

## Genomic analysis of early SARS-CoV-2 isolates introduced in Mexico

Blanca Taboada<sup>1\*</sup>, Joel Armando Vazquez-Perez<sup>2\*</sup>, José Esteban Muñoz Medina<sup>3\*</sup>, Pilar Ramos Cervantes<sup>4\*</sup>, Marina Escalera-Zamudio<sup>5\*</sup>, Celia Boukadida<sup>6</sup>, Alejandro Sanchez-Flores<sup>7</sup>, Pavel Isa<sup>1</sup>, Edgar Mendieta Condado<sup>8</sup>, José Arturo Martínez-Orozco<sup>2</sup>, Eduardo Becerril-Vargas<sup>2</sup>, Jorge Salas-Hernández<sup>2</sup>, Ricardo Grande<sup>7</sup>, Carolina González-Torres<sup>9</sup>, Francisco Javier Gaytán-Cervantes<sup>9</sup>, Gloria Vazquez<sup>7</sup>, Francisco Pulido<sup>7</sup>, Adnan Araiza Rodríguez<sup>8</sup>, Fabiola Garcés Ayala<sup>8</sup>, Cesar Raúl González Bonilla<sup>10</sup>, Concepción Grajales Muñiz<sup>11</sup>, Víctor Hugo Borja Aburto<sup>12</sup>, Gisela Barrera Badillo<sup>8</sup>, Susana López<sup>1</sup>, Lucía Hernández Rivas<sup>8</sup>, Rogelio Perez-Padilla<sup>2</sup>, Irma López Martínez<sup>8</sup>, Santiago Ávila-Ríos<sup>6</sup>, Guillermo Ruiz-Palacios<sup>4</sup>, José Ernesto Ramírez-González<sup>8\*</sup>, and Carlos F. Arias<sup>1\*</sup>

<sup>1</sup>Departamento de Genética del Desarrollo y Fisiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico; <sup>2</sup>Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Ciudad de México, Mexico; <sup>3</sup>División de Laboratorios de Vigilancia e Investigación Epidemiológica, Instituto Mexicano del Seguro Social, Ciudad de México, México; <sup>4</sup>Instituto Nacional de Ciencias Médicas y Nutrición, Ciudad de México, Mexico; <sup>5</sup>Department of Zoology, Oxford University, Oxford, UK; <sup>6</sup>Centro de Investigación en Enfermedades Infecciosas, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Ciudad de México, Mexico; <sup>7</sup>Unidad Universitaria de Secuenciación Masiva y Bioinformática, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico; <sup>8</sup>Instituto de Diagnóstico y Referencia Epidemiológicos, Dirección General de Epidemiología, Ciudad de México, Mexico; <sup>9</sup>División de Desarrollo de la Investigación, Instituto Mexicano del Seguro Social, Ciudad de México, México; <sup>10</sup>Coordinación de Investigación en Salud, Instituto Mexicano del Seguro Social, Ciudad de México, México; <sup>11</sup>Coordinación de Control Técnico de Insumos, Instituto Mexicano del Seguro Social, Ciudad de México, México; <sup>12</sup>Dirección de Prestaciones Médicas, Instituto Mexicano del Seguro Social, Ciudad de México, Mexico.

\*These authors contributed equally to this work.

+Corresponding authors: José Ernesto Ramírez-González, and Carlos F. Arias

## ABSTRACT

The COVID-19 pandemic has affected most countries in the world. Studying the evolution and transmission patterns in different countries is crucial to implement effective strategies for disease control and prevention. In this work, we present the full genome sequence for 17 SARS-CoV-2 isolates corresponding to the earliest sampled cases in Mexico. Global and local phylogenomics, coupled with mutational analysis, consistently revealed that these viral sequences are distributed within 2 known lineages, the SARS-CoV-2 lineage A/G, containing mostly sequences from North America, and the lineage B/S containing mainly sequences from Europe. Based on the exposure history of the cases and on the phylogenomic analysis, we characterized fourteen independent introduction events. Additionally, three cases with no travel history were identified. We found evidence that two of these cases represent local transmission cases occurring in Mexico during mid-March 2020, denoting the earliest events described for the country. Within this local transmission cluster, we also identified the H49Y amino acid change in the Spike protein. This mutation is a homoplasy occurring independently through time and space, and may function as a molecular marker to follow on any further spread of these viral variants throughout the country. Our results depict the general picture of the SARS-CoV-2 variants introduced at the beginning of the outbreak in Mexico, setting the foundation for future surveillance efforts.

## INTRODUCTION

The 2019 coronavirus disease (COVID-19), declared a pandemic by the WHO on March 11<sup>th</sup>, 2020<sup>1</sup>, is caused by a novel betacoronavirus known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), detected in December of 2019 in the province of Wuhan in China<sup>1</sup>. This is the third outbreak related to zoonotic betacoronaviruses known to occur in humans in the last two decades, after SARS (severe acute respiratory syndrome) in 2002 and MERS (Middle East respiratory syndrome) in 2012. After its emergence in China, SARS-CoV-2 spread initially to other parts of the world by people with a travel history to China, but gradually shifted to local transmissions<sup>2</sup>. Viral spread was first detected in Thailand, South Korea and Japan, and by the second half of January, the first positive cases appeared in the USA and Europe (France, Italy and Spain). The current SARS-CoV-2 genome analysis in the Nextstrain site<sup>3</sup>, points out that viral transmission is now mainly community-driven<sup>4,5</sup>.

In many countries, despite diagnostic efforts and initial control strategies, SARS-CoV-2 spread went on undetected until a critical number of cases requiring hospitalization and intensive care was reached, alerting the authorities in charge. As for May 4<sup>th</sup> 2020, SARS-CoV-2 has infected more than 3,578,000 people and caused around 251,000 deaths worldwide<sup>3</sup>. In Mexico, the first case of SARS-CoV-2 was detected on February 27<sup>th</sup> 2020, corresponding to a person who travelled back to Mexico from Italy, and who was in direct contact with a confirmed SARS-CoV-2 case. Soon after, additional cases were detected among travelers that returned from the USA and Europe, increasing every day. By May 4<sup>th</sup>, there were over 23,400 confirmed cases and 2,150 deaths within the country, indicating local transmission<sup>6</sup>. Understanding the introduction, spread and establishment of SARS-CoV-2 within distinct human populations is crucial to implement effective control strategies. In this work, we studied the early introduction dynamics of the first SARS-CoV-2 cases in Mexico. For this, we used a whole genome sequencing and phylogenomic approach. We obtained 17 full viral genome sequences, including the first case detected and sampled within the country. Phylogenomic placement showed that these viruses belong to the A2/G and B/S lineages, two of the three circulating viral lineages reported so far. Our analysis also confirms that there have been multiple independent introduction events in Mexico from travelers abroad. We also found evidence for early local transmission of viral isolates having the mutation H49Y in the Spike protein, that could be further used as a molecular marker to follow viral spread within the country.

## METHODS

### **Ethical statement**

All clinical samples were processed at the “Instituto de Diagnóstico y Referencia Epidemiológicos” (InDRE), following official procedures<sup>7</sup>. All samples used for this work are considered part of the national response to COVID-19, and the data collected is directly related to disease control.

### **Sample collection and diagnostics**

All samples used in this study were collected under the Mexican Official Norm NOM-024-SSA2-1994 for prevention and control of acute respiratory infections in the primary health attention, as part of the early diagnostics scheme for SARS-CoV-2 in public health laboratories and hospitals in Mexico City (Red Nacional de Laboratorios Estatales de Salud Pública, RNLSP; Instituto Nacional de Enfermedades Respiratorias, INER; Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, INCMNSZ; and Instituto Mexicano del Seguro Social, IMSS). Oro- and naso-pharyngeal swabs were collected and placed in virus transport medium upon collection, following InDRE official procedures<sup>8</sup>. A tracheal aspirate was also obtained from one patient and it was frozen at -70°C until use. Diagnosis was done using validated protocols for SARS-CoV 2, as approved by InDRE and by the World Health Organization (WHO)<sup>9</sup>.

### **Sample processing and Whole Genome Sequencing**

All samples were prepared for RNA extraction, as described<sup>10,11</sup>. Briefly, centrifuged and filtered supernatants were treated with Turbo DNase and RNase. Nucleic acids were then extracted using the PureLink™ Viral RNA/DNA Kit (ThermoFisher), following the manufacturer’s instructions and using linear acrylamide (Ambion) as RNA carrier. cDNA was synthesized using the SuperScript III Reverse Transcriptase System (ThermoFisher) and primer A (5'-GTTTCCCAGTAGGTCTCN9-3'). The second strand was generated by two rounds of synthesis with Sequenase 2.0, followed by 15 cycles of amplification using Phusion DNA polymerase using primer B (5'-GTTTCCCAGTAGGTCTC-3). Next, cDNA was purified using the DNA Clean & Concentrator Kit (Zymo Research) and used as input material for generating sequencing libraries, following the Nextera XT DNA Library Preparation Kit<sup>12</sup> (Illumina). Finally, all samples were sequenced on the Illumina NextSeq 500 platform using a 150 cycle High Output Kit v2.5 to obtain paired end reads of 75 base pairs. Sequencing yields are reported in SI Table 1.

### **Bioinformatic analysis**

### *Data quality control and processing*

Read quality control was carried out using FAST-QC<sup>13</sup> under the default parameters. Adapter sequences and low-quality bases were removed using Fastp v0.19<sup>14</sup>. Low complexity reads, those with a length shorter than 40 bases, and duplicates were excluded using CD-HIT-DUP v.4.6.8<sup>15</sup>. Off-target reads were then filtered out using Bowtie2 v2.3.4.3 with the default parameters against the human genome version GRCh38.p13, and the SILVA database<sup>16</sup> as a reference to filter out human DNA and ribosomal sequences.

### *Viral genome assembly*

The reads obtained were used as input to assemble viral genomes using the Wuhan-Hu-1 reference genome sequence (MN908947). For this, the reads obtained for each sample were mapped against the reference using Bowtie2 v2.3.4.3. Aligned reads were then used for de novo assembly using SPAdes v17. Consensus genome sequences were generated using the majority threshold criterion. Only sequences with a coverage above 80% and a mean depth of  $\geq 8X$  were considered for the analyses (SI Table 1).

## **Phylogenetic analyses**

### *Data collation*

From 4698 complete SARS-CoV-2 genomes deposited in the Global Initiative on Sharing All Influenza Data (GISAID) platform on the morning of April 7<sup>th</sup> 2020, a total of 3014 sequences genomes ( $>29,000$  nt and only high coverage) were downloaded to generate a local database. As the collection dates of the Mexican samples range from late February to March, we filtered sequences collected between 02/01/2020 to 03/31/2020 for our database (Table 1). From these, unique sequences were extracted and those identical were collapsed, leaving 2633. We then included the 17 consensus viral genomes determined in this study and the Wuhan-Hu-1 reference genome sequence, yielding a total of 2651 sequences. We aligned the whole genome nt dataset using MUSCLE v3.8<sup>18</sup> under default parameters, and then used the *getorf*s script from the EMBOSS suite<sup>19</sup> to extract complete ORFs (Open Reading Frames) above 300 nt (Orf1a, Orf1b, Spike, M, Orf3a, Orf7a, Orf8 and N), that were then individually re-aligned as described<sup>18</sup>. Finally, to exclude UTRs and non-coding intergenic regions from the phylogenomic analyses, individual ORFs were concatenated to generate an additional 28,320 nt-long whole genome (WG) alignment.

### *Data subsampling and tree inference*

The individual and concatenated alignments were then reversed to nucleotides and used for estimating maximum likelihood (ML) trees using RAxML v8<sup>20</sup> under the following parameters: -T 2 -f a -x 390 -m GTRGAMMA -p 580 -N 100. All trees were rooted on Wuhan-Hu-1 reference genome sequence<sup>21</sup>. Given that the SARS-CoV-2 virus shows a low degree of genetic variation<sup>22</sup>, lineage definition must be based on consensus branching patterns within different trees and by shared nucleotide substitution patterns<sup>23</sup>, in addition to bootstrap support values. Based on these criteria, the position of the Mexican sequences was determined within the whole genome (WG) and individual ORF1a, ORF1b and S trees (SI Table 2), and was then confirmed on the global phylogeny available in Nextstrain<sup>24</sup>.

To visualize details on the phylogenetic relatedness of the Mexican sequences, we then subsampled the previous large-scale WG alignment in a phylogenetically-informed manner, as based on the position of the Mexican isolates within the large-scale trees and by selecting sequences using pairwise genetic distances<sup>25,26</sup>. Briefly, all Mexican sequences were retained together with their immediate ancestors and descendants, and 184 sequences were further selected based on the minor pairwise genetic distance in relation to the Mexican genomes under a threshold of 99.5% (SI Table 3). The subsampled WG alignment was scanned for recombinant sequences using the GARD algorithm<sup>27</sup> in the Datamonkey server<sup>28,29</sup>. A total of 201 sequences, including the Wuhan-Hu-1 reference genome, were used to re-estimate subsampled trees, as described above.

Global phylogenetic analysis and shared nucleotide substitution patterns<sup>23</sup> were used to confirm the position of the Mexican sequences based on the consensus clustering patterns observed within the large-scale WG tree, the subsampled WG tree, and the individual large-scale ORF trees, using a bootstrap value >50 for branch support, when possible (SI Table 2). In general, we observed consistency within the global tree and our large-scale and subsampled trees, as depicted by a conserved general structure at an internal branch level. Finally, analysis of the phylogenetic relationship between the Mexican and other viral sequences to identify groups of introduction events (IE) and local transmissions (LT) was done based on the following local definition, in which and IE or LT must include: *i*) one or more Mexican sequences, *ii*) a minimum of one closest related sister sequence(s), *iii*) and/or the immediate common ancestor<sup>30,31</sup>.

#### *Mutation identification*

Snippy<sup>32</sup> was used to identify all mutations unique to the Mexican genomes, as compared to the reference genome sequence of isolate Wuhan-Hu-1. The large-scale WG alignment in nucleotides (including UTR

and intergenic regions) was used as input for SI Table 4, while the large-scale WG concatenated ORF alignment was used as input for SI Table 5. The frequency and distribution for nucleotide and amino acid changes were determined using a normalized Sequence Logo, implemented by JalView<sup>33</sup>. Non-conservative amino acid changes were determined based on amino acid properties. Finally, the list of amino acid changes obtained was compared to the list of known sites scored to be evolving under pervasive, episodic or directional positive selection, as tested under several *dN/dS* models that have been fine tuned for SARS-CoV-2 datasets, in which an inflated *dN/dS* due to the occurrence of intra-species / intra- host polymorphism that may not be attributable by positive selections can be mitigated by restricting the site-specific analyses to internal branches <sup>34,35</sup>.

## RESULTS AND DISCUSSION

### Multiple introduction events of SARS-CoV-2 variants from two different lineages

A total of 17 full viral genome sequences were obtained from selected Mexican samples representing the earliest sampled cases detected in the country (Figure 1). From the epidemiological data associated with the Mexican samples, 15 of the cases corresponded to introduction events from travelers returning from abroad that entered the country through Mexico City airport, with 5 of them then relocating to other places within the country using local transportation (either aerial or terrestrial). Two additional cases reported no travel history (Table 1). Global phylogenetic analysis<sup>23</sup> confirmed that 8 Mexican isolates (samples 8, 17, 19, 24, 27, 28, 30, 31) grouped within the SARS-CoV-2 lineage B (also called lineage S, composed by sequences predominantly from the Americas). The remaining 9 Mexican isolates (samples 2, 5, 6, 7, 13, 16, 22, 32, and 33) grouped within lineage A (also called lineage G, includes sublineages A2 and A2a, and composed by sequences predominantly from Europe) (Figure 2)<sup>36</sup>. Lineage A has been defined as those viruses that share two nucleotide substitutions (positions 8782 in ORF1ab and 28144 in ORF8) that are closest to the root of the tree, and that are most similar to reference sequence Wuhan/WH04/2020 (EPI\_ISL\_406801). Lineage B viruses comprises viruses that share nucleotide substitutions C18060T and are most similar to reference sequence Wuhan-Hu-1 as an early representative (REF 36). The letters G (for A) and S (for B) are equivalents assigned in the GSAID/Nexstrain genomic epidemiological report (REF23). Both the A2/G and B/S lineages are contemporary, as the estimated date of emergence for lineage B/S is 12/29/2019 (95% CI: 12/20/2019, 01/03/2020), while for lineage A2/G is 01/18/2020 (95%

CI 01/02/2020, 01/19/2020)<sup>23</sup>. The collection dates for the Mexican samples that fall within lineage B1 range from 03/04/2020 to 03/15/2020, while those that fall within lineage A2 range from 02/27/2020 to 03/15/2020, including the isolate corresponding to the first reported case in México<sup>37</sup> REF (sample 33). This suggests an initial co-circulation of both the A2/G and B/S lineages in Mexico (Figure 2). Viruses belonging to the third lineage reported, V, were not identified in this study.

Consensus clustering patterns observed within the large-scale and subsampled trees (SI Table 2), showed eight well-supported independent introduction events (IE1, IE3, IE5, IE7, IE9, IE10, IE12 and IE13; Figure 2). For IE2, IE4, IE6 and IE8, we were unable to determine their origins and the immediate phylogenetic relatedness of the Mexican sequences, due to low support values and inconsistent clustering patterns across trees (SI Table 2). The current resolution of phylogenomic analyses is limited by the low diversity of the SARS-CoV-2 virus. Thus, for the characterized Mexican viral isolates, we could only determine with confidence the geographical origins at a regional level, rather than at country-level resolution. Altogether, these observations suggest that the virus variants identified in this study were more closely related to viral sequences circulating at the time in the USA and Europe rather than in China or South East Asia.

### **Evidence for early local transmission**

Sequences 27 and 31 correspond to two individuals that shared travel history to Vail, CO, USA, and that were in direct contact with case 28 in the return flight to Mexico. Nonetheless, the distribution of sequence 28 in an independent group (IE12) in relation to sequences 27 and 31 (Figure 2, SI Table 2), suggests that there were at least two different viral variants co-circulating within this specific location in the USA. On the other hand, sequences 8, 27, 30 and 31 grouped together, representing a local transmission cluster (LT11) (Figure 2). This observation is supported by phylogenetic consistency within our local trees and the global tree<sup>23</sup>, and a high support value (bs: 100) observed for LT11 in all cases (S1 Table 2, Figure 2). Case 8 corresponds to an individual with no epidemiological relationship or contact with cases 27, 30 and 31, and with no travel history. Case 30 had travel history to Europe, but no direct exposure with cases 27 and 31. For case 30, the possibility of this person acquiring the virus whilst being abroad cannot be ruled out. However, our analysis shows evidence that this person may have contracted the virus whilst already being in Mexico. Thus, LT11 strongly supports the occurrence of at least one independent local transmission event in Mexico City (Case 8), occurring as early as the second week of March 2020.



## Genetic variation within the Mexican viral genomes

When compared to the Wuhan-Hu-1 reference genome, the Mexican sequences displayed between 4 and 10 nucleotide substitutions, and between 1 and 5 amino acid changes. This is consistent with the reported rate of evolution of  $\sim 8 \times 10^{-4}$  nucleotide substitutions per site per year, equivalent to  $\sim 2$  substitutions per month<sup>38,39</sup>. Collectively, 46 nucleotide substitutions and 20 amino acid residues were identified within the Mexican viral genomes (SI Table 4 and 5). As expected, the majority of these variants were not conserved through the genomes, and only 15 nucleotide changes and 6 amino acid changes were shared by more than one sequence. These results exclude sequences 6, 13 and 16, which showed a considerably lower coverage and depth when compared to the remaining 14 high-quality viral genomes obtained. Thus, most of the variability observed could be explained by errors introduced during reverse transcription, PCR amplification, sequencing or assembly (SI Table 1 and 4).

Consistent with our phylogenomic analysis, all Mexican sequences belonging to lineage A2 showed two lineage-specific nucleotide substitutions, C241T and A23403G. A23403G results in the D614G amino acid change in the Spike protein (Table 2). All lineage A2a sequences (including the Mexican isolates) had the additional nucleotide substitution C14408T, resulting in the amino acid change P314L in Orf1b<sup>23</sup>. Sequences within lineage B had the nucleotide substitution T28144C rendering the L84S amino acid change in the Orf8 protein. Similarly, lineage B1 sequences also showed the nucleotide substitution C18060T (Table 2). No evidence for recombination was found within any of the alignments, in agreement with previous observations<sup>28</sup>. Taken together, our results suggest the Mexican viral sequences display the genetic changes according to their phylogenetic placement.

According to the natural selection analyses for SARS-CoV-2 enabled by data from GSAID (REF 34,35), most of the variable sites detected in our analyses are likely to be evolving under negative or neutral evolution, as expected for RNA viruses (REF). However, two of these sites (614 in the Spike protein and site 84 in Orf8) have been predicted to be evolving under positive selection (Table 2, Supplementary Table 5). Site 614 in Spike has been scored as evolving under pervasive positive selection and belongs to predicted CTL linear epitope that may be recognized by one or more HLA alleles (REFS 34,35, Nueva). Mutation D/G at this site has been speculated to be involved in an increased spread of the virus (REF NUEVA). Similarly, site 84 in Orf8 is also evolving under pervasive positive selection, may show intra-host variation, and again, belongs to predicted CTL linear epitope (REF 34,35, Nueva). Nonetheless,

observations on the functional properties of these mutations are still debatable, and data available up to date is yet inconclusive for any changes in biological and/or evolutionary properties of the virus.

#### **H49Y amino acid change in Mexican sequences within LT11**

We further identified within all Mexican sequences grouping with the local transmission cluster (LT11), the C21707T nucleotide substitution (following the whole genome and nucleotide alignment numbering), that corresponds to a H to Y amino acid change in position 49 of the Spike protein. Mapping this nucleotide substitution onto the branches of the global virus phylogeny<sup>23</sup> (with dates ranging from December 2019 to April 2020), revealed that C21707T has occurred with a frequency of 0.4% (20/4533) within the viral genomes available as of May 6<sup>th</sup> 2020. It first occurred within a single cluster of 14 sequences from China (representative sequence: Jiangsu/JS02/2020 EPI ISL 411952). The estimated date of emergence for this cluster that eventually died off (stopped circulating and had no more descendants) was of 01/12/2020 (95% CI:01/08/2020-01/16/2020). This mutation emerged again in all the sequences within the LT11 from Mexico (Table 1, Figure 2). Since then, C21707T has also appeared independently as a singleton in different isolated circulating worldwide. ~~three viruses from Australia (EPI\_ISL\_426898, EPI\_ISL\_430632 and EPI\_ISL\_430633), two viruses from the UK (EPI\_ISL\_433757 and EPI\_ISL\_432818), one virus from Taiwan (EPI\_ISL\_417518) and two viruses from the USA (EPI\_ISL\_408010 and EPI\_ISL\_430050), with collection dates ranging from 01/29/2020 to 04/05/2020 (sequences deposited on GISAID platform by May 10<sup>th</sup> 2020).~~

All of the sequences showing this nucleotide substitution within the global phylogeny (REF23) belong to different viral lineages (either, A2, B1 or others), confirming that there is no phylogenetic correlation between them (e.g. founder effect), and rather supporting for the independent occurrence of this change as a homoplasy. Despite C21707T appearing several times as a singleton (represented by tips on the tree), it has only been fixed in two lineages comprising independent viral subpopulations (occurring on internal branches or nodes of the tree): China and Mexico. Both global and local phylogenetic analyses show that the closest viral sequences from abroad related to LT11 do not have this nucleotide substitution. Thus, at least for LT11, this change most likely originated when this viral variant was initially introduced in the country. We do not know if any variants derived from LT11 continued circulating in the country, or if this cluster eventually died off. Therefore, it would be interesting to use genomic surveillance to follow up on

the frequency of occurrence of C21707T in viruses currently circulating in Mexico, either to determine if these are sequencing errors<sup>41</sup>, or to know if this is a real, low-frequency homoplasy.

The H49Y mutation resulting from the C21707T nucleotide substitution represents a non-conservative amino acid change located within the N-terminal domain (NTD) of the S glycoprotein trimer, a protein region that has not been fully studied so far<sup>40</sup>. No evidence of episodic or directional positive selection was found for this site, as tested under local and global analyses to estimate  $dN/dS$ <sup>34,35</sup>. It would be relevant to explore if this change is associated with any biological properties or the virus, as has been shown for other virus populations<sup>25</sup>. However, further structural biology analysis and experimental data would be needed to determine if this site has any functional impact, or to determine any implications of it in local transmission dynamics.

### **Data Availability**

The generated sequences SARS-CoV-2 from Mexico can be found in GISAID and in Nextstrain<sup>24</sup>. The corresponding GISAID accession numbers are listed in Table 1.

### **ACKNOWLEDGEMENTS**

This work was partially supported by grant “Epidemiología genómica de los virus SARS-CoV-2 circulantes en México” from the National Council for Science and Technology (CONACyT)-Mexico to CFA, and by grants from the Mexican Government (Comisión de Equidad y Género de las Legislaturas LX-LXI y Comisión de Igualdad de Género de la Legislatura LXII de la H. Cámara de Diputados de la República Mexicana) to SAR and CB. MEZ is supported by a Leverhulme Trust ECR Fellowship (ECF-2019-542). We thank the “Unidad de Secuenciación Masiva y Bioinformática” of the “Laboratorio Nacional de Apoyo Tecnológico a las Ciencias Genómicas” (CONACyT #260481) for their support in sequencing services. We thank all the staff of the Technological Development and Molecular Research Unit, Virology Department, and Sample Control and Services Department at InDRE for their technical assistance. The findings and conclusions in this report are only the responsibility of the authors and do not necessarily of the institutions involved.

## FIGURE LEGENDS

### **Figure 1. Epidemiological positioning of the SARS-CoV-2 samples from Mexico**

Epidemiological curve for the early SARS-CoV-2 epidemic in Mexico, dating from late February until early April. The rise in cumulative cases is shown in red, whilst the daily incidence is shown with orange bars. Dates of collection for the samples used in this study are indicated with the black arrows. One arrow might indicate the collection of more than two samples (dates shown in Table 1).

### **Figure 2. Phylogenomic positioning of the SARS-CoV-2 isolates from Mexico.**

RAxML tree estimated from the reduced whole genome alignment built using the concatenated main viral ORFs, and rooted using the Wuhan-Hu-01 isolate. Sequences corresponding to the viral isolates from Mexico sequenced in this work are shown in red, while the region of origin for the identified closest related immediate ancestors and/or sister isolates is indicated in black. Support values for the branches of interest are indicated with numbers next to the branches. The clusters identified in this work representing introduction events (IE1-IE10 and IE12 and IE13) and the local transmission (LT11) are shown next to the isolate names.

## REFERENCES

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
2. Pinotti, F. *et al.* Lessons learnt from 288 COVID-19 international cases: importations over time, effect of interventions, underdetection of imported cases. *medRxiv* 2020.02.24.20027326 (2020).
3. auspice. <https://nextstrain.org/narratives/ncov/sit-rep/2020-01-23>.
4. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
5. Liu, J. *et al.* Community Transmission of Severe Acute Respiratory Syndrome Coronavirus 2, Shenzhen, China, 2020. *Emerg. Infect. Dis.* **26**, (2020).
6. [No title].  
[https://www.gob.mx/cms/uploads/attachment/file/547271/Comunicado\\_Tecnico\\_Diario\\_COVID-19\\_2020.04.18.pdf](https://www.gob.mx/cms/uploads/attachment/file/547271/Comunicado_Tecnico_Diario_COVID-19_2020.04.18.pdf).
7. de Salud, S. Sistema Nacional de Vigilancia Epidemiológica. *gob.mx*  
<http://www.gob.mx/salud/acciones-y-programas/sistema-nacional-de-vigilancia-epidemiologica>.
8. [http://www.censida.salud.gob.mx/descargas/biblioteca/documentos/manual\\_para\\_la\\_toma\\_envio\\_y\\_recepcion\\_de\\_muestras.pdf](http://www.censida.salud.gob.mx/descargas/biblioteca/documentos/manual_para_la_toma_envio_y_recepcion_de_muestras.pdf).
9. [https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf?sfvrsn=a9ef618c\\_2](https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf?sfvrsn=a9ef618c_2).
10. Taboada, B. *et al.* Is There Still Room for Novel Viral Pathogens in Pediatric Respiratory Tract Infections? *PLoS One* **9**, e113570 (2014).
11. Taboada, B. *et al.* The Geographic Structure of Viruses in the Cuatro Ciénegas Basin, a Unique Oasis in Northern Mexico, Reveals a Highly Diverse Population on a Small Geographic Scale. *Appl. Environ. Microbiol.* **84**, (2018).
12. Nextera XT DNA Sample Prep Kit Documentation.  
[https://support.illumina.com/sequencing/sequencing\\_kits/nextera\\_xt\\_dna\\_kit/documentation.html](https://support.illumina.com/sequencing/sequencing_kits/nextera_xt_dna_kit/documentation.html).
13. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.  
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
14. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
15. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation

- sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
16. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
  17. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
  18. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1–19 (2004).
  19. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
  20. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
  21. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
  22. Pond, S. Genomic diversity and divergence of SARS-CoV-2/COVID-19. *Observable* <https://observablehq.com/@spond/current-state-of-sars-cov-2-evolution> (2020).
  23. <https://nextstrain.org/ncov/global?dmax=2020-04-08>.
  24. <https://nextstrain.org/ncov/global>.
  25. Escalera-Zamudio, M. *et al.* Parallel Evolution in the Emergence of Highly Pathogenic Avian Influenza A Viruses. *bioRxiv* 370015 (2019) doi:10.1101/370015.
  26. Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS* **31**, 1211–1222 (2017).
  27. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006).
  28. Robertson, D. I. *et al.* nCoV's relationship to bat coronaviruses & recombination signals (no snakes) - no evidence the 2019-nCoV lineage is recombinant. *Virological* <http://virological.org/t/ncovs-relationship-to-bat-coronaviruses-recombination-signals-no-snakes-no-evidence-the-2019-ncov-lineage-is-recombinant/331> (2020).
  29. Datamonkey Adaptive Evolution Server. <http://datamonkey.org/>.
  30. Shiino, T. *et al.* Molecular Evolutionary Analysis of the Influenza A(H1N1)pdm, May–September, 2009: Temporal and Spatial Spreading Profile of the Viruses in Japan. *PLoS One* **5**, e11057 (2010).

31. Baillie, G. J. *et al.* Evolutionary Dynamics of Local Pandemic H1N1/2009 Influenza Virus Lineages Revealed by Whole-Genome Analysis. *J. Virol.* **86**, 11–18 (2012).
32. Seemann, T. tseemann/snippy. *GitHub* <https://github.com/tseemann/snippy>.
33. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
34. Pond, S. Natural selection analysis of SARS-CoV-2/COVID-19. *Observable* <https://observablehq.com/@spond/natural-selection-analysis-of-sars-cov-2-covid-19> (2020).
35. Dashboard | SARS-CoV-2 Admin. <http://covid19.datamonkey.org/2020/04/01/covid19-analysis>.
36. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* 2020.04.17.046086 (2020) doi:10.1101/2020.04.17.046086.
37. Garcés-Ayala, F. *et al.* Full Genome Sequence of the first SARS-CoV-2 detected in Mexico. *Arch. Virol.* (2020). *In press*.
38. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv* 2020.04.02.20051417 (2020).
39. Rambaut, A. *et al.* Phylodynamic Analysis | 176 genomes | 6 Mar 2020. *Virological* <http://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356> (2020).
40. Gui, M. *et al.* Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res.* **27**, 119–129 (2016).
41. <http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473>.