ORCID 0000-0002-8797-2667 Oliver G Pybus oliver.pybus@zoo.ox.ac.uk
ORCID 0000-0002-7749-0365 Guillermo de Anda Jáuregui gdeanda@inmegen.edu.mx
Plinio Guzmán pgzmnk@gmail.com
ORCID 0000-0003-1620-7885 Marisol Flores-Garrido mflores@enesmorelia.unam.mx
ORCID 0000-0002-2381-7035 Óscar Fontanelli ofontanelli@ciencias.unam.mx
ORCID 0000-0003-1878-1223 Maribel Hernández Rosales maribel.hr@cinvestav.mx
ORCID 0000-0003-1976-6199 Amilcar Meneses ameneses@cs.cinvestav.mx
ORCID 0000-0003-2122-7506 Gabriela Olmedo-Alvarez golmedo@cinvestav.mx
ORCID 0000-0002-4589-6870 Alfredo Herrera-Estrella alfredo.herrera@cinvestav.mx
ORCID 0000-0002-1744-0083 Celia Boukadida celia.boukadida@cieni.org.mx
ORCID 0000-0003-0476-3139 Alejandro Sánchez-Flores alejandro.sanchez@ibt.unam.mx
ORCID 0000-0002-1289-4457José Esteban Muñoz-Medina eban10@hotmail.com
ORCID 0000-0002-2368-1529 Andreu Comas-García andreu.comas@uaslp.mx
ORCID 0000-0002-3695-3597 Bruno Gómez-Gil bruno@ciad.mx
ORCID 0000-0003-1034-204X Selene Zárate selene.zarate@uacm.edu.mx
ORCID 0000-0003-1896-5962 Blanca Taboada btaboada@ibt.unam.mx
ORCID 0000-0001-6336-9209 Susana López susana.lopez@ibt.unam.mx
ORCID 0000-0003-3130-4501 Carlos F Arias arias@ibt.unam.mx
ORCID 0000-0001-8838-7147 Moritz Kraemer moritz.kraemer@zoo.ox.ac.uk
ORCID 0000-0001-7365-8557 Antonio Lazcano alar@ciencias.unam.mx
ORCID Marina Escalera-Zamudio marina.escalerazamudio@zoo.ox.ac.uk

[†] HGCS, LD, RPDI, SD and BG contributed equally to this work
* Correspondence to: Marina Escalera-Zamudio marina.escalerazamudio@zoo.ox.ac.uk

ABSTRACT (150/150 WORDS)

Up to November 2021, over 200 different SARS-CoV-2 lineages circulated in Mexico. To investigate lineage replacement dynamics, we applied a phylodynamic approach to explore the evolutionary trajectories of five dominant lineages that circulated during the first year of the local epidemic. For most lineages, peaks in sampling frequencies coincided with different epidemiological waves of infection in the country. Lineages B.1.1.222 and B.1.1.519 showed comparable dynamics, represented by clades likely originating in Mexico and persisting for over a year. Lineages B.1.1.7, P.1 and B.1.617.2 also displayed similar dynamics, characterized by multiple introduction events leading to a few successful extended local transmission chains that persisted for several months. We further explored viral movements across the country, applied within the largest clades identified (belonging to lineage B.1.617.2). Many clades were located within the south region of the country, suggesting that this area played a key role in the spread of SARS-CoV-2 in Mexico.

MAIN TEXT (7179/5000 WORDS)

INTRODUCTION

Genome sequencing efforts for the surveillance of the Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) has granted public access to a massive number of virus genomes generated worldwide (https://www.gisaid.org/). Exploring SARS-CoV-2 genome data using molecular evolution tools has enabled to characterize an increasing virus diversity [1], to track emerging viral subpopulations, and to explore virus evolution in real-time both at local and global scales (for examples see [2–6]). Throughout the development of the COVID-19 pandemic, viral variants have emerged and circulated across different regions of the world [2,7], displaying specific mutations that define their phylogenetic clustering patterns [8,9]. SARS-CoV-2 virus lineages have been classified using a dynamic nomenclature system ('Pango system', Phylogenetic Assignment of Named Global Outbreak Lineages), developed to consistently assign newly generated virus genomes to existing lineages, and to designate novel virus lineages according to their phylogenetic identity and epidemiological relevance [10,11].

The emergence and spread of SARS-CoV-2 virus lineages has been routinely monitored since early 2021 (https://www.gisaid.org/), prompted by public health authorities to inform on the ongoing responses to the pandemic [12]. Virus lineages that may pose an increased risk to global health are classified as Variants of Interest (VOIs), Variants under Monitoring (VUMs), and Variants of Concern (VOC), and may display some of the following biological properties [9,13] : increased transmissibility [14], decreasing the effectiveness of available diagnostics or therapeutic agents (such as monoclonal antibodies)[15], and evasion of immune responses (including vaccine-derived immunity) [12,16,17]. Up to date, five distinct SARS-CoV-2 lineages (including their descending sub-lineages) have been designated as VOC: B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), B.1.617.2 (Delta), and more recently, the B.1.1.529 lineage (Omicron) [12,18,19].

The spread of variants in different regions of the world has been linked to human movement, with VOC often replacing previously circulating virus lineages to become dominant (for example, see [20]). Virus lineages that dominate across distinct geographic regions are likely to denote an evolutionary advantage, driven in part by a genetic background prompting some increase in fitness (such as enhanced transmissibility and/or immune escape) [20–24]. However, whilst some lineages may become dominant in some geographic regions, others fail to do so [20]. Exploring lineage replacement dynamics remains challenging, as it represents a complex scenario impacted by numerous factors, such as: i)

31 differential and stochastic growth rates linked to lineage-specific spread patterns varying across
32 geographic regions, ii) a genetic background and a shifting immune structure of the host population
33 (linked to viral pre-exposure levels, as well as vaccination rates),[20,24] and iii) changing behavioural
34 patterns, such as fluctuating human mobility and the implementation of local non-pharmaceutical
35 interventions across time [20,25,26].

36 Mexico has been severely impacted by the COVID-19 pandemic, evidenced by the high number
37 of cumulative deaths reported relative to other countries in Latin America [27]. Since the first introductions
38 of the virus into the country during early 2020 [28], the development of the local epidemic has been
39 characterized by an epidemiological curve fluctuating between three waves of infection observed in July
40 2020, January 2021, and August 2021 (up to November 2021) [27–29]. Prior to the peak of the first wave,
41 non-pharmaceutical interventions including social distancing and suspension of non-essential activities
42 were implemented at a national scale (from March 23rd 2020 to May 30th 2020), whilst a reopening plan
43 for the country was announced on May 13th 2020. Further, the national vaccination campaign did not
44 began until December 24th 2020 [27].

45 During February 2021, a sequencing consortium for the surveillance of SARS-CoV-2 in Mexico
46 was launched (Mexican Consortium for Genomic Surveillance, abbreviated CoViGen-Mex [30]). In an
47 attempt to further homogenize the spatial and temporal representation of the viral genomes sequenced,
48 the sampling scheme was upgraded in May 2021, for which approximately 1,200 representative
49 samples selected across all states following official case reports were sequenced monthly [30]. This
50 resulted in approximately one third of the virus genome sequences from Mexico publicly available.
51 Independent to CoViGen-Mex, other national institutions significantly contributed to the genomic
52 surveillance of the virus (https://www.gisaid.org/). Sequencing efforts at a local scale have enabled to
53 characterize viral diversity, and to identify the viral lineages circulating within national territory [30–32]. Over
54 200 different virus lineages have been detected in Mexico, including all VOC [19,33]. Many lineages also
55 co-circulated across time and space, an observation of particular importance in the context of the
56 recently described recombinant SARS-CoV-2 linages that emerged in North America during 2021 [34].

57 Using a phylodynamic approach, we used publicly available virus genome data collected in
58 Mexico to investigate the introduction, spread and replacement patterns of predominant virus lineages
59 (B.1.1.222, B.1.1.519, B.1.1.7, P.1 and B.1.617.2) that circulate(d) in the country up to November 2021.
60 To generate datasets, we implemented a migration and a phylogenetically-informed subsampling

approach, aiming to increase robustness for tailored phylogeographic analyses. We further explored the correlation between human mobility and the inferred viral movements for the largest clades (C5d and C6d) belonging to the B.1.617.2 lineage, which also represented the most extended transmission chains. Analysis revealed similar dynamics for the B.1.1.222 and B.1.1.519, with both of these likely originating in Mexico and showing a sustained transmission over a year. For the P.1, B.1.1.7 and the B.1.617.2, multiple introduction events were identified, with few large transmission chains representing successful viral spread across the country. For C5d, spread occurred from the south, with subsequent movement into the central and northern region of the country.

RESULTS

Prior to the introduction of 'Omicron' (B.1.1.529) into Mexico, only the B.1.1.222, B.1.1.519, B.1.1.7 (Alpha), P.1 (Gamma) and B.1.617.2 (Delta) virus lineages showed a progressive replacement pattern to become dominant over other co-circulating virus lineages, with peaks in sampling frequency often coinciding with the epidemiological waves of infection (except for B.1.1.7 and the P.1; **Figure 1a and b**) [19,31,32]. During this time, Mexico reported a daily mean test rate ranging between 0.13-0.18 test/1,000 inhabitants, approximately 19 times lower than the global daily median test rate recorded at the time [29]. Despite testing limitations, the number of viral genomes generated at a national scale correlates with the cumulative number of cases recorded throughout the development of the epidemic (up to March 2022), with the majority of genomes coming from Mexico City (**Supplementary Figure 1**). In average, 100 genomes per 10,000 cases have been generated, corresponding to approximately 1% of the official COVID-19 cases reported. The distribution of genomes sequenced across states apparently coincides with the reported numbers of active and cumulative confirmed cases at the time [35,36,37]. Nonetheless, this is likely impacted by an uneven genome sequencing across states (particularly evident during the first year of the epidemic), and by differences in the sequencing intensity across time.

- B.1.1.222

The B.1.1.222 lineage circulated between April 2020 and September 2021, mostly sampled from the USA (77% of all sampled sequences) and Mexico (22% of all sequences), with a limited detection in other regions of the world (as reported in https://cov-lineages.org/) [38]. However, the apparent proportion of genome sequences sampled from different countries reflects sequencing biases across locations.

91 The first B.1.1.222-assigned genome was collected from Mexico on the 04/04/2020 (Mexico/CMX-
92 INER-0026/2020-04-04) [38], whilst the last B.1.1.222-assigned genome was collected from the USA on
93 the 14/09/2021 (USA/CA-CDPH-1002006730/2021-09-14). The last B.1.1.222-assigned genome from
94 Mexico dates to the 26/07/2021 (Mexico/CHH_INER_IMSS_1674/2021-07-26), two months before the
95 latest collection date for the B.1.1.222 lineage at an international scale.

96 The B.1.1.222 lineage was continuously sampled from the country between April 2020 and May
97 2021, followed by a steady decline in detection after July 2021 (**Figure 1b**). Most of the genome
98 sampling corresponds to the central region of the country, represented by Mexico City (CMX; **Figure**
99 **2a**). Data corresponding to the first year of the epidemic showed that by February 2021, the B.1.1.222
100 had reached a cumulative frequency of approximately 10% relative to other virus lineages co-circulating
101 within national territory [32]. Our results show that the sampling frequency of the B.1.1.222 in Mexico
102 progressively increased after June 2020, initially coinciding with the peak of the first epidemiological
103 wave in the country, further reached a highest of up to 35% in October 2020 **(Figure 1b)**. Thus, the
104 B.1.1.222 lineage is likely associated with the observed increase in the number of cases at that time.
105 An initial frequency underestimation for the B.1.1.222 lineage is attributable to a limited and centralized
106 virus genome sequencing during the first year of the epidemic in the country.

107 Phylogenetic analysis revealed a main clade deriving from a single MRCA (most recent
108 common ancestor; here denoting lineage emergence), with subsequent introductions into other
109 countries (namely the USA; **Figure 2a**). The inferred date for this earliest MRCA corresponds to March
110 2020, a month before the earliest collection date of the first B.1.1.222-assigned genome (Mexico/CMX-
111 INER-0026/2020-04-04; **Figure 2a**). During this time, the B.1.1.222 was likely to be seeded into the
112 USA from Mexico multiple times; thus, all subsequent inferred introduction events are interpreted as
113 're-introductions'. In this context, our analysis revealed a minimum of 237 distinct introduction events
114 from Mexico into the USA (95% HPD interval = [225-250]), and a minimum of 106 distinct events from
115 the USA into Mexico (95% HPD interval = [95-122]; **Figure 2a**). The dates inferred for re-introduction
116 events ranged from October 2020 to July 2021. The B.1.1.222 lineage displayed a total persistence of
117 up to 16 months.

118

119 - B.1.1.519

120 Descending from the B.1.1.222, the B.1.1.519 lineage was first detected in North America, being mostly
121 sampled in the USA (58% of all sequences) and Mexico (32% of all sequences, as reported in
122 https://cov-lineages.org/) [32,39–41]. Again, the proportion of genomes from different countries reflect
123 sequencing biases across geographic regions. The first B.1.1.519-assigned genome was collected in
124 the USA on the 31/07/2020 (USA/TX-HMH-MCoV-45579/2020-07-31) [42], with the lineage being last
125 detected on December 2021 [43]. During our initial phylogenetic assessment, most of the B.1.1.519-
126 assigned genomes collected after November 2021 were identified as outliers, and were excluded from
127 further analyses. Of note, these genomes were sampled from countries where the lineage was not
128 reported to be locally circulating (namely Turkey and Africa).

129 In Mexico, the B.1.1.519 lineage was first detected on the 28/08/2020 (Mexico/YUC-NYGC-
130 39037-20/2020-08-28) and was last detected on the 01/11/2021 (Mexico/CAM_IBT_IMSS_4057/2021-
131 11-01) [32]. By late February 2021, the B.1.1.519 represented the majority of the viral genomes from the
132 country [32,41]. The spread of the B.1.1.519 lineage raised awareness in public health authorities, [32,40,41],
133 leading to its designation as a VUM in June 2021 [12]. Cumulative virus genome data shows that by
134 January 2021, the B.1.1.519 progressively displaced other co-circulating virus lineages, representing
135 up to 75% of all virus genomes generated from the country, and dominating during the peak of the
136 second wave of infection (**Figure 1a and b**). Most of the genome sampling again corresponds to the
137 central region of the country, represented by the state of Puebla (PUE; **Figure 2b**).

138 Phylogenetic analysis revealed a pattern similar to the one observed for the B.1.1.222, with an
139 earliest MRCA leading to a single clade from which subsequent dispersal events into other countries
140 occurred (namely USA; **Figure 2b**). This earliest MRCA dates to July 2020, again representing lineage
141 emergence within the country. Our analysis suggests that the B.1.1.519 circulated cryptically in Mexico
142 for four months prior to its initial detection, being seeded back and forth between the USA and Mexico
143 during this time (**Figure 2b**). Such observations are congruent with the collection date of the earliest
144 B.1.1.519 genome sampled globally (July 2020, USA/TX-HMH-MCoV-45579/2020-07-31), positioned
145 basally to the B.1.1.519 clade. As only good-quality sequences were included within our dataset, two
146 earlier genome sequences from Mexico excluded from our datasets collected between August and
147 October 2020 (Mexico/YUC-NYGC-39037-20/2020-08-28 and Mexico/CMX-INER-IBT-0485/2020-10-
148 13) suggest that the cryptic circulation period for the lineage could have been shorter. We estimated a
149 minimum number of 121 introduction events from the USA into Mexico (95% HPD interval = [109-132]),

150 compared to a minimum number of 391 introduction events from Mexico into the USA (95% HPD interval
151 = [380-402]), all occurring between July 2020 and November 2021 (the latter represented by sequence
152 Mexico/CAM_IBT_IMSS_4057/2021-11-01; **Figure 2b**). The B.1.1.519 lineage displayed a total
153 persistence of approximately 16 months (earliest collection date: 28/08/2020, latest collection date:
154 01/11/2021).
155
156 - B.1.1.7

157 The B.1.1.7 lineage was first detected in the UK during September 2020, spreading to over 175
158 countries in over a year [44]. The earliest B.1.1.7-assigned genome from Mexico was first collected on
159 the 31/12/2020 (Mexico/TAM-InDRE-94/2020-12-31), and the latest on the 13/10/2021
160 (Mexico/QUE_InDRE_FB47996_S8900/2021-10-13). Cumulative data showed a continuous lineage
161 detection for over ten months, with a sampling peak observed around May 2021, followed by a steady
162 decline with almost no detection by late August 2021 (**Figure 1b**). Relative to the epidemic in the
163 country, the circulation of the B1.1.7 in Mexico coincided with a lower number of recorded cases
164 observed during May-June 2021 [27,29,45]. Overall, the B.1.1.7 lineage reached a lower genome sampling
165 frequency (of up to 25%), compared to the other virus lineages studied here (**Figure 1b**). Most of the
166 genome sampling corresponds to the northern region of the country, represented by the state of
167 Chihuahua (CHH; **Figure 3a**).
168 Phylogenetic analysis revealed a minimum of 224 introduction events into Mexico (95% HPD
169 interval = [219-231]; **Figure 3a**), with an earliest MRCA dating to October 2020, denoting a cryptic
170 circulation period of approximately two months. Subsequent introduction events occurred between
171 March and June 2021, with a total lineage persistence of ten months. Seven clades (C1a to C7a) were
172 identified to represent large local transmission chains (composed of >20 sequences from Mexico with
173 ancestral nodes supported by a PP value >.80; **Figure 3a, Supplementary Data 2**). The earliest
174 genome sequences were collected from the north-east region of the country (mostly represented by the
175 state of Chihuahua, CHH) [32,46] and fall within the C2a identified here, with an MRCA dating to late
176 January 2021 (**Supplementary Data 2**). Of note, the collection date of the earliest B.1.1.7 genome from
177 Mexico coincides with at least four independent introduction events that occurred synchronously during
178 December 2020 (**Figure 3a**).

179 The C3a and C7a were identified as the largest clades, with both MRCAs dating to February 2021,
180 denoting again independent and synchronous introduction events (**Figure 3a**). C3a was sampled from
181 22/32 states in the country, mostly from Mexico City (CMX), followed by southern states of Yucatán
182 (YUC) and Quintana Roo (ROO; **Supplementary Figure 4, Supplementary Data 2**). The MRCA for
183 C3a dates to late February 2021, showing a persistence of three months (earliest collection date:
184 02/03/2021, latest collection date: 28/06/2021). Within C7a, ~70% of the genomes were sampled from
185 20/32 states of the country (mostly represented by the southern state of Tabasco; TAB), followed by
186 the norther-eastern state of Tamaulipas (TAM; **Supplementary Data 2**). As for the C3a, the MRCA
187 node for C7a dates to late February 2021, showing a persistence of four months (earliest collection
188 date: 10/03/2021, latest collection date: 05/07/2021).
189
190 - P.1

191 The P.1 lineage was first detected in Brazil during October 2020 [47], after which it diverged into >20 sub-
192 lineages circulating in different countries[19]. Relevant to the epidemic in Mexico, the P.1.17 virus sub-
193 lineage was sampled mostly from the USA (66% of all sequences) and Mexico (29% of all sequences;
194 https://cov-lineages.org/) [19], again keeping in mind the sequencing bias across geographic regions. We
195 detected at least 13 P.1 sub-lineages in the country, with the majority of genomes assigned to the
196 P.1.17 sub-lineage (66%), and to a lesser extent to the parental P.1 lineage (25%) [33]. As our dataset
197 comprises genomes assigned to the P.1 and descending sub-lineages, it is henceforth referred here as
198 a P.1+. The earliest P.1+ genome from Mexico was collected in January 2021 (Mexico/JAL-
199 InDRE_245/2021-01-28) and the latest in November 2021 (Mexico/ROO_IBT_IMSS_4502/2021-11-
200 19). Cumulative data shows that genome sampling frequency for the P.1+ peaked around June 2021,
201 with almost no detection after September 2021. Most of the genome sampling corresponds to the
202 southern region of the country, represented by the states Yucatan and Quintana Roo (YUC and ROO;
203 **Figure 3b**). As for the B.1.1.7, the P.1+ showed an overall lower genome sampling frequency (of 25%)
204 relative to the other lineages studied here, with spread coinciding with a decreasing number of cases
205 following the second wave of infection in the country (**Figure 1b)** [31–33].
206 Phylogenetic analysis revealed a minimum number of 126 introduction events into Mexico (95%
207 HPD interval = [120-132]; **Figure 3b**). We identified two well-supported clades named here C1_P1
208 (corresponding to P.1) and C1_P1_17 (corresponding to P.1.17; **Figure 3b, Supplementary Data 2**).

209 Genomes within C1_P1 were mostly collected from the central region of the country (>40% of the
210 sequences from Mexico City and the State of Mexico; CMX and MEX), and directly descend from
211 genomes from South America (**Supplementary Figure 5**). The MRCA of clade C1_P1 dates to March
212 2021, showing a persistence of seven months (earliest collection date: 07/03/2021, latest collection
213 date: 01/10/2021). Subsequent smaller introduction events were identified for the P.1, with most MRCAs
214 positioned closely together (referred here as clade C1_P1 MRCAs 1 and 2; **Figure 3b**).
215 Genomes within clade C1_P1_17 we mostly sampled from Mexico (75%), and to a lesser extent
216 from the USA (20%). 'Mexico' genomes are positions basally, and were collected predominantly from
217 the southern region of the country (90%, represented by the states of Quintana Roo and Yucatán, ROO
218 and YUC; **Supplementary Figure 5**). The MRCA of C1_P1_17 dates to October 2020, corresponding
219 to the earliest introduction event inferred in the tree. However, the long branch between the MRCA and
220 the earliest sequence within the clade suggests a significant under-sampling, indicating a lag in
221 detection after lineage emergence/introduction (**Figure 3b**). Considering tip dates only, the C1_P1_17
222 showed a persistence of five months (earliest collection date: 01/04/2021, latest collection date:
223 17/09/2021). As for P.1, subsequent P.1.17 introduction events were identified as smaller clusters, with
224 their MRCAs positioned closely together (referred here as C1_P1_17 MRCAs; **Figure 3b**). Our results
225 indicate that the P.1 parental lineage was introduced independently to the P.1.17 and from distinct
226 geographic locations. Contrasting to the P.1 parental lineage, the P.1.17 showed a limited number of
227 introductions but spread successfully, showing a persistence of up to five months and establishment
228 within the southern region of the country.
229
230 - B.1.617.2

231 Initially detected in India during October 2020, the B.1.617.2 lineage was associated with an increase
232 in the number of COVID cases recorded globally after March 2021, spreading globally to become
233 dominant [48,49]. The parental B.1.617.2 lineage further diverged into >230 descending sub-lineages
234 (designated as the AY.X sub-lineages) that circulate(d) in different regions of the world [19,48,50]. Again,
235 as our dataset comprises both B.1.617.2 and AY. X-assigned genomes, it is henceforth referred here
236 as a B.1.617.2+. The first 'B.1.617.2-like' genome from Mexico was collected on September 2020
237 (Mexico/AGU-InDRE_FB18599_S4467/2020-09-22), followed by a sporadic genome sampling
238 throughout January 2021 (with <10 sequences) [51]. However, lineage expansion only occurred after April

239 2021 (representative earliest sequence: Mexico/JAL-InDRE_FB13752_S1722/2021-04-21). By the end
240 of August 2021, the B.1.617.2+ reached a relative frequency of >95%, coinciding with the peak of the
241 third wave of infection recorded in the country (**Figure 1b**) [52]. Over 80 AY.X sub-lineages were detected
242 in Mexico, with most genomes being assigned to the AY.20 (22%), AY.26 (13%), and AY.100 (5%),
243 followed by the AY.113 and other sub-lineages (AY.62 and AY.3) [53]. All of these were initially
244 characterized as being most prevalent within North America [54–58]. Compared to the other virus lineages
245 studied here, genome sampling of the B.1.617.2+ lineage in Mexico displayed a more homogeneous
246 distribution across states, associated with an enhanced genome sampling and sequencing after May
247 2021 (**Figure 4**).

248 Phylogenetic analysis revealed a minimum number of 142 B.1.617.2+ introduction events into
249 Mexico (95% HPD interval = [125-148]), and a total lineage persistence of seven months (up to
250 November 30th 2021). We identified six major clades (C1d-C6d), with C1d, C3d, C5d and C6d being
251 the largest (**Figure 4, Supplementary Data 2**). At least four synchronous introduction events were
252 identified as the earliest MRCAs dating to April (including the introduction events leading to C3d, C4d
253 and C6d). Within the smaller clades, C2d comprises genomes from Mexico assigned to the AY.62,
254 sampled mainly from the state of Yucatán. Clade C4d comprises genomes from Mexico assigned to the
255 AY.3, sampled from the north, central and south of the country (**Supplementary Data 2**). Within the
256 larger clades, C1d is composed of genomes from Mexico assigned to AY.26, sampled mostly from the
257 north (>60%; BCS, SIN, JAL), followed by central (CMX) and south-eastern states (VER, ROO and
258 YUC). The MRCA of C1d dates to May 2021, showing a persistence of six months (earliest collection
259 date: 17/05/2021, latest collection date: 12/11/2021). Comparably, C3d is composed of genomes from
260 Mexico assigned to AY.26, sampled mostly from the north (37%; SIN, BCS and SON), and to a lesser
261 extent from central, east and southern states. The MRCA of C3d dates to April 2021, displaying a
262 persistence of seven months (earliest collection date: 15/05/2021, latest collection date: 30/11/2021;
263 **Figure 4, Supplementary Data 2**). Thus, C1d and C3d represent two independent introductions events
264 of the AY.26 sub-lineage into the country.

265 Clade C5d is one of the largest clades identified (n=~1500), represented by genomes assigned
266 to the AY.100 (44%), to the parental B.1.617.2 (40%), and to the AY.113 sub-lineage (12%). Within
267 C5d, the AY.100 and B.1.617.2 genomes are separated by a central sub-cluster of AY.113-assigned
268 genomes (**Figure 4, Supplementary Data 2**). Approximately, 70% of the genomes within C5d were

sampled from Mexico (mostly assigned to the AY.100 and AY.113), and 30% from the USA (mostly assigned to the B.1.617.2). Genomes from Mexico are positioned basally and distally, and were sampled from all 32 states, but predominantly from the north, centre and south of the country (>50%; represented by CHH, DUR, NLE, CMX, MEX, CAM, YUC, TAB, CHP and ROO; **Figure 4, Supplementary Data 2**). Thus, C5d represents the most genetically and geographically diverse clade identified. The MRCA of C5d dates to May, showing a persistence of up to six months (earliest collection date: 28/05/2021, latest collection date: 29/11/2021). Finally, C6d is the second largest clade (n=~1150), composed of genomes mostly assigned AY.20 sub-lineage, mainly sampled from central region of the country (>60%; represented by CMX, MEX, MOR, MIC and HID), suggesting a restricted geographic distribution for this transmission chain [59]. The MRCA of C6d dates to April, displaying a persistence of seven months (earliest collection date: 06/05/2021, latest collection date: 30/11/2021; **Figure 4**).

- Spread of the B.1.617.2 lineage

Being the largest clades, we further explored viral spread for the C5d and C6d under a phylogeographic approach (see Methods section 4). For C5d, viral spread within the country is likely to have occurred from the south (represented by the states of Chiapas and Campeche; CHP and CAM) into the rest of the country (**Figure 4**). Well-supported transitions (scored under a with a BFR>100 and a PP>.90, **Supplementary Table 1**) were mostly inferred from the southern state of Campeche (CAM) into central and northern states, and subsequently from the northern state of Chihuahua (CHH) into the central and northern region of the country (with some bidirectionality observed). Additional well-supported transitions were observed from Baja California into Chihuahua (BCN/BCS into CHH), and from Chihuahua into the USA (arbitrarily represented by the geographic coordinates of the state of California; **Figure 4**). Contrastingly, for C6d, viral spread within the country is likely to have occurred from the central region (represented by Mexico City, CMX) into central, northern and southern states (again with some bidirectionality observed). Additional well-supported transitions were inferred from the southern state of Chiapas into central and northern states (**Figure 4, Supplementary Table 1**). Animated visualizations of the spread of C5d and C6d across Mexico is available as **Supplementary Data 3 and 4**.

- Linking virus spread and human mobility data

Analysis of human mobility data collected between 01/01/2020 and 31/12/2021 at a national scale (see Methods section 5) revealed two peaks in human mobility observed across time (**Figure 1c**). The first occurred between February and April 2021, coinciding with the introduction and spread of the B.1.1.7 and P.1+ lineages, and with the withdrawal of the B.1.1.519. The second peak was observed between August and November 2021, coinciding with the spread of the B.1.617.2+. Increased movements (represented by the cumulative number of trips into a given state) were observed for Mexico City, followed by Jalisco, the State of Mexico, Nuevo León and Puebla, and to a lesser extent for Coahuila, Guanajuato and Veracruz (**Figure 1c**). Mean intra-state connectivity within national territory revealed intensified movements from Mexico City into the State of Mexico, Morelos, Hidalgo, Puebla, Veracruz and Jalisco, and from Jalisco into Michoacán and Guadalajara (**Supplementary Figure 6**). However, for both the C5d and C6d clades, no correlation between transitions and the mean intrastate connectivity was observed (C5d: Adjusted R-squared: 0.006577, F-statistic: 7.15 on 1 and 928 DF, p-value: 0.007628. C6d: Adjusted R-squared: 0.3216, F-statistic: 470.8 on 1 and 990 DF, p-value: < 2.2e-16), nor with the pairwise distance between states (C5d: Adjusted R-squared: 0.01086, F-statistic: 4.051 on 1 and 277 DF, p-value: 0.04511. C6d: Adjusted R-squared: 0.02296, F-statistic: 2.715 on 1 and 72 DF, p-value: 0.1038). Many of the clades identified were localized within the southern region of the country (clades C3a and C7a belonging to the B.1.1.7, C1_P1_17 belonging to the P.1.17, and C2d and C5d belonging to the B.1.617.2). Ranking connectivity between the southern region of the country (represented by the states of Yucatán, Quintana Roo, Chiapas and Campeche) and the remaining 28 states did reveal a consistently high number of bidirectional movements between CAM, CMX and VER (**Supplementary Figure 6**).

DISCUSSION

Our analyses revealed contrasting evolutionary dynamics between lineages, with some identifiable patterns. The B.1.1.222 and B.1.1.519 are likely to have originated in Mexico, with viral spread characterized by single and extended transmission chains sustained for over a year. During this time, these lineages were seeded back and forth into Mexico and the USA, remaining endemic, but never dominating across the USA. The B.1.1.222 lineage reached a maximum national frequency of 3.5% in the USA, whilst in Mexico it reached a maximum of 35% (up to October 2020). Similarly, the B.1.1.519

329 lineage reached a maximum national frequency of 5% in the USA (up to April 2021), whilst in Mexico it
330 reached a maximum of 76% (up to March 2021) [38,42]. Thus, a greater number of genomes sampled from
331 the USA compared to the numbers from Mexico (https://cov-lineages.org/), reflect sequencing
332 disparities across countries, an important consideration for genomic epidemiology analyses [1].

333 The introduction and spread B.1.1.7, P.1 and B.1.617.2 lineages in Mexico was characterized by
334 multiple introduction events resulting in a few successful extended local transmission chains, with many
335 clades constrained to specific regions, and displaying a persistence shorter periods of time. Further
336 similarities were observed between the P.1 and B.1.1.7, which overlapped both spatially and temporally,
337 coinciding with a decrease in the number of cases following the second wave of infection in the country.
338 Compared to the B.1.1.222 and the B.1.1.519 (and later to the B.1.617.2), and contrasting to what was
339 observed for other countries (like UK, USA and Brazil) [20,47,60], the P.1 and B.1.1.7 lineages did not
340 dominate in Mexico [32,46]. Nonetheless, the epidemiological dynamics of these VOC in the country are
341 comparable to those observed for other regions of the world [20,47,60]. For instance, in the USA, the earliest
342 introductions of the B.1.1.7 were synchronous to those observed in Mexico (occurring between October
343 and November 2020), with a similar spread pattern characterized by a few extended transmissions
344 chains, often constrained to specific states [60]. Thus, viral spread does not seem to be linked to a higher
345 number of introduction events, but rather to the extent and distribution of successful transmission
346 chains, with the size of clades likely reflecting virus transmissibility [61].

347 Increasing growth rates (*Rt;* the instantaneous reproduction number measuring how an infection
348 multiplies [62]) for viral variants within different geographic locations are partially explained by a genetic
349 background increasing viral fitness [20,47,60]. For instance, in the UK, the B.1.617.2 lineage sweep
350 occurred at a growth rate of around 60% greater than for the B.1.1.7, resulting from a combination
351 specific mutations increasing transmissibility and immune evasion from the host population [20]. Within
352 Mexico, the B.1.1.519 likely displayed an enhanced fitness over the previously dominating B.1.1.222.
353 In agreement with this observation, an *Rt* of 2.9 was estimated for the B.1.1.519, compared to an *Rt* of
354 1.93 for the B.1.1.222 (calculated during epidemic week 2020-46, coinciding with the early expansion
355 of the B.1.1.519 in the country [40]). As these two lineages were directly phylogenetically related, they
356 were almost genetically identical [41]. Relative to the parental B.1.1.222, the B.1.1.519 displayed only two
357 amino acid changes within the Spike protein: T478K and P681H [41]. Mutation T478K is located within
358 the RBD domain, with a possible impact in antibody-mediated neutralization [63,64]. On the other hand,

359 mutation P681H is located upstream the furin cleavage site within an epitope signal hotspot [65] and may
360 potentially enhance viral cell entry [66], reduce antibody-mediated recognition [65], and confer Type I
361 interferon resistance [67].

362 We speculate that at least one of these mutations may have contributed to the dominance of the
363 B.1.1.519 lineage over the B.1.1.222 by increasing the *Rt*, as has been documented for other SARS-
364 CoV-2 subpopulations [21,22,68]. Noteworthy, the mutations observed in the B.1.1.519 were not exclusive
365 to the lineage, as P681H emerged later and independently in the B.1.1.7 (and to a lesser extent in P.1,
366 corresponding to 5% of all sampled genomes) [69,70]. Moreover, mutation T478K subsequently appeared
367 in the B.1.617.2 [63,66,67]. In this scenario, we propose that a somewhat 'shared' genetic background
368 between the B.1.1.519 and the B.1.1.7 (represented by mutation P681H) may have limited the spread
369 of this VOC in Mexico. Although the impact of specific mutations on viral fitness would need to be
370 assessed experimentally, current data derived from the natural virus population evidences that amino
371 acid changes observed at site 681 of the Spike protein have been predominantly observed in VOCs,
372 and are likely to yield an evolutionary advantage [63,66,67,71–73].

373 As for other regions of the world, lineage replacement dynamics during the first year of the epidemic
374 in Mexico were probably driven by increased virus transmissibility. However, lineage establishment was
375 likely to be shaped by the changing immune landscape of the local host population [74]. Relatively
376 constant exposure levels to widespread and genetically similar virus subpopulations for extended
377 periods of time (represented by the B.1.1.222 and the B.1.1.519 lineages), may have yielded
378 consistently increasing immunity levels in a somewhat still naïve population (assuming a lower infection
379 attack rate at that time). As more genetically divergent virus lineages were introduced and began to
380 spread (the P.1 and B.1.1.7 lineages, and later the B.1.617.2), this resulted in a shift in the local immune
381 landscape, impacted by a virus genetic background prompting (at least a partial) evasion of the existing
382 immune responses. In Mexico, a vaccination rate of >50% (reflecting the proportion of the population
383 immunized with at least two doses) was only reached after December 2021 [29,75], suggesting that
384 immunity levels in the population depended mostly on virus pre-exposure levels. By December 2020, a
385 nationwide representative seroprevalence was estimated to be of ~33.5% [76,77].

386 Differences in the proportion of genomes sequenced across states did not always mirror the
387 geographic distribution and extension of the transmission chains identified, but are likely reflecting an
388 increasing sampling and sequencing intensity through time and space. Many clades identified were

389 distributed within the south region of the country, suggesting this area has played a key role in the
390 spread of SARS-CoV-2, comparable to observations made for other viral epidemics in Mexico
391 (Gutierrez *et al.*, *in preparation*). Although geographic distances and human mobility trends at state
392 level resolution did not correlate with the spread pattern of clades C5d and C6d, connectivity measures
393 between the southern region and Mexico City do mirror some of the transitions observed. Mobility
394 network analyses were also consistent with decrease in the number cases following epidemiological
395 peaks (as observed for the second and third wave of infection in the country). Thus, fluctuating human
396 mobility patterns likely reflect changes in the colour-coded system used to regulate restrictions
397 leveraged by the risk of infection (including travel), applied at a state level [78].

398 As geographic distances and human mobility within the country are not potential predictors of viral
399 spread, the patterns observed are most likely to be explained (at least partially) by both terrestrial and
400 aerial movements across borders. Nonetheless, determining the geographic origin of introduction
401 events is limited by genome sampling biases both at a global and local scale [79,80]. Such biases also
402 affects the viral spread reconstructions we present, likely to be impacted by *i*) an uneven genome
403 sampling across all Mexico states, by *ii*) an uneven sampling of foreign locations at a regional level, and
404 by *iii*) a poor viral genome representation for many countries (*i.e.*, Central and South America) [51,81].
405 Taking this into consideration, it has been hypothesized that the introduction and spread of SARS-CoV-
406 2 in Mexico would be strongly linked to mobility between Mexico and the USA (for an example, see [46]).
407 Although human movements across these countries have indeed contributed to spread of some
408 lineages (clearly for the B.1.1.222 and the B.1.1.519), some of the viral diffusion patterns observed
409 across national territory are more likely to be congruent with human mobility linked to migration routes
410 from South and Central America into Mexico. This raises the question of the potential role of irregular
411 migration on virus dispersal across countries, and further highlights the need of using multiple sources
412 for human mobility data across borders (such as mobile-usage, in addition to air traffic) to inform
413 phylogeographic analyses [82–85]. Our study further provides the basis for in-depth exploration of SARS-
414 CoV-2 genomic data from Mexico, and highlights the need to investigate the spread of transmission
415 chains at a state level resolution.

416

417 **METHODS (1592/3000 WORDS)**

418 **1. Data collation and initial sequence alignments**

419 Global genome datasets assigned to each PANGO lineage [10] under investigation (B.1.1.222, B.1.1.519,
420 B.1.1.7, P.1 and B.1.617.2) were downloaded with their associated metadata from the GISAID platform
421 as of November 30th 2021 [51,86]. The total number of sequences retrieved sampled globally for each virus
422 lineage were the following: B.1.1.222 = 3,461, B.1.1.519 = 19,246, B.1.1.7 = 913,868, P.1 = 87,452,
423 and B.1.617.2 = 2,166,874. This also included the virus genome sequences available from Mexico up
424 the sampling date of this study. Sampling date represents the peak for B.1.617.2 frequency in Mexico
425 before the introduction of 'Omicron' into the country [30,31]. Sequences were further excluded if they
426 displayed incomplete collection dates, and/or if >1000 nt shorter than full genome length, and/or if
427 showing >10% of sites coded as ambiguities (including N or X). Individual datasets were then processed
428 using the Nextclade pipeline to filter sequences according to sequence quality [87]. In addition, a set of
429 the earliest SARS-CoV-2 sequences sampled from 2019-2020 (including reference sequence Wuhan-
430 Hu-1, GenBank accession ID: MN908947.3), and a set of viral genomes representing early 'Pango'
431 diversity sampled up to 31/05/2020 were added for rooting purposes
432 (https://github.com/BernardoGG/SARS-CoV-2_Genomic_lineages_Ecuador). Output datasets were
433 mapped to the reference sequence Wuhan-Hu-1 (GenBank: MN908947.3) using Minimap2 [88] to
434 generate whole genome alignments. The main viral ORFs (Orf1ab and S) were extracted to generate
435 reduced-length alignments of approximately 25,000 bases long, comprising only the largest and most
436 phylogenetically informative coding genome regions, to exclude smaller ORFs, UTRs, and short
437 intergenic sequences.
438
439 **2. Migration data and phylogenetically-informed subsampling**
440 To provide an overview for global introductory events into Mexico as a proxy for dataset reduction, we
441 used openly available data describing anonymized relative human mobility flow into different
442 geographical regions based on mobile data usage [89,90]. For any given dataset, all 'non-Mexico'
443 sequences are sorted according to their location, selecting only the top 5 countries representing the
444 most intense human mobility flow into Mexico. Then, 'Mexico' sequences are subsampled to ~4,000 in
445 proportion to the total number of cases reported in the country across time (corresponding to
446 epidemiological weeks) according to publicly accessible epidemiological data [91]. This enables a
447 homogeneous epidemiological and temporal representation of genomes. In case of a given lineage
448 reporting sub-lineages, datasets are further reduced by selecting the top 5 sub-lineages that circulate(d)

449 within the country. For 'globally sampled sequences', datasets are further reduced randomly to an
450 approximate 1:1 ratio relative to the number of 'Mexico' sequences (keeping sequences with the earliest
451 and latest collection dates sampled both from Mexico and globally). This migration-informed approach,
452 coupled with balancing the number of sequences from different countries, aims to mitigate the effects
453 of geographical over-representation impacting phylogeographic analyses, in which clades tend to not
454 segregate fully, and certain regions may appear as more frequent seeding sources than they are in
455 reality [1].

456 For each dataset, a preliminary Maximum Likelihood (ML) tree is inferred using IQ-TREE (command
457 line: iqtree -s -m GTR+I+G -alrt 1000) [92]. The pipeline then uses a modified version of Treemmer v0.3
458 (https://github.com/fmenardo/Treemmer/releases) to reduce the size and redundancy of the resulting
459 ML trees with a minimal loss of diversity [93]. Subsampling is based on phylogenetic clustering patterns,
460 reducing the noise derived from clonal and phylogenetically overrepresented sequences. For this, the -
461 lm command is initially used to protect 'Mexico' sequences and those added for rooting purposes.
462 During the pruning iterations, the -pp command is then used to protect 'Mexico' clusters and pairs of
463 'non-Mexico' sequences that are immediately ancestral or directly descending from these. This results
464 in reduced datasets that render local computational analyses feasible, in which clades may appear to
465 be smaller relative to the raw counts of genomes publicly available, but reflect the sampled viral genetic
466 diversity. Datasets are used for re-estimating ML trees as described, to be then inputted for a time-
467 scaled phylogenetic analysis (see Methods section 4). Plots representing genome data distribution
468 before and after implementing our subsampling approach are available in the **Supplementary Figure**
469 **2**. The pipeline is publicly accessible at (https://github.com/rhysinward/Mexico_subsampling).
470

471 **3. Dataset construction**

472 Given the reduced size of the original B.1.1.222 dataset, all sequences retained after initial quality
473 filtering were used for analyses. This resulted in a 3,849-sequence alignment (including 760 genomes
474 from Mexico) used for ML tree inference using IQ-TREE (iqtree -s -m GTR+I+G -alrt 1000;
475 **Supplementary Data 1**). The resulting tree was further inputted for a time-scaled analysis (see
476 Methods section 4). All other datasets (B.1.1.519, B.1.1.7, P.1 and B.1.617.2) were processed under
477 the pipeline described above to render informative datasets for phylogeographic analysis (see Methods
478 section 2 and 4). The B.1.1.519 final dataset resulted in a 5,001-sequence alignment, including 2,501

479 genomes from Mexico. The B.1.1.7 final dataset resulted in a 7,049-sequence alignment, including
480 1,449 genomes from Mexico. The P.1 final dataset resulted in a 5,493-sequence alignment, including
481 2,570 genomes from Mexico. The B.1.617.2 final dataset resulted in a 5,994-sequence alignment,
482 including 3,338 genomes from Mexico (**Supplementary Data 1**). Individual datasets were then used
483 for ML tree inference as described, with the resulting trees used as input for time-scaled analyses
484 (described in Methods section 4).
485
486 **4. Time-scaled analysis**
487 Output ML trees were assessed for temporal signal using TempEst v1.5.3 [94], removing outliers and re-
488 estimating trees when necessary. Trees were then time-calibrated informed by tip sampling dates using
489 TreeTime [95] (command line: treetime -aln --tree --clock-rate 8e-4 --dates --keep-polytomies --clock-filter
490 0). Due to a low temporal signal, a fixed clock rate corresponding to a reported virus evolutionary rate
491 estimate of 8×10^{-4} substitutions per site per year was used [96,97]. Root-to-tip regression plots for the ML
492 trees (prior to time calibration and excluding rooting sequences) are available in the **Supplementary**
493 **Figure 2**. In order to investigate viral introductions into Mexico, the time-calibrated trees were used as
494 input for a discrete trait analysis (DTA, or 'discrete phylogeographic inference') using BEAST v1.10.4
495 to generate maximum clade credibility (MCC) trees [98–100]. Using fixed 'time-calibrated' tree as an input
496 for DTA in an effective way of circumventing the restrictions of computationally-expensive analyses on
497 large datasets [98]. However, although this approach allows inferring dates and number for introduction
498 events into the study area, it does not consider the uncertainty associated with phylogenetic inference
499 (*i.e.*, the dates of inferred MRCAs do not come with a credibility intervals). Introduction events into the
500 country are associated with the most recent common ancestor (MRCA) node of each 'Mexico' clade.
501 Following a similar strategy as described in du Plessis et al.[3] 'Mexico' clades were identified as those
502 composed by a minimum of two sister 'Mexico' sequences directly descending from another 'Mexico'
503 sequence. For all introduction events identified in the MCC trees, the mean and associated HPD interval
504 were assessed.
505 Analyses for the B.1.1.7 and the P.1 and B.1.617.2 datasets were performed to estimate the
506 number of introduction events into Mexico from other (unknown) geographic regions, and only two
507 locations were considered: 'Mexico' and 'other'. For the B.1.1.222 and B.1.1.519 dataset, we estimated
508 the number of transitions between Mexico and the USA, thus considering three distinct geographic

509 locations: 'Mexico', 'USA' and 'other'. Each Monte Carlo Markov Chain (MCMC) was run for 10^6
510 iterations, sampling every 10^3 iterations. To assess if sufficient effective sample size values (*i.e.,*
511 ESS>200) were associated with the estimated parameters, we then inspected MCMC convergence and
512 mixing using the program Tracer 1.7 [101]. To explore viral spread across the country, two additional DTAs
513 were performed within each of the largest monophyletic clades identified within the B.1.617.2+ tree
514 (C5d and C6d, see Results section). These additional DTAs were based on the definition of 33 distinct
515 sampling locations (all 32 states from Mexico + 'other' for genomic sequences sampled from outside
516 the country). Visualization of viral spread patterns was performed using SpreadViz
517 (https://spreadviz.org/home), an updated web implementation of the Spatial Phylogenetic
518 Reconstruction of Evolutionary Dynamics software SpreaD3 [102], also used to estimate Bayes Factor
519 Rate (BFR) values to identify well-supported transitions between locations [100] (**Supplementary Table**
520 **1**).
521

522 **5. Human mobility data analysis and exploring correlations with genomic data**

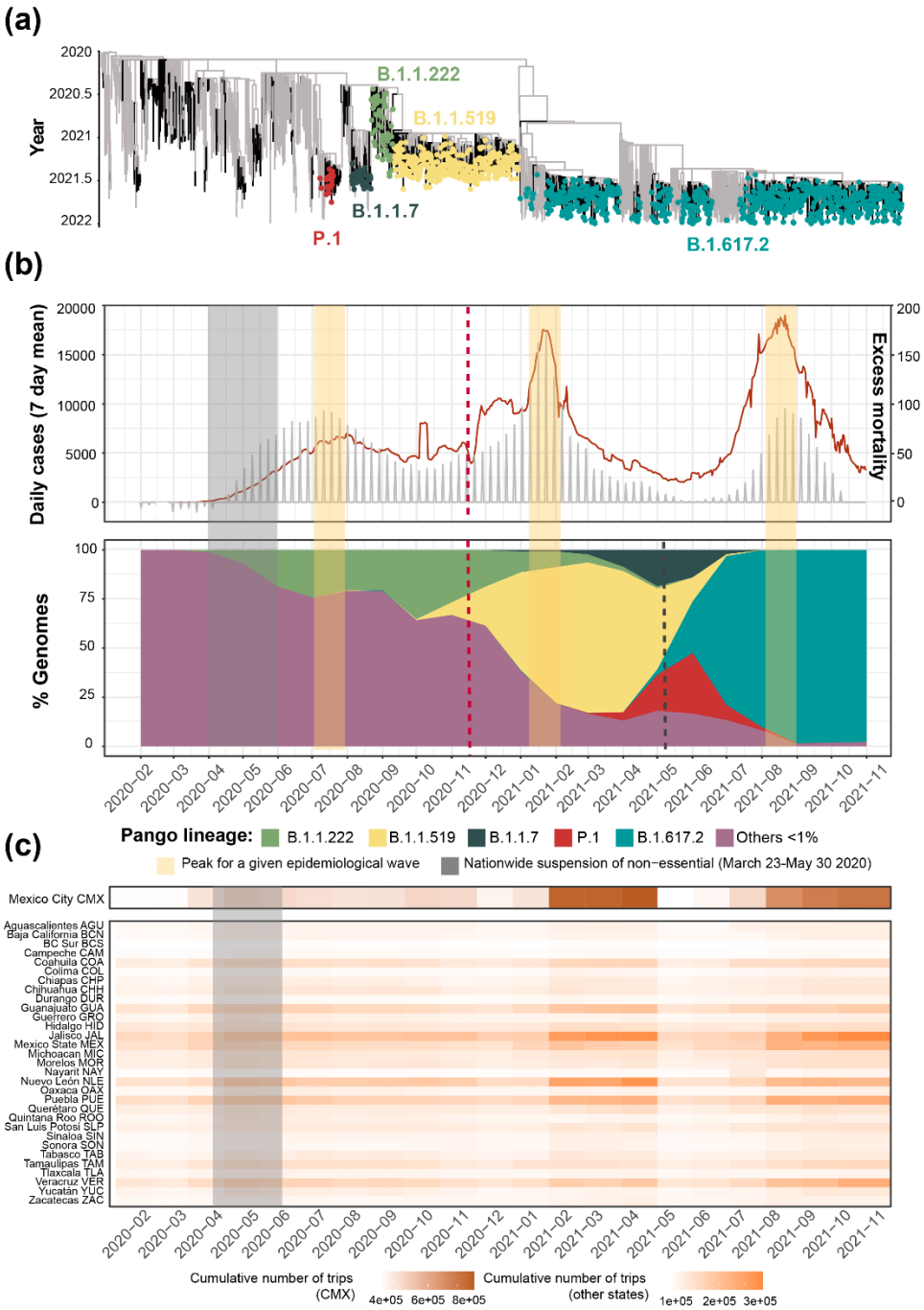
523 Human mobility data used for this study derived from anonymized mobile device locations provided by
524 the company Veraset collected between 01/01/2020 and 31/12/2021 within national territory. The
525 source dataset includes anonymized identifiers for mobile devices, geographical coordinates (latitude
526 and longitude) and a timestamp. The dataset was used to construct aggregated intra-state mobility
527 networks, where nodes are defined as each of the 32 states from the country, whilst (weighed and
528 directed) edges represent the normalized volume of observed trips between nodes [103]. The resulting
529 networks were then used to quantify the number of cumulative trips from any state into a given specific
530 state across time, the geographic distances among states, the mean intrastate connectivity observed
531 between April 2021 and November 2021 (corresponding to the expansion period for the B.1.617.2
532 lineage, see **Figure 4b**), and for ranking connectivity between the southern region of the country
533 (represented by Yucatán, Quintana Roo, Chiapas and Campeche) and the remaining 28 states (see
534 **Supplementary Figure 8,** matrixes available as **Supplementary Data 5**). The connectivity measure
535 was defined as the sum of the weights for edges that go from any given node into other node(s),
536 reflecting the number of trips in any direction. We then used the 'PhyCovA' software tool
537 (https://evolcompvir-kuleuven.shinyapps.io/PhyCovA/) to perform preliminary analysis for exploring the
538 human mobility data from the country as a potential predictor of viral transition among locations [104].

539 'PhyCovA' was chosen as an explanatory approach over a fully-integrated GLM implemented in the
540 Bayesian BEAST framework, as the last one would imply a high computational burden related to our
541 datasets [99].
542
543 **DATA AVAILABILITY**
544 Virus genome IDs and GISAID accession numbers for the sequences used in each dataset are provided
545 in the **Supplementary Data 1** file. All genomic and epidemiological data supporting the findings of this
546 study is publicly available from GISAID/GenBank, from the Ministry Of Health Mexico[105], and/or from
547 the 'Our world in data coronavirus pandemic web portal [29]. For the GISAID data used, the corresponding
548 acknowledgements table is available on the 'GISAID Data Acknowledgement Locator' under the
549 EPI_SET_20220405qd and EPI_SET_20220215at keys [51]. Our bioinformatic pipeline implementing a
550 migration data and phylogenetically-informed sequence subsampling approach is publicly available at
551 https://github.com/rhysinward/Mexico_subsampling.
552
553 **AUTHOR CONTRIBUTIONS**
554 HGCS, LD, BG, RPDI, SD, and MEZ performed research and analysed data. RPDI developed the
555 bioinformatic pipeline, whilst RPDI, NMV and CB processed data. MK, BG and RPDI, together with the
556 Mobility Team in Mexico [MHR, AM, MF, OF, PG, GOA, GAJ] performed the analysis of human mobility
557 data integrated with genomic data. Other members of CoViGen-Mex and ACG, BT, SZ, SL, CFA, AL
558 and MEZ contributed to the initial experimental design by leading evolutionary analyses part of the
559 duties of the consortium. OGP and SD contributed to revising the methodological approach. MEZ
560 supervised and coordinated data analysis. MEZ wrote the manuscript, with contributions and comments
561 from all authors.
562
563 **COMPETING INTERESTS**
564 The authors declare no competing interests.
565
566 **ACKNOWLEDGEMENTS**
567 HGCS is supported by funding through the "Vigilancia Genómica del Virus SARS-CoV-2 en México"
568 grant from the National Council for Science and Technology-México (CONACyT). BG is supported by

the 2017 Universities of Academic Excellence Scholarship Program of the Secretariat for Higher Education, Science, Technology and Innovation of the Republic of Ecuador (ARSEQ-BEC-003163-2017) and by Universidad San Francisco de Quito. SD acknowledges support from the *Fonds National de la Recherche Scientifique* (F.R.S.-FNRS, Belgium; grant n°F.4515.22), from the Research Foundation - Flanders (*Fonds voor Wetenschappelijk Onderzoek-Vlaanderen*, FWO, Belgium; grant n°G098321N), and from the European Union Horizon 2020 project MOOD (grant agreement n°874850). MEZ is supported currently supported by Leverhulme Trust ECR Fellowship (ECF-2019-542). OGP acknowledges support of the Oxford Martin School. MK and RPDI acknowledge support from the European Union Horizon 2020 project MOOD (#874850). The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of the European Commission. The mobility team [MHR, AM, OF, MF, PG, GO, GAJ] and AHE are supported by 'Fondo Conjunto de Cooperación México-Uruguay' (Agencia Mexicana de Cooperación Internacional para el Desarrollo). CFA acknowledges support from grants "Vigilancia Genómica del Virus SARS-CoV-2 en México-2022" (PP-F003) from the National Council for Science and Technology-México (CONACyT), grant 057 from the "Ministry of Education, Science, Technology and Innovation (SECTEI) of Mexico City", and grant "Genomic surveillance for SARS-CoV-2 variants in Mexico" from the AHF Global Public Health Institute at the University of Miami. AL work was supported by DGAPA-PAPIIT (IN214421) and DGAPA-PAPIME (PE204921) of UNAM. We thank Verity Hill, Philippe Lemey, Tim Blokker and Sam Hong for their valuable advice on the technical details related to the methodology used for the time-scaled analyses. We thank all members of the Consorcio Mexicano de Vigilancia Genómica (CoViGen-Mex) for their efforts on sample collection and generating genetic sequence and metadata. Particularly, we thank León Martínez-Castilla and José Campillo Balderas for their contributions in the initial collation of preliminary data. We gratefully acknowledge all data contributors for the GISAID sequence data: *i.e.* the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID initiative, on which this research is based.

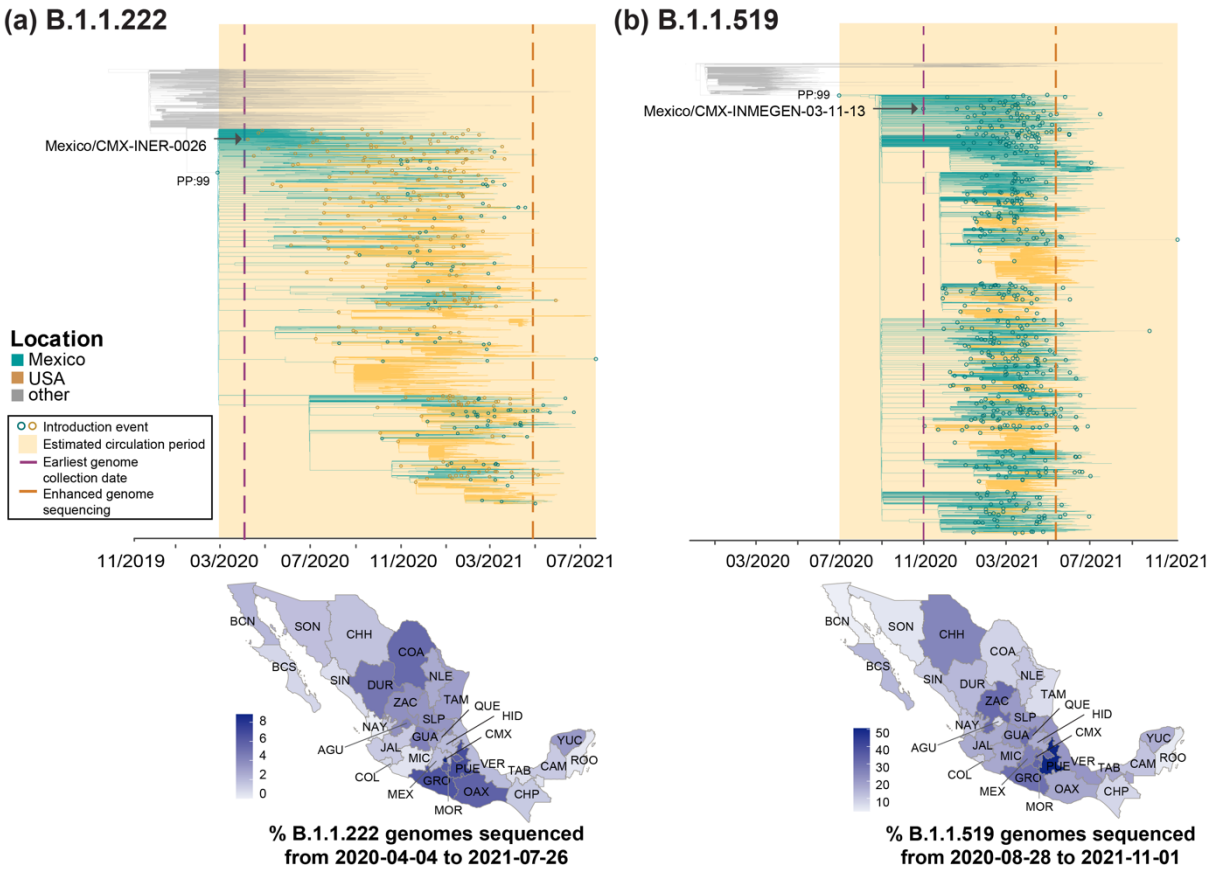
FIGURES

Figure 1. Overview of the SARS-CoV-2 epidemic in Mexico



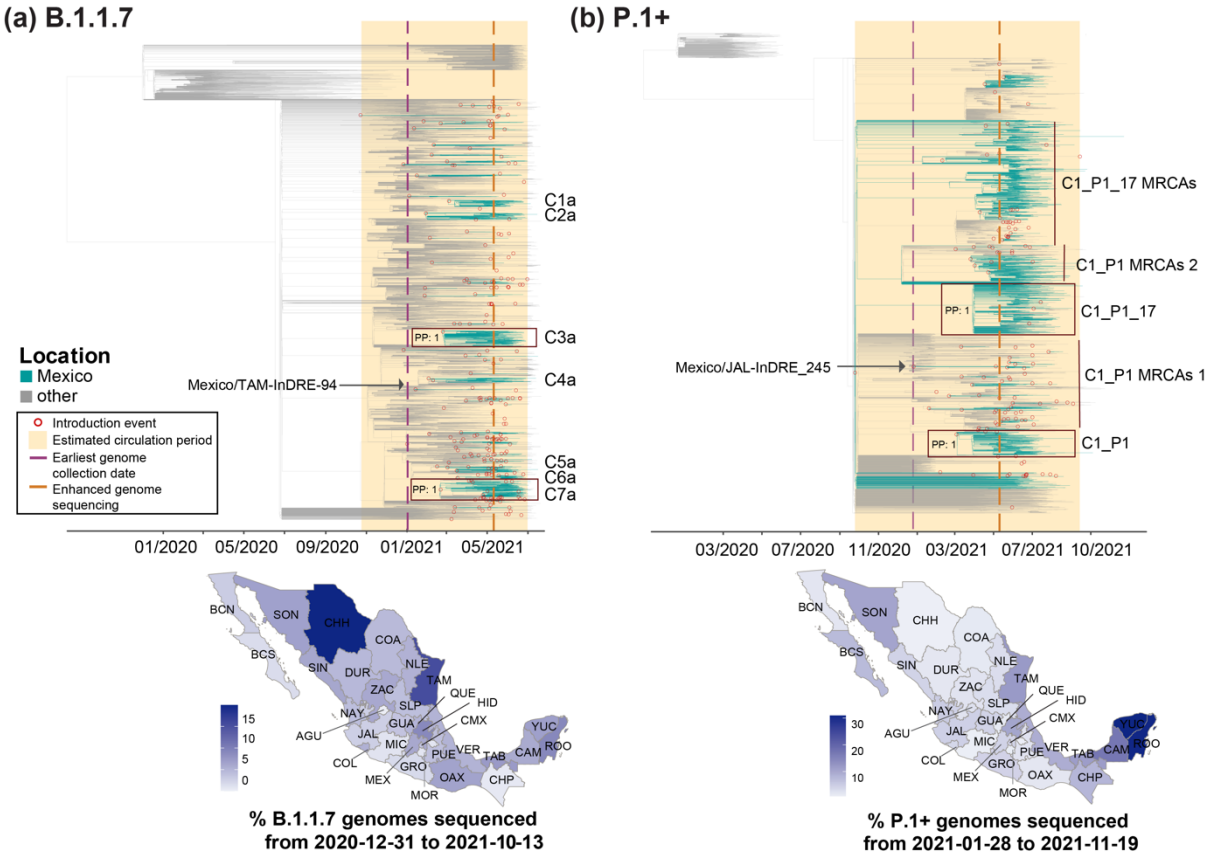
(a) Time-scaled phylogeny for representative viral genomes from Mexico within a global context, highlighting the positioning of the B.1.1.222, B.1.1.519, B.1.1.7, P.1 and B.1.617.2 virus lineages. (b) The epidemic curve for COVID-19 in Mexico from January 2020 up to November 2021, showing the average number of daily cases (red line) and the associated excess mortality (grey line). The peak for the first, second and third waves of infection are highlighted in yellow shadowing, whilst the lower panel represents the percentage of SARS-CoV-2 genomes from Mexico assigned to distinct 'Pango' lineages within the given timeframe. Lineage B.1.1.222 is shown in light green, B.1.1.519 in yellow, P.1 in red (Gamma), B.1.1.7 (Alpha) in dark green, and B.1.617.2 (Delta) in teal. Other lineages displaying a lower sampling frequency are jointly shown in purple. The time period for the implementation of non-pharmaceutical interventions at national scale is highlighted in grey shadowing. (c) Heatmap displaying the volume of trips into a given state from any other state recorded from January 2020 up to November 2021 derived from anonymized mobile device geolocated and time-stamped data. The dashed red line corresponds to the start date national vaccination campaign (December 2020), whilst the dashed black line represents the implementation date of a broader viral genome sampling and sequencing scheme for the surveillance of SARS-CoV-2 in Mexico (after May 2021).

Figure 2. Time-scaled phylogenetic analyses for the B.1.1.222 and B.1.1.519 lineage



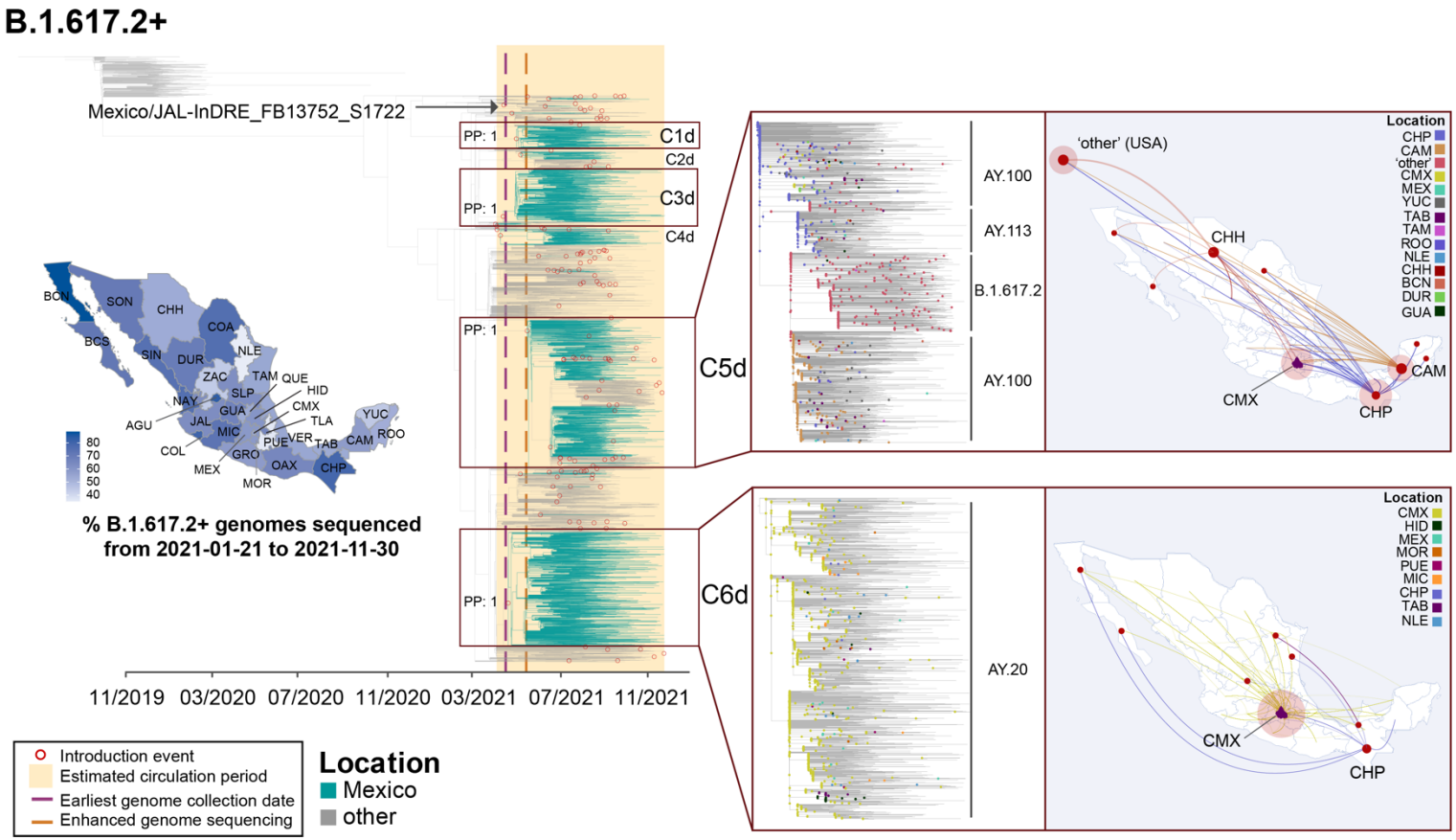
Maximum clade credibility (MCC) trees for the (a) B.1.1.222 and (b) B.1.1.519 lineages, in which the main clades corresponding to distinct introduction events into Mexico are highlighted. Nodes shown as outline circles correspond to the most recent common ancestor (MRCA) of clades resulting from a distinct introduction event into Mexico (in teal) or the USA (in ochre). The estimated circulation period in Mexico for a given lineage is highlighted in yellow shadowing. The dashed purple line represents the collection date for the earliest lineage-assigned genome sampled from Mexico, with its position indicated with an arrow. The dashed yellow line represents the implementation date (May 2021) of a broader viral genome sequencing scheme for the surveillance of SARS-CoV-2 in Mexico. Below, map graphs show the cumulative proportion of viral genomes from Mexico assigned to each virus lineage relative to the total number of genome sequences reported for any given state (taken from GISAID, corresponding to raw data before subsampling).

Figure 3. Time-scaled phylogenetic analyses for the B.1.1.7 and P.1+ lineages



Maximum clade credibility (MCC) trees for the (a) B.1.1.7 and the (b) P.1+ B.1.617.2 lineages, in which major clades identified as distinct introduction events into Mexico are highlighted. Nodes shown as red outline circles correspond to the most recent common ancestor (MRCA) of clades resulting from a distinct introduction event into Mexico. The estimated circulation period for a given lineage is highlighted in yellow shadowing. For each panel, the dashed purple line represents the collection date for the earliest lineage-assigned genome sampled from Mexico, with its position within the tree indicated with an arrow. The dashed yellow line represents the implementation date (May 2021) of a broader viral genome sampling and sequencing scheme for the surveillance of SARS-CoV-2 in Mexico. Below, map graphs show the cumulative proportion of viral genomes from Mexico assigned to each virus lineage of interest relative to the total number of genome sequences reported for any given state (taken from GISAID, corresponding to raw data before subsampling).

Figure 4. Time-scaled and phylogeographic analysis for the B.1.617.2+ lineage



Maximum clade credibility (MCC) tree for the B.1.617.2+ lineage, in which major clades identified as distinct introduction events into Mexico are highlighted. Nodes shown as red outline circles correspond to the most recent common ancestor (MRCA) of clades resulting from a distinct introduction event into Mexico. The estimated circulation period is highlighted in yellow shadowing. The dashed purple line represents the collection date for the earliest lineage-assigned genome sampled from Mexico, with its position within the tree indicated with an arrow. The dashed yellow line represents the implementation date (May 2021) of a broader viral genome sampling and sequencing scheme for the surveillance of SARS-CoV-2 in Mexico. On the left, a map graph sows the cumulative proportion of sequenced viral genomes from Mexico assigned as B.1.617.2+, relative to the total number of genome sequences reported for any given state (taken from GISAID, corresponding to raw data before subsampling). On the right, a zoom-in to the C5d and C6d clades shows sub-lineage composition and the inferred location for nodes. The geographic spread of these clades across Mexico is shown from the inferred discrete phylogeographic analysis (see Methods section 4). Viral lineage transition events between Mexican states are represented by curved lines coloured according to sampling location, showing only well-supported transitions (Bayes factor rate >100 and a posterior probability >0.9). Bayes factor rate values are all reported in the Supplementary Table 1.

REFERENCES

1. Hill, V., Ruis, C., Bajaj, S., Pybus, O. G. & Kraemer, M. U. G. Progress and challenges in virus genomic epidemiology. *Trends Parasitol.* **37**, 1038–1049 (2021).
2. Kraemer, M. U. G. *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021).
3. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
4. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
5. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
6. COVID-19 Genomics UK (COG-UK) consortiumcontact@cogconsortium.uk. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100 (2020).
7. Chen, Z. *et al.* A global analysis of replacement of genetic variants of SARS-CoV-2 in association with containment capacity and changes in disease severity. *Clin. Microbiol. Infect.* **27**, 750–757 (2021).
8. Li, J., Lai, S., Gao, G. F. & Shi, W. The emergence, genomic diversity and global spread of SARS-CoV-2. *Nature* **600**, 408–418 (2021).
9. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).
10. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020).
11. Pango Network – Helping track the transmission and spread of SARS-CoV-2. https://www.pango.network/.
12. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.
13. CDC. Science brief: Emerging SARS-CoV-2 variants. *Centers for Disease Control and Prevention* https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/scientific-brief-emerging-variants.html?CDC_AA_refVal=https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html (2022).
14. Horby, P. *et al.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/961037/NERVTAG_note_on_B.1.1.7_severity_for_SAGE_77__1_.pdf.
15. Weisblum, Y. *et al.* Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife* **9**, e61312 (2020).
16. Greaney, A. J. *et al.* Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.* **12**, 4196 (2021).
17. Greaney, A. J. *et al.* Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* **29**, 44-57.e9 (2021).
18. Classification of Omicron (B.1.1.529): SARS-CoV-2 variant of Concern. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern.
19. Cov-lineages. https://cov-lineages.org/.
20. Vöhringer, H. S. *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).
21. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv* 2021.02.23.21252268 (2021).
22. Escalera-Zamudio, M. *et al.* Identification of site-specific evolutionary trajectories shared across human betacoronaviruses. *bioRxivorg* (2021) doi:10.1101/2021.05.24.445313.
23. Kumar, S. *et al.* An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. *Mol. Biol. Evol.* **38**, 3046–3059 (2021).
24. Caniels, T. G. *et al.* Emerging SARS-CoV-2 variants of concern evade humoral immune responses from infection and vaccination. *Sci. Adv.* **7**, eabj5365 (2021).
25. Zhang, X. *et al.* Viral and host factors related to the clinical outcome of COVID-19. *Nature* **583**, 437–440 (2020).
26. Boutin, S. *et al.* Host factors facilitating SARS-CoV-2 virus infection and replication in the lungs. *Cell. Mol. Life Sci.* **78**, 5953–5976 (2021).
27. https://globalhealthsciences.ucsf.edu/sites/globalhealthsciences.ucsf.edu/files/mexico-covid-19-case-study-english.pdf.
28. Taboada, B. *et al.* Genomic analysis of early SARS-CoV-2 variants introduced in Mexico. *J. Virol.* **94**, (2020).
29. Ritchie, H. *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).
30. MexCoV2. http://mexcov2.ibt.unam.mx:8080/COVID-TRACKER/.
31. Vigilancia de variantes del virus SARS-CoV-2. *Vigilancia de variantes del virus SARS-CoV-2* https://salud.conacyt.mx/coronavirus/variantes/.
32. Taboada, B. *et al.* Genetic analysis of SARS-CoV-2 variants in Mexico during the first year of the COVID-19 pandemic. *Viruses* **13**, 2161 (2021).
33. MexCoV2. http://mexcov2.ibt.unam.mx:8080/COVID-TRACKER/tablero.
34. Gutierrez, B. *et al.* Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. *Cell Host Microbe* (2022) doi:10.1016/j.chom.2022.06.010.
35. https://www.gob.mx/cms/uploads/attachment/file/571250/Comunicado_Tecnico_Diario_COVID-19_2020.08.19__1_.pdf.
36. https://www.gob.mx/cms/uploads/attachment/file/626111/Comunicado_Tecnico_Diario_COVID-19_2021.03.30.pdf.
37. https://www.gob.mx/cms/uploads/attachment/file/648612/Comunicado_Tecnico_Diario_COVID-19_2021.06.25.pdf.
38. Cov-lineages. https://cov-lineages.org/lineages/lineage_B.1.1.222.html.
39. https://cov-lineages.org/lineages/lineage_B.1.1.519.html,.
40. Cedro-Tanda, A. *et al.* The evolutionary landscape of SARS-CoV-2 variant B.1.1.519 and its clinical impact in Mexico city. *Viruses* **13**, 2182 (2021).
41. Rodríguez-Maldonado, A. P. *et al.* Emergence and spread of the potential variant of interest (VOI) B.1.1.519 of SARS-CoV-2 predominantly present in Mexico. *Arch. Virol.* **166**, 3173–3177 (2021).
42. Cov-lineages. https://cov-lineages.org/lineages/lineage_B.1.1.519.html.
43. Outbreak.Info. *outbreak.info* https://outbreak.info/situation-reports?pango=B.1.1.519.
44. Cov-lineages. https://cov-lineages.org/global_report_B.1.1.7.html.
45. https://www.gob.mx/cms/uploads/attachment/file/655969/Comunicado_Tecnico_Diario_COVID-19_2021.07.21.pdf.
46. Zárate, S. *et al.* The Alpha Variant (B.1.1.7) of SARS-CoV-2 Failed to Become Dominant in Mexico. *Microbiol Spectr* **10**, e0224021 (2022).

47. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).
48. Cov-lineages. https://cov-lineages.org/global_report_B.1.617.2.html.
49. Tian, D., Sun, Y., Zhou, J. & Ye, Q. The global epidemic of the SARS-CoV-2 Delta variant, key spike mutations and immune escape. *Front. Immunol.* **12**, 751778 (2021).
50. https://cov-lineages.org/lineage_list.html,.
51. GISAID - initiative. https://www.gisaid.org/.
52. SSA/SPPS/DGE/DIE/InDRE/UIES/Informe técnico. COVID. https://www.gob.mx/cms/uploads/attachment/file/684495/Comunicado_Tecnico_Diario_COVID-19_2021.11.30.pdf.
53. Taboada, B. *et al.* Dominance of three sublineages of the SARS-CoV-2 Delta variant in Mexico. *Viruses* **14**, 1165 (2022).
54. Cov-lineages. https://cov-lineages.org/lineage.html?lineage=AY.20.
55. Cov-lineages. https://cov-lineages.org/lineage.html?lineage=AY.26.
56. Cov-lineages. https://cov-lineages.org/lineage.html?lineage=AY.100.
57. Cov-lineages. https://cov-lineages.org/lineage.html?lineage=AY.113.
58. Cov-lineages. https://cov-lineages.org/lineage.html?lineage=AY.3.
59. Taboada, B. *et al.* Dominance of three sublineages of the SARS-CoV-2 Delta variant in Mexico. (2022) doi:10.20944/preprints202205.0131.v1.
60. Washington, N. L. *et al.* Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **184**, 2587-2594.e7 (2021).
61. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
62. https://rss.org.uk/RSS/media/File-library/Publications/Special.
63. Liu, C. *et al.* Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. *Cell* **184**, 4220-4236.e13 (2021).
64. Wilhelm, A. *et al.* Antibody-mediated neutralization of authentic SARS-CoV-2 B.1.617 variants harboring L452R and T478K/E484Q. *Viruses* **13**, 1693 (2021).
65. Haynes, W. A., Kamath, K., Lucas, C., Shon, J. & Iwasaki, A. Impact of B.1.1.7 variant mutations on antibody recognition of linear SARS-CoV-2 epitopes. *bioRxiv* 2021.01.06.20248960 (2021) doi:10.1101/2021.01.06.20248960.
66. Lubinski, B. *et al.* Functional evaluation of the P681H mutation on the proteolytic activation the SARS-CoV-2 variant B.1.1.7 (Alpha) spike. *bioRxivorg* (2021) doi:10.1101/2021.04.06.438731.
67. Lista, M. J. *et al.* The P681H mutation in the Spike glycoprotein confers Type I interferon resistance in the SARS-CoV-2 alpha (B.1.1.7) variant. *bioRxiv* 2021.11.09.467693 (2021) doi:10.1101/2021.11.09.467693.
68. Dolan, P. T., Whitfield, Z. J. & Andino, R. Mapping the evolutionary potential of RNA viruses. *Cell Host Microbe* **23**, 435–446 (2018).
69. Emergence and spread of SARS-CoV-2 P.1 (Gamma) lineage variants carrying Spike mutations Δ141-144, N679K or P681H during persistent viral circulation in Amazonas, Brazil. *Virological* https://virological.org/t/emergence-and-spread-of-sars-cov-2-p-1-gamma-lineage-variants-carrying-spike-mutations-141-144-n679k-or-p681h-during-persistent-viral-circulation-in-amazonas-brazil/722 (2021).
70. CoVariants: 20J (gamma, V3). https://covariants.org/variants/20J.Gamma.V3.
71. auspice. https://nextstrain.org/groups/neherlab/ncov/S.P681?c=gt-S_681.
72. https://observablehq.com/@spond/spike-trends,.
73. CoVariants: S:P681. https://covariants.org/variants/S.P681.
74. Gupta, S. & Anderson, R. M. Population structure of pathogens: the role of immune selection. *Parasitol. Today* **15**, 497–501 (1999).
75. Bhatia, G., Dutta, P. K., McClure, J. & Reuters Graphics. Mexico: the latest coronavirus counts, charts and maps. *Reuters* (2020).
76. Basto-Abreu, A. *et al.* Nationally representative SARS-CoV-2 antibody prevalence estimates after the first epidemic wave in Mexico. *Nat. Commun.* **13**, 589 (2022).
77. Muñoz-Medina, J. E. *et al.* SARS-CoV-2 IgG antibodies seroprevalence and Sera neutralizing activity in MEXICO: A national cross-sectional study during 2020. *Microorganisms* **9**, 850 (2021).
78. https://coronavirus.gob.mx/semaforo/.
79. Kalkauskas, A. *et al.* Sampling bias and model choice in continuous phylogeography: Getting lost on a random walk. *PLoS Comput. Biol.* **17**, e1008561 (2021).
80. De Maio, N., Wu, C.-H., O'Reilly, K. M. & Wilson, D. New routes to phylogeography: A Bayesian STructured coalescent Approximation. *PLoS Genet.* **11**, e1005421 (2015).
81. No title. https://academic.oup.com/ve/article/7/2/veab051/6292076?login=true.
82. Kraemer, M. U. G. *et al.* The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **368**, 493–497 (2020).
83. Migration data in central America. *Migration data portal* https://www.migrationdataportal.org/regional-data-overview/migration-data-central-america.
84. Migrants trapped in the Mexican vertical border. *Oxford Law Faculty* https://www.law.ox.ac.uk/research-subject-groups/centre-criminology/centreborder-criminologies/blog/2018/06/migrants-trapped (2018).
85. París-Pombo, M. D. & El Colegio de la Frontera Norte, México. Trayectos peligrosos: inseguridad y movilidad humana en México. *Papeles Poblac.* **22**, 145–172 (2016).
86. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, (2017).
87. Aksamentov, I., Roemer, C., Hodcroft, E. & Neher, R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021).
88. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
89. Kraemer, M. U. G. *et al.* Mapping global variation in human mobility. *Nat. Hum. Behav.* **4**, 800–810 (2020).
90. Inward, R. P. D., Parag, K. V. & Faria, N. R. Using multiple sampling strategies to estimate SARS-CoV-2 epidemiological parameters from genomic sequencing data. *bioRxiv* 2022.02.04.22270165 (2022) doi:10.1101/2022.02.04.22270165.
91. COVID-19 tablero México. *COVID - 19 Tablero México* https://datos.covid-19.conacyt.mx/.
92. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

93. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, (2018).
94. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
95. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
96. Su, Y. C. F. *et al.* Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. *MBio* **11**, (2020).
97. MacLean, O. A., Orton, R. J., Singer, J. B. & Robertson, D. L. No evidence for distinct types in the evolution of SARS-CoV-2. *Virus Evol.* **6**, veaa034 (2020).
98. Dellicour, S. *et al.* A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. *Mol. Biol. Evol.* **38**, 1608–1613 (2021).
99. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
100. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520 (2009).
101. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
102. Bielejec, F. *et al.* SpreaD3: Interactive visualization of spatiotemporal history and trait evolutionary processes. *Mol. Biol. Evol.* **33**, 2167–2169 (2016).
103. Fontanelli, O. *et al.* Intermunicipal travel networks of Mexico (2020-2021). *arXiv [cs.SI]* (2022).
104. Blokker, T., Baele, G., Lemey, P. & Dellicour, S. Phycova - a tool for exploring covariates of pathogen spread. *Virus Evol.* **8**, veac015 (2022).
105. https://covid19.sinave.gob.m.