AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

| Epidemiology | Research Article

Genomic analysis reveals dynamics of SARS-CoV-2 during the initial phases of the COVID-19 outbreak in the Dominican Republic

Allie Kreitman,[1] Alexandra Mushegian,[1] Martha I. Nelson,[2] Stephanie Banakis,[1] Christopher Mederos,[1] Matthew Chung,[1] Allison Roder,[1] Armando Peguero,[3] Sayira Mueses,[3] Paula Cuevas,[4] Robert Paulino-Ramírez,[3] Elodie Ghedin[1]

AUTHOR AFFILIATIONS See affiliation list on p. 12.

ABSTRACT Genomic sequencing of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been widely instituted during the coronavirus disease 2019 (COVID-19) pandemic to track the emergence and spread of new lineages. While this powerful tool can influence public health decisions and therapeutic development, not all regions of the world have had equal access to sequencing capacity, affecting surveillance. One such underrepresented region is the Caribbean islands, including the Dominican Republic (DR). To determine retrospectively what lineages were circulating in the DR in the first year of the pandemic, when there were strict travel restrictions imposed, we sequenced SARS-CoV-2 from nasal swab and saliva samples collected between July 2020 and February 2021. We investigated whether COVID-19 outbreaks were seeded by single or multiple introductions and established epidemiological linkages to other countries. Using 98 newly sequenced samples, we identified 16 SARS-CoV-2 lineages in the DR, indicating many independent introductions from diverse geographic areas. Further, we show that analyzing both globally prevalent and globally rare lineages within the DR highlights different aspects of international disease transmission.

IMPORTANCE This study uses genome sequencing of severe acute respiratory syndrome coronavirus 2 samples collected in an undersampled region of the world—the Caribbean, specifically the Dominican Republic—to make novel inferences about the dynamics of disease spread during a period of the coronavirus disease 2019 pandemic when many diverse lineages were co-circulating globally.

KEYWORDS COVID-19 pandemic, genomic epidemiology, SARS-CoV-2 genomics, Dominican Republic

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in late 2019 from Wuhan, China, and is the etiological agent of coronavirus disease 2019 (COVID-19). The virus quickly spread globally and was declared a pandemic by the World Health Organization in March 2020. Viral whole genome surveillance sequencing and sharing of genomic data were instituted internationally on a scale never seen in previous disease outbreaks (1) to track the dynamics of SARS-CoV-2 evolution and transmission. This effort helped in the timely identification of new emerging SARS-CoV-2 lineages (2–5), in adapting infection control intervention, in mitigating effects on public health infrastructure, and in tracking the spread of viral lineages across the world (6–8). However, the availability of surveillance sequencing has been highly heterogeneous (1), with thousands of SARS-CoV-2 samples sequenced per week in some countries (9), and only tens of samples per week in others (10). Given the power of sequencing to detect the circulation of novel viral lineages and inform the development of diagnostic

Editor My V. T. Phan, Arizona State University College of Health Solutions, Phoenix, Arizona, USA

Address correspondence to Elodie Ghedin, elodie.ghedin@nih.gov.

Allie Kreitman and Alexandra Mushegian contributed equally to this article. Author order was determined in order of increasing seniority.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 10 April 2025
Accepted 15 December 2025
Published 23 January 2026

tools, therapeutics, and non-pharmaceutical interventions (11), increasing the breadth of viral surveillance sequencing across the world should be a high priority as sequencing capacity develops.

One such under-surveilled region was the Caribbean islands, including La Hispaniola, a shared island between the Dominican Republic (DR) and Haiti (12). The DR is the third most populous country in the region, with more than 11 million citizens, and a popular tourism destination. The first case of COVID-19 in the DR was identified on 1 March 2020 from an individual who had traveled to Italy (13). Of the over 200,000 cases reported in the DR in the first 1.5 years of the pandemic, from the beginning of 2020 through June 2021, only 289 SARS-CoV-2 whole genome sequences were publicly available on Global Initiative on Sharing All Influenza Data (GISAID) in that same timeframe. The limited number of sequences is particularly noteworthy, especially in contexts where strict travel restrictions were imposed during the early stages of the pandemic.

To retrospectively understand how tourism and regional factors drove the dynamics of the early COVID-19 pandemic in the DR and provide a better understanding of how human behaviors and travel impacted the Caribbean region, we performed whole genome sequencing of samples collected between 29 July 2020 and 25 February 2021. We used phylogenetic analysis and international travel data to report what SARS-CoV-2 lineages were circulating, whether SARS-CoV-2 outbreaks in the DR were seeded by a single or multiple introductions, and from where those introductions originated.

MATERIALS AND METHODS

Sample collection

Saliva and nasopharyngeal swab samples were obtained from patients with respiratory symptoms and their contacts. Samples were collected at different primary care and hospital facilities. After PCR testing, all results were returned for clinical decisions, and an anonymized aliquot was frozen for sequencing analysis. RNA extraction was performed following the protocol provided by the IVD RADI PREP swab and stool DNA/RNA extraction kit (KH Medical Co., Republic of Korea). The sample was confirmed to be SARS-CoV-2 positive (cycle threshold [Ct] value of 26) by real-time quantitative PCR (RT-qPCR) using the virellaSARS-CoV-2 seqc amplification protocol (gerbion GmbH & Co., Germany).

SARS-CoV-2 whole genome sequencing

Amplification of viral genomes, library construction, and genomic analysis were done according to the protocols available at https://github.com/GhedinSGS/SARS-CoV-2_analysis. Briefly, SARS-CoV-2 genomes were amplified using Qiagen OneStep Ahead RT-PCR kit (cat 220211). Libraries were constructed using Illumina Nextera XT (cat FC-131-1096) at 0.25× the volume described in the protocol. Libraries were pooled equimolarly and sequenced on the Illumina NextSeq500 and NextSeq2000 using the 2 × 150 bp paired-end protocol. Adapters and primers were trimmed, reads were aligned to the Wuhan/Hu-1 strain (NC_045512.2), and the two libraries for each sample were merged. Consensus sequences were assembled using the Timo and GATK pipelines, with a minimum of 5× read coverage required to call nucleotides at each site.

Summarization of publicly available SARS-CoV-2 data from GISAID

Table S1 includes all SARS-CoV-2 sequences from GISAID that were collected in the DR from the beginning of the pandemic through June 2021. These sequences were collected from humans and classified by GISAID as complete, high coverage, and collection date complete.

Lineage calling

SARS-CoV-2 genome sequences with coverage of at least 20,000 nucleotides was input into pangolin v. 3.1.20 (14, 15) for lineage calling and Nextclade (16) v. 2.12.0 for clade classification.

Phylogenetics analysis

A maximum likelihood phylogenetic tree was constructed using the Nextstrain CLI (17) v. 3.0.6 ncov workflow v12, which uses nextalign for sequence alignment and IQTree v. 2.2.0 with the GTR substitution model (18) for tree building. Trees were plotted in R using the packages ggplot (19) and ggtrees (20).

Global phylogenetic tree

All DR samples sequenced for this project with 25,000 nucleotide coverage or higher were included in a phylogenetic tree on a diverse global background. The background includes a random sample of SARS-CoV-2 sequences publicly available on GISAID that were collected between 1 July 2020 and 31 March 2021. The background samples were chosen by randomly selecting 500 samples from North America, South America, Asia, Africa, Europe, Oceania, and Central America. Up to 100 samples were randomly selected from each Caribbean Island, except for the DR where all available samples were included. The distribution of samples per country is in Table S2. All background samples fulfilled the GISAID criteria of being high quality, complete, included completed collection dates, and were collected from humans (Table S3). Background data set metadata and FASTA sequences were downloaded from GISAID on 28 March 2023. A maximum likelihood phylogenetic tree was made as described above.

B.1.575 phylogenetic tree

A total of 3,681 B.1.575 sequences was downloaded from GISAID on 4 November 2024 to construct a background data set. The majority of sequences (89.0%) were from the United States (USA), and 62.7% were collected in the eastern USA region, primarily New York. The SAMPI subsampling tool, available at https://github.com/jlcherry/SAMPI, was used to randomly subsample the data set over space and time, including no more than 50 sequences from the following locations: Asia, Europe, Oceania, South America, Central America, Caribbean (not including DR), USA (East region), USA (Midwest), USA (South), USA (West), and DR (not including this study). Fewer than 50 sequences were available for South America (n = 42), the Caribbean (n = 35), and Central America (n = 9). The final background data set included 469 publicly available global sequences from GISAID, sampled over a 13-month period from 23 October 2020 (hCoV-19/USA/TX-HMH-MCoV-15455/2020) to 17 November 2021 (hCoV-19/USA/NY-NYULH3718/2021). The final data set included these 469 GISAID sequences plus 26 samples sequenced for this study (Table S4).

To infer the transmission dynamics of the B.1.575 lineage, an ancestral state reconstruction was performed using the Markov Chain Monte Carlo (MCMC) methods available in the Bayesian Evolutionary Analysis Sampling Trees (BEAST) v.1.10.4 package (21). Sequences were aligned using Nextalign CLI (16) v2.12.0, then untranslated regions and intergenic regions were trimmed manually. A relaxed uncorrelated lognormal (UCLN) clock was used, with a Hasegawa–Kishino–Yano (HKY) model of nucleotide substitution with gamma-distributed rate variation among sites. An exponential population growth model was used during this early stage of SARS-CoV-2 epidemic growth in an immunologically naïve population. The MCMC chain was run separately four times using the BEAGLE 3 (22) library to improve computational performance, until all parameters reached convergence, as assessed visually using Tracer v.1.7.2. At least 10% of the chain was removed as burn-in, and runs for the same data set were combined using LogCombiner v1.10.4. A MCC tree was summarized using TreeAnnotator v.1.10.4 and visualized using the ggtrees (20) package in R.

A lineage phylogenetic tree

All publicly available viral sequences from human samples from the A lineage with GISAID's high coverage designation, including its sub-lineages, were downloaded from GISAID on 21 February 2023 (Table S5). A maximum likelihood phylogenetic tree was made as described above, but relaxing the coverage criteria to a minimum of 21,000 nucleotides in the Nextstrain build parameters.

COVID-19 case data

The number of COVID-19 cases was obtained from Our World In Data (https://ourworldindata.org/) on 8 November 2022.

COVID-19 non-pharmaceutical interventions data

Date ranges for non-pharmaceutical interventions in the DR, including international border closures, restrictions on internal movement, school closures, stay-at-home orders, restrictions on large gatherings, and masking requirements were obtained from Our World in Data (https://ourworldindata.org/) on 8 November 2022. Total border closures and bans on high-risk regions were included for classification as having a closed border. Recommended and required movement restrictions were both considered as part of "movement within country restricted." "School closure" includes when schools were required to close at some or all grade levels, and "stay-at-home" includes when stay at home was recommended or required. No large gatherings classification includes any gathering over 100 people required or recommended to be canceled. "Masking requirements" include when masking was required whenever outside the home or when required in public places.

Travel data

To investigate the international spread of SARS-CoV-2, we obtained international flight passenger data from OAG (https://www.oag.com). OAG collects passenger bookings data, which accounts for true origin and destination of flight. We obtained the monthly number of passengers traveling into the Dominican Republic by air from all countries. The air passenger data used in this study are proprietary and were purchased from OAG Aviation Worldwide Ltd. These data were used under a license for the current study and so are not publicly available.

RESULTS

Descriptive statistics of the data set

The Caribbean region had a relatively low case burden, potentially due to its island nation characteristics, airport closures, and to the early use of non-pharmaceutical interventions, such as mask mandates, school closures, and limits to large gatherings. On the other hand, case detection could have been low due to limited availability of laboratory infrastructure, technical personnel, reagents due to border closures, and testing availability. The DR accounted for 43% of confirmed cases in the Caribbean region in the first 16 months after the first COVID-19 case was reported (23) (Fig. 1A), although they make up 25% of the population, and 289 complete SARS-CoV-2 genomes collected through June 2021 are publicly available from the DR on the GISAID database as of 7 May 2024 (Table S1). To better understand the contribution of the DR to regional and international spread of the virus, we sequenced SARS-CoV-2 genomes from nasopharyngeal swab and saliva samples collected between 29 July 2020 and 25 February 2021. Of these, 98 samples yielded high enough quality sequence data to be classified by Pango (14) and be included in further analyses. The primary difference between successfully and unsuccessfully sequenced samples was in their viral titer (Ct value) ($P < 0.001$). For the individuals for which samples had sufficient sequencing coverage, the mean age was 38 years old, and 47% of sampled individuals identified as

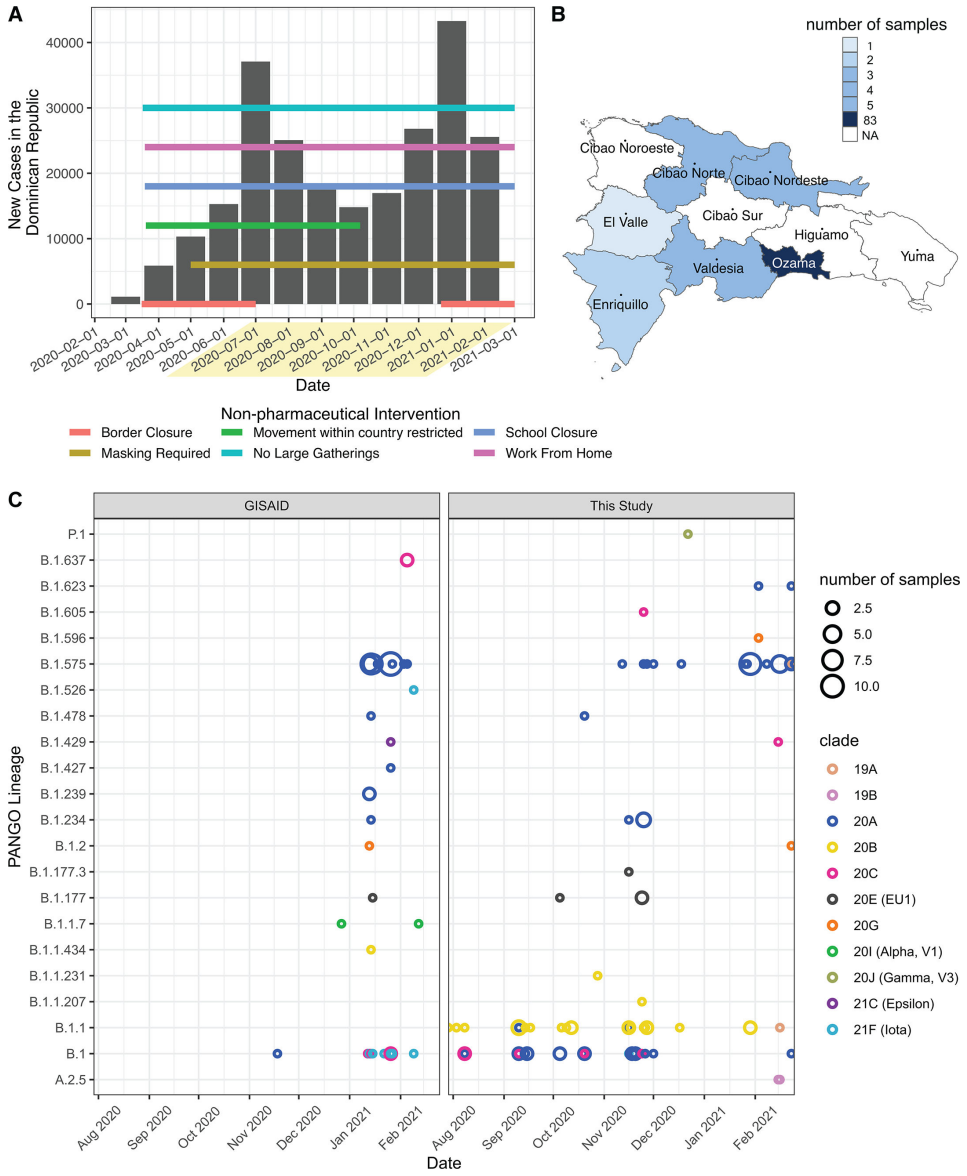

FIG 1 COVID-19 precautions and sample collection in the DR. (A) Plot of new COVID-19 cases in the DR with timeline of government restrictions in the DR. Orange indicates international border closures; dark yellow indicates masking requirements in public spaces; dark green indicates restriction on movement within the DR; turquoise indicates restrictions on gatherings of 100 people or greater; blue indicates public school closures; and pink indicates non-essential workers required or recommended to work from home and study dates are highlighted with yellow box (July 2020 to February 2021). (B) Map of the regions of the DR. Darker shades of blue indicate greater number of samples collected from that region for this study. (C) Comparison of SARS-CoV-2 lineages collected from the DR and made available on GISAID vs samples sequenced during the sampling period of this study. Color represents nextclade lineage, while Y axis lists PANGO lineages. Size indicates the number of genome sequences of that lineage on that date.

male (Table 1). These samples were collected across seven regions of the DR, but the majority (84%) were from Ozama (Fig. 1B), which includes Santo Domingo, the largest city and capital. Sampled patients were both symptomatic (n = 34) and asymptomatic (n = 64), but there was no statistically significant difference between age, sex, locality, or Pango lineage between these two groups. Most sequences were classified as Pango lineages B.1 (27%), B.1.1 (25%), and B.1.575 (26.5%) (Table 1). These samples also spanned a large diversity of Nextstrain clades (16), a more general lineage classification system. This data set includes many of the lineages previously identified to be circulating in the

TABLE 1 Summary statistics of the study participants whose samples had high enough coverage for lineage calling[a]

Characteristic	Asymptomatic (*n* = 64)	Symptomatic (*n* = 34)
Age, mean; range (SD)	38.2; 2–81 (15.5)	38.4; 20–89 (14.6)
Sex, *n* (%)		
Female	35 (35.7)	17 (17.4)
Male	29 (29.6)	17 (17.4)
Patient location, *n* (%)		
Cibao Nordeste	4 (4.1)	0
Cibao Norte	2 (2.04)	1 (1.02)
Enriquillo	2 (2.04)	0
Ozama	51 (52.04)	32 (32.7)
Valdesia	5 (5.1)	0
El Valle	0	1 (1.02)
PANGO lineage, *n* (%)		
B.1	20 (20.4)	7 (7.1)
B.1.1	17 (17.35)	8 (8.2)
B.1.1.231	1 (1.02)	0
B.1.177.3	1 (1.02)	0
B.1.234	4 (4.1)	0
B.1.429	1 (1.02)	0
B.1.478	1 (1.02)	0
B.1.575	16 (16.3)	10 (10.2)
B.1.596	1 (1.02)	0
B.1.623	2 (2.04)	0
P.1	0	1 (1.02)
A.2.5	0	2 (2.04)
B.1.1.207	0	1 (1.02)
B.1.177	0	3 (3.1)
B.1.2	0	1 (1.02)
B.1.605	0	1 (1.02)

[a]Age, sex, and symptomology were reported by the participant at time of sample collection; region represents the region in which the sample was collected, and lineage was determined by pangolin.

DR at that time and provides support for earlier transmission of lineages, such as B.1.575 and B.1.1, than previously reported (Fig. 1C).

SARS-CoV-2 lineages and diversity in a global context

To understand how the diversity of SARS-CoV-2 lineages in the DR compared with the global diversity of SARS-CoV-2 during the time of sample collection, we constructed a maximum likelihood phylogenetic tree using samples from all Caribbean islands, including those of our sequenced samples with sequence coverage over at least 25,000 nucleotides, and a random sample of global SARS-CoV-2 sequences (more detailed subsampling in Materials and Methods) that represent a large diversity of SARS-CoV-2 variants (Fig. 2A). The samples collected from the DR were spread across many clades of the global tree of SARS-CoV-2 samples (Fig. 2B). Uneven sequencing of SARS-CoV-2 globally has led to some regions and timeframes being underrepresented in the tree, and thus there are gaps in viral evolutionary paths. Because of these gaps, and due to limited genetic diversity in the early pandemic, there are several polytomies present, indicating that more data may be needed to differentiate these closely related lineages. Nevertheless, our results suggest many independent introductions of SARS-CoV-2 into the DR in the first year of the pandemic, despite some travel restrictions.

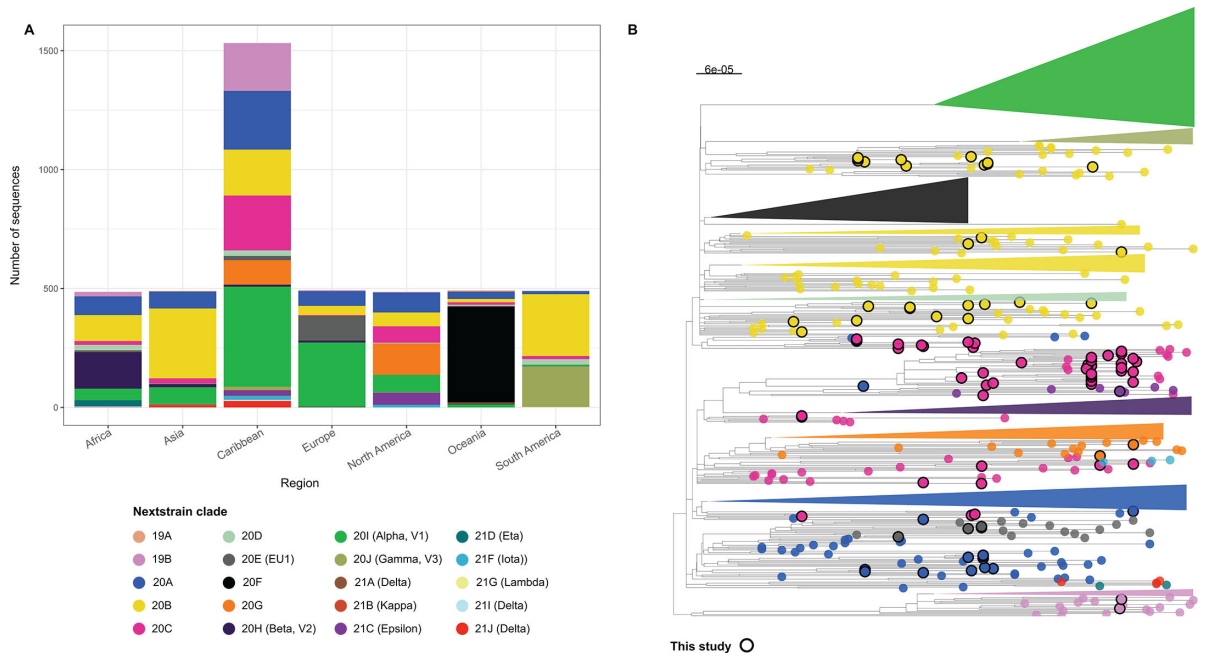

FIG 2 Nextclade lineage of SARS-CoV-2 samples by region of sample collection. (A) Distribution of publicly available SARS-CoV-2 whole genome sequences from GISAID colored by NextStrain clade (see Materials and Methods for more detailed sampling method) used for global phylogenetic tree. (B) Maximum likelihood phylogenetic tree of all high-coverage samples sequenced for this study on a randomly sampled global background. Background was then downsampled to 500 sequences. The color of the tree tip represents the NextStrain clade of the sample. Nodes circled in black represent samples collected and sequenced for this study. Branch lengths are measured in substitutions/nucleotide site/year. Newick files are included as supplemental data files for additional viewing of this tree before downsampling.

B.1.1 phylogenetics

First, we investigated the transmission of one of the most widespread lineages in the DR, the B.1.1 lineage. This lineage was highly represented in our data set and globally widespread with over 40,000 cases sequenced in the timeframe of this study across 131 countries (covSPECTRUM, https://cov-spectrum.org/). We used the phylogenetic tree of DR samples on a global and Caribbean-centered background to focus on the transmission dynamics of the B.1.1 lineage (corresponding to NextStrain clades 20A, 20B, and 20E; Fig. 2B). The samples from the DR sequenced as part of this study clustered as a mix of small clusters (Fig. 3A, C, and D) and singletons (Fig. 3B). The number of distinct clusters suggests several introductions of B.1.1 into the DR. The samples cluster with samples collected across large geographic regions, including from North and South America, Asia, Africa, and other Caribbean islands. While limited sampling and polytomies within the larger B.1.1 clade prevent confident determination of where introductions to the DR began, this evidence suggests that the B.1.1 lineage was introduced many times, likely from diverse geographic regions, and circulated within the DR broadly throughout late 2020 and early 2021. This also corresponds to when the DR reduced their border restrictions in the summer and fall of 2020 (Fig. 1A), increasing the opportunities for travelers to have spread SARS-CoV-2 lineages from other countries.

B.1.575 lineage phylogenetics

Internationally, the B.1.575 lineage (corresponding to the NextStrain clade 20C) only made up 0.13% of cases in the same sampling period (covSPECTRUM, https://cov-spectrum.org/), leading us to investigate the transmission dynamics of this lineage and whether it represented an outbreak originating from within the DR. We identified 26 cases of the B.1.575 lineage in our sample set from the DR, making up 26.5% of cases we sequenced. Globally, some of the earliest B.1.575 cases were detected in DR (e.g.,

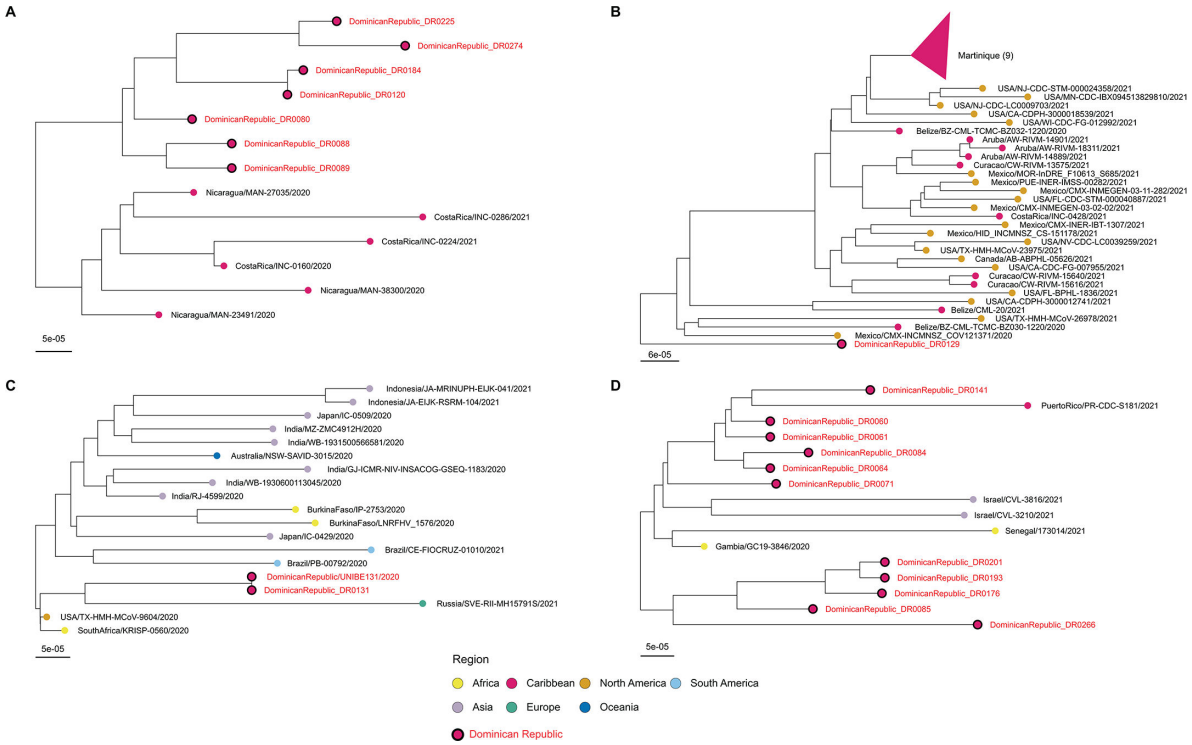

FIG 3 Multiple introductions of the B.1.1 lineage. Sub-trees of the B.1.1 lineage (20A, B, and E clades) from Fig. 2, but without the subsampling described in Fig. 2B. (A–D) Instances within the full tree (available as a newick file) where DR B.1.1 samples clustered together. Color represents the region the sample was collected; nodes with bold borders represent samples sequenced for this study (sample names in red), while nodes with lighter borders represent samples downloaded from GISAID. Branch lengths are measured in substitutions/nucleotide site/year. Newick files are included for additional viewing of the full tree.

DR0114, sampled 12 November 2020). We constructed a Bayesian timescale phylogenetic tree that included the 26 B.1.575 viruses sequenced in DR for this study, plus a random sample of 469 publicly available sequences from GISAID of the B.1.575 lineage (Table S4). DR B.1.575 sequences are dispersed through the B.1.575 tree, exhibiting high genetic diversity for such a small country (Fig. 4A). As a comparison, all 18 samples of the B.1.575 lineage from the Caribbean island of Bonaire form a single clade, suggesting a single point-source outbreak during February–March 2021, with no observable onward transmission to other locations (Fig. 4B). The high genetic diversity of B.1.575 in DR, represented by a large number of singletons and small clusters positioned in different regions of the tree (Fig. 4A), could arise from (i) multiple independent virus introductions into DR, primarily exported from the United States (Fig. 4C), (ii) B.1.575 originating in DR and subsequently spreading to the United States and other locations, and (iii) a combination of both, with transmission back and forth between DR and the United States. Many DR samples cluster with viruses from the United States, especially New York (Fig. 4B), which is to be expected as 89% (n = 3,275) of all B.1.575 sequences publicly available on GISAID were reported from the United States, and 29% (n = 977) of those from New York state, the most from any single state. (By prevalence as a percentage of total sequences, the top five U.S. states with B.1.575 samples were in descending order Pennsylvania, New Jersey, New York, New Hampshire, and Rhode Island.) Additionally, we observe several instances of samples from the DR closely clustering with samples from the Caribbean or Central America, indicating close genetic relatedness and highlighting potentially more complex transmission chains that include multiple Caribbean islands and Central American countries.

In analyzing the Markov jumps between regions, a measure of gene flow, we see that the highest rates of gene flow are between the USA and DR, and vice versa (Fig. 4C). There was not sufficient resolution in the phylogeographic analysis, particularly in

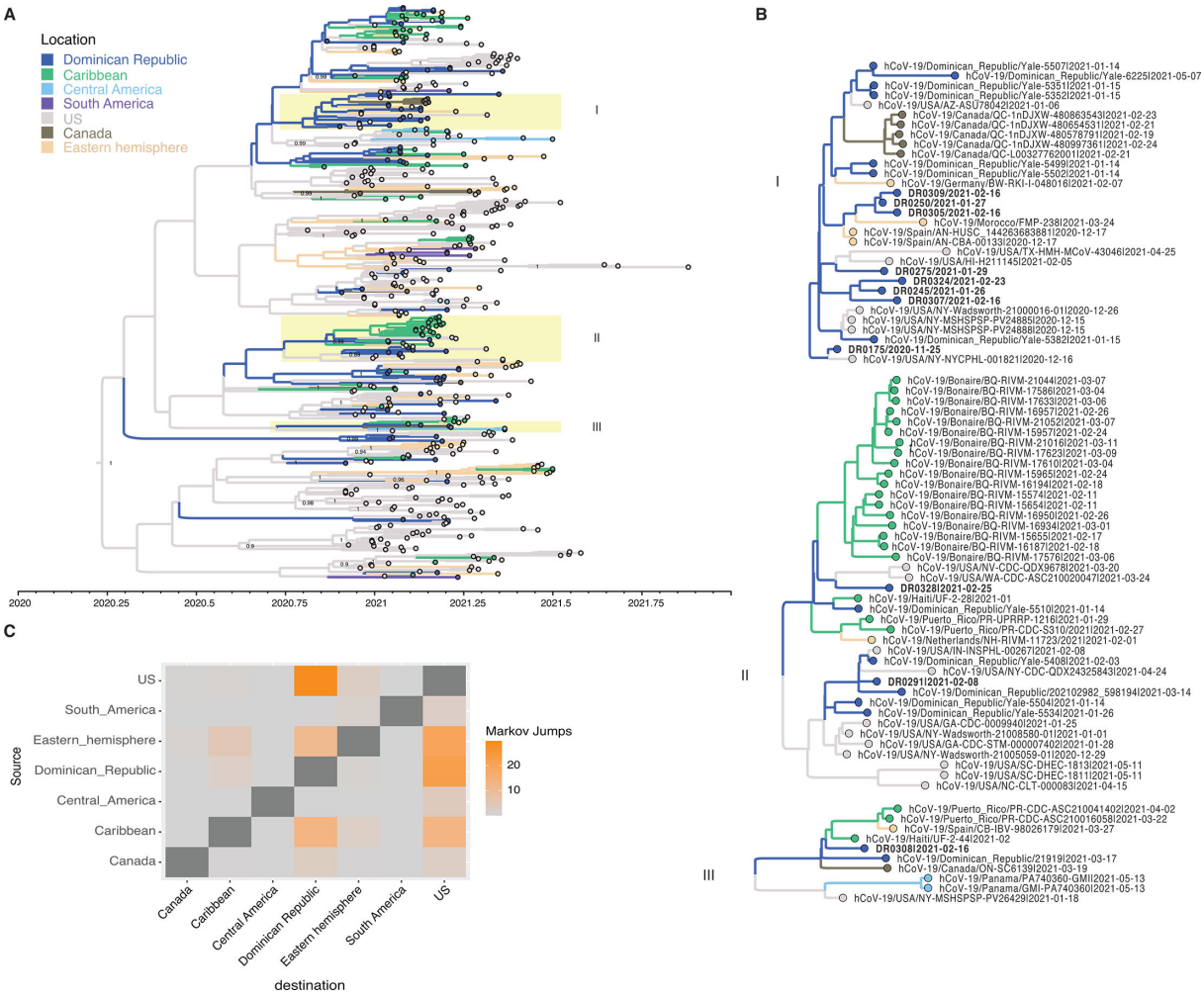

FIG 4 Multiple introductions of the B.1.575 lineage. (A) A Bayesian time-scaled phylogeny was constructed using BEAST. Color represents the location where the sample was collected, and date is plotted on the *x* axis. Clusters that are enlarged in panel B are highlighted with yellow boxes. (B) Clusters I, II, and III, respectively from panel A, enlarged to see regions and tree phylogeny more closely. (C) Heatmap of the Markov jumps from the phylogenetic tree in panel A, where darker orange represents greater number of Markov jumps, and light gray represents low number of Markov jumps.

the location state of trunk branches (often <0.5), to confidently determine the ancestral origin of B.1.575 and whether it originated in the DR or United States (possibly NY). Due to under-sequencing of SARS-CoV-2 from many Caribbean and Central American countries, it was also not possible to determine if these transmission chains comprised movement between additional countries or island states not included in this phylogeny. Regardless, the diversity of the strain and high frequency of detection of this rare lineage in the DR indicate that, although B.1.575 was rare globally, it had an outsized impact on the SARS-CoV-2 epidemic in DR in late 2020 and early 2021 and traveled frequently between the DR and the USA.

A lineage phylogenetics

We also investigated two A.2.5 samples that were collected in February of 2021 in Ozama, DR. To put these two samples into global context, we built a maximum likelihood phylogenetic tree of all publicly available A lineage samples from this time period (Fig. 5A). Based on their distance within the phylogenetic tree, the A.2.5 samples were likely introduced independently into the DR. Both samples are found in clusters including samples from the USA and Panama (Fig. 5B), and one sample clusters most closely to

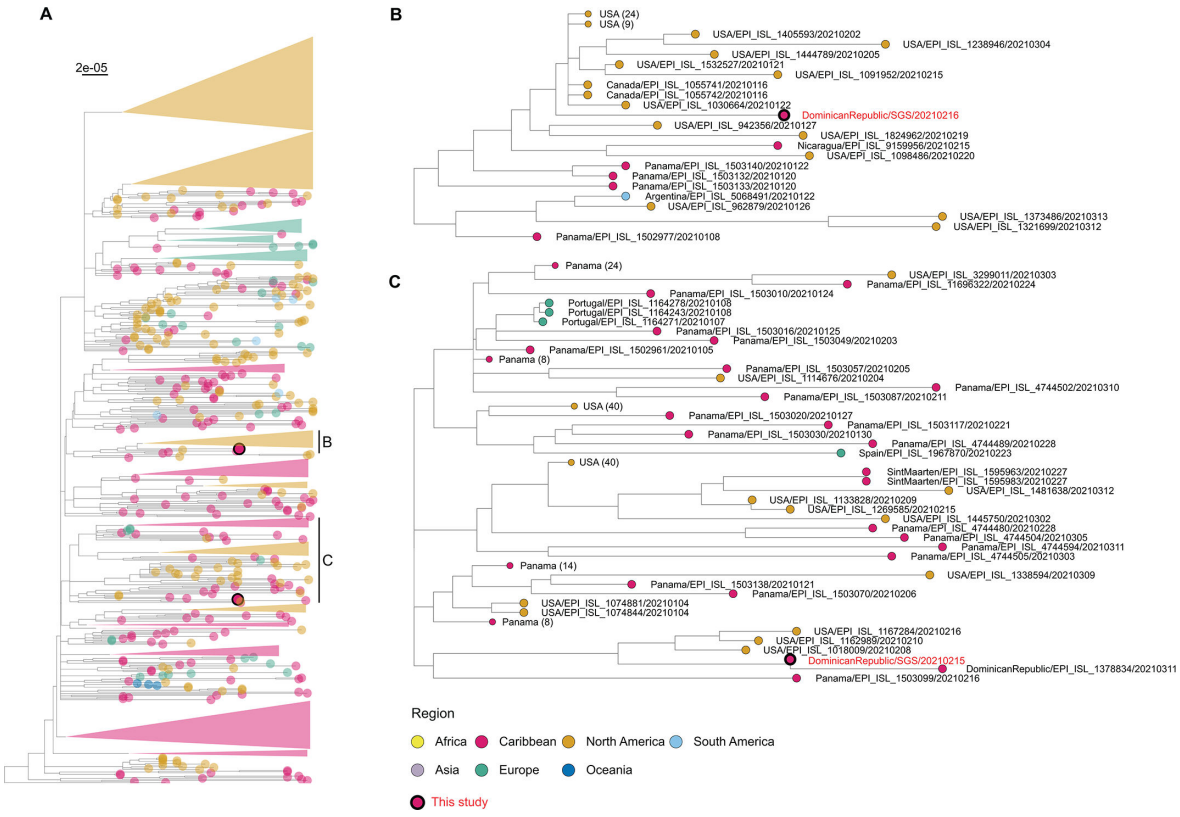

FIG 5 Two introductions of the A lineage into the DR and Caribbean focused spread. Maximum likelihood phylogenetic tree of the A lineage. (A) A phylogenetic tree built on all A lineage samples that are publicly available. The background was randomly downsampled by 25%, then the parental clade containing both samples sequenced in this study was selected. Branch lengths are measured in substitutions/nucleotide site/year. Newick files are included as supplemental data files for additional viewing of this tree. (B and C) Subsections of the full tree in which samples we sequenced cluster. Color represents the region the sample was collected; nodes with bold borders in panels B and C represent samples sequenced for this study (sample names in red), while nodes with lighter borders represent samples downloaded from GISAID.

another DR A lineage sample found on GISAID, suggesting potential transmission within the DR (Fig. 5C).

These transmission patterns between the United States, Panama, and the DR are also supported by patterns in international travel. When looking at the number of passengers flying to the DR between 1 January 2021 and 31 March 2021 around the time the two Lineage A samples were collected, the greatest number of travelers to the DR were on flights that originated in the United States and Panama, making up 70.6% (315,837) and 5.1% (22,788) of all travelers to the DR, respectively. Specifically, 50% (159,905) of all US flights came from NY state, potentially providing further support for transmission chains between NY and DR. Overall, a lineage transmission was rare globally, but more prevalent in regions near the DR, especially Panama and the United States. Tracking of uncommon lineages highlights transmission between and within the United States, Central America, and the Caribbean islands during the beginning of 2021.

DISCUSSION

Among the 98 SARS-CoV-2 genome sequences we analyzed from samples collected between July 2020 and February 2021 in the DR, we detected 16 different lineages. We used a maximum-likelihood phylogenetic tree analysis to determine that multiple introductions of SARS-CoV-2 occurred from North and South America, Asia, Africa, and the Caribbean, and that some lineages, such as B.1.1, were introduced many times independently. B.1.1 was a dominant lineage in the Americas at this time, and it is not surprising that B.1.1 would be introduced into DR frequently. It is less apparent

why a rare lineage such as B.1.575 would be detected at such high frequency into DR, raising the possibility that the lineage originated in DR, rather than being imported multiple times independently. Despite >900 B.1.575 sequences available from New York on GISAID, none were collected prior to December 2020, whereas B.1.575 was sampled several weeks earlier in DR (12 November 2020). Prior studies found evidence of novel SARS-CoV-2 lineages originating in undersampled developing countries (24), supporting this as a possibility. New York has a large Dominican American population (almost 40% of the US Dominican population) (25), which is consistent with high rates of B.1.575 movement between DR and NY. Going forward, sampling in passengers arriving and departing from airport hubs and recording metadata on the port of origin and destination could provide crucial insights into the distribution of viral genetic diversity around the world and the direction of virus movement.

We initially anticipated that many samples of a rare lineage, such as B.1.575, in a small region like the DR sampled during the same epidemiological weeks would suggest a country-wide outbreak. Rather, we found that the within-lineage diversity was consistent with many introductions and small outbreaks. These data suggest a surprising dynamic where a globally rare lineage could be actively moving in and out of the Dominican Republic over the course of months, potentially suggesting missing information and sampling from other regions that could better explain this phenomenon. For example, on epidemiological week 71 a B.1.575 sublineage, B.1.575.2, was reported in the Dominican Republic and later associated with an outbreak in Pamplona, Spain (26). The B.1.575 lineage contains the membrane protein change V70L, which was shown to have greater biological fitness via increased glucose uptake during viral replication compared to the parental B.1 lineage (27). The B.1.575.2 sublineage reported in Spain in the summer of 2021 contained the Spike E484K mutation seen in many variants of concern (28). Genomic surveillance identified changes in fitness and mutational signatures in the B.1.575 lineage, some of which were also found in variants of concern. But important gaps in SARS-CoV-2 genomic surveillance leave open questions regarding why such a rare lineage as B.1.575 would be introduced repeatedly into the DR when global prevalence was low. One potential explanation could be the frequent traveling between the two countries (25), but another explanation could be that other regions with large B.1.575 lineage outbreaks were severely undersampled, such as the neighboring country Haiti (29).

Interestingly, two A.2.5 samples were discovered in our data set from February 2021. The A and B lineages were briefly co-circulating early in the pandemic with the B lineage taking over globally. In the first quarter of 2021, lineage samples made up only 0.38% of all SARS-CoV-2 samples sequenced globally, but 86% of SARS-CoV-2 samples sequenced in Panama (COV-Spectrum, https://cov-spectrum.org/). Although our A.2.5 lineage samples had relatively low sequencing coverage, they appeared to be closely related to ones collected from Panama and the United States, lending support to the idea that Central America and the Caribbean maintained transmission of A lineages even as B lineages became globally dominant.

Some limitations to this study include that 84% of the samples are from Ozama, the region containing the capital city, and thus pandemic dynamics involving rural residents may have been missed. The primer tiling scheme used for sequencing in this study relied on relatively large (~2 kb) amplicons, resulting in lower average coverage than might be expected because failure of just one or a few primers results in larger drop-outs than in a scheme utilizing smaller amplicons, compounding the problem of biospecimen instability. As international consortia develop and adapt best practices for sequencing, measures of robustness should take into account the possibility of variable sample quality due to the conditions under which samples are collected, stored, and shared. Additionally, uneven background sequencing globally still leaves Caribbean-wide dynamics challenging to address and the source of introductions impossible to confidently identify.

Future virus genomic sequencing of under-surveilled regions could provide additional understanding of how a pandemic progresses through those regions, and thus what can be done to prevent future outbreaks. Further, developing global viral sequencing programs can better equip the global community to address emergence and spread of future viral variants before they become widespread.

Conclusion

We sequenced SARS-CoV-2 samples collected from the DR during the early COVID-19 pandemic to better understand the dynamics of lineage transmission across under-surveilled regions, such as the Caribbean islands. We found the presence of 16 different lineages and evidence of multiple SARS-CoV-2 introductions from multiple regions. Even this relatively small data set reflected the presence of what appeared to be a long-lived hotspot of lineage A descendants concentrated in Central America and the Caribbean. The globally prevalent B.1.1 lineage was repeatedly introduced into the DR from several different countries, possibly reflecting the role of international tourism. Finally, the repeated introduction of the globally rare B.1.575 lineage into the DR from the United States is potentially consistent with ongoing epidemiological linkages between immigrant communities and their countries of origin, though absence of contact tracing or individual demographic data precludes definitive conclusions. Altogether, these results highlight surprising epidemiological phenomena and the important role of widespread surveillance sequencing, even when limited in depth, to understand transmission dynamics of SARS-CoV-2 globally.

Highlights

- The A lineage circulated in Central America and the Caribbean throughout 2020, despite a rapid decline in its global prevalence in the early pandemic.
- The globally widespread B.1.1 lineage was likely repeatedly introduced into the DR from several countries.
- Some of the earliest sequences of the rare B.1.575 lineage were found in the DR, either imported from the United States or possibly emerging first in the DR.

ACKNOWLEDGMENTS

This work was supported in part by the Division of Intramural Research (DIR) of the NIAID/NIH and utilized computational resources of the NIH High Performance Computing (HPC) Biowulf cluster (http://hpc.nih.gov). This work was also supported by the Centers of Excellence for Influenza Research and Response, National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services, under contract 75N93021C00014, and by the Intramural Research Program of the US National Library of Medicine at the NIH.

We also want to express our gratitude to the COVID-19 early responders, the Dominican Ministry of Economics, Servicio Nacional de Salud and the Ministry of Health for all their supports in an interinstitutional collaborative agreement to accelerate the Dominican Republic's pandemic response.

A.P., S.M., P.C., and R.P.-R. collected and shared samples from the DR. A.K., A.M., C.M., and S.B. performed WGS sequencing on the samples using a protocol that was designed by S.B., A.R., A.K., A.M., and C.M. Computational variant calling pipelines were run and designed by M.C. and A.R. Analysis was performed by A.K., A.M., and M.I.N. A.K. and A.M. wrote the paper and made figures, and the paper was edited by E.G., A M., M.I.N., and R.P.-R. This work was supported by funding collected by E.G. and R.P.-R.

AUTHOR AFFILIATIONS

[1]Systems Genomics Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland, USA

[2]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, Maryland, USA
[3]Instituto de Medicina Tropical & Salud Global, Universidad Iberoamericana (UNIBE), UNIBE Research Hub, Santo Domingo, Dominican Republic
[4]Department of Applications, Bionuclear, Santo Domingo, Dominican Republic

AUTHOR ORCIDs

Allie Kreitman http://orcid.org/0000-0002-0390-5425
Robert Paulino-Ramírez http://orcid.org/0000-0002-3676-0357
Elodie Ghedin http://orcid.org/0000-0002-1515-725X

FUNDING

Funder	Grant(s)	Author(s)
NIH Intramural Research Program	AI001324	Elodie Ghedin

AUTHOR CONTRIBUTIONS

Allie Kreitman, Conceptualization, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review and editing | Alexandra Mushegian, Conceptualization, Data curation, Formal analysis, Investigation, Supervision, Writing – review and editing | Martha I. Nelson, Data curation, Formal analysis, Methodology, Writing – review and editing | Stephanie Banakis, Investigation, Project administration, Resources | Christopher Mederos, Investigation | Matthew Chung, Data curation, Software | Allison Roder, Investigation, Resources | Armando Peguero, Investigation, Resources | Sayira Mueses, Investigation, Resources, Writing – review and editing | Paula Cuevas, Investigation | Robert Paulino-Ramírez, Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – review and editing | Elodie Ghedin, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Visualization, Writing – review and editing

DATA AVAILABILITY

Data and scripts, excluding the OAG flight data, are available on github (https://github.com/GhedinSGS/SARS-CoV-2-Genomic-Epidemiology-in-the-Dominican-Republic). Raw sequence data are available on SRA under the Bioproject number PRJNA1223001. Assembled sequences are available on GenBank under accession numbers PV101376 to PV101456. GISAID IDs of sequences are provided in Table S6.

ETHICS APPROVAL

Secondary analysis of de-identified samples previously collected for diagnostic or surveillance purposes was approved by Universidad Iberoamericana (UNIBE) Institutional Review Board CEI-2020-16, and the National Bioethical Committee (CONABIOS 020-2021). These entities explicitly approved the export shipment of extracted RNA to the United States for genome sequencing.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Data Set S1 (Spectrum01105-25-s0001.xlsx). Metadata for global newick tree in Fig. 2.

Data Set S2 (Spectrum01105-25-s0002.xlsx). Metadata for the full newick tree in Fig. 5.
Text S1 (Spectrum01105-25-s0003.txt). Newick file of the global tree in Fig. 2B.
Text S2 (Spectrum01105-25-s0004.txt). Newick file of the full tree in Fig. 5.
Supplemental legend (Spectrum01105-25-s0005.docx). Description of nwk tree and metadata from Nextstrain.
Supplemental tables (Spectrum01105-25-s0006.xlsx). Tables S1 to S6.

REFERENCES

1. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, Xu X, Wu Y, Lu W, Ge S, Zhao Z, Yang J, Leung DT, Domman DB, Yu H. 2022. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet 54:499–507. https://doi.org/10.1038/s41588-022-01033-y
2. Bugembe DL, Phan MVT, Ssewanyana I, Semanda P, Nansumba H, Dhaala B, Nabadda S, O'Toole ÁN, Rambaut A, Kaleebu P, Cotten M. 2021. Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. Nat Microbiol 6:1094–1101. https://doi.org/10.1038/s41564-021-00933-9
3. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, Crispim MAE, Sales FCS, Hawryluk I, McCrone JT, et al. 2021. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science 372:815–821. https://doi.org/10.1126/science.abh2644
4. Tegally H, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, Subramoney K, Makatini Z, Moyo S, Amoako DG, et al. 2022. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. Nat Med 28:1785–1790. https://doi.org/10.1038/s41591-022-01911-2
5. Wilkinson E, Giovanetti M, Tegally H, San JE, Lessells R, Cuadros D, Martin DP, Rasmussen DA, Zekri A-RN, Sangare AK, et al. 2021. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. Science 374:423–431. https://doi.org/10.1126/science.abj4336
6. Dellicour S, Hong SL, Vrancken B, Chaillon A, Gill MS, Maurano MT, Ramaswami S, Zappile P, Marier C, Harkins GW, Baele G, Duerr R, Heguy A. 2021. Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. PLoS Pathog 17:e1009571. https://doi.org/10.1371/journal.ppat.1009571
7. Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D, Bloom JD, Veesler D, Mateo D, Hernando A, Comas I, González-Candelas F, SeqCOVID-SPAIN consortium, Stadler T, Neher RA. 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature 595:707–712. https://doi.org/10.1038/s41586-021-03677-y
8. López MG, Chiner-Oms Á, García de Viedma D, Ruiz-Rodriguez P, Bracho MA, Cancino-Muñoz I, D'Auria G, de Marco G, García-González N, Goig GA, et al. 2021. The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant. Nat Genet 53:1405–1414. https://doi.org/10.1038/s41588-021-00936-6
9. Anonymous. 2022. UK Completes over 2 Million SARS-CoV-2 whole genome sequences. Available from: https://www.gov.uk/government/news/uk-completes-over-2-million-sars-cov-2-whole-genome-sequences
10. Mallapaty S. 2021. India's neighbours race to sequence genomes as COVID surges. Nature 593:485–486. https://doi.org/10.1038/d41586-021-01287-2
11. Oude Munnink BB, Worp N, Nieuwenhuijse DF, Sikkema RS, Haagmans B, Fouchier RAM, Koopmans M. 2021. The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. Nat Med 27:1518–1524. https://doi.org/10.1038/s41591-021-01472-w
12. Lucien MAB, Forde MS, Isabel MR, Boissinot M, Isabel S. 2023. Infectious diseases genomic surveillance capacity in the Caribbean: a retrospective analysis of SARS-CoV-2. Lancet Reg Health Am 18:100411. https://doi.org/10.1016/j.lana.2022.100411
13. Paulino-Ramirez R, Riego E, Vallejo-Degaudenzi A, Calderon VV, Tapia L, León P, Licastro D, Dal Monego S, Rajasekharan S, Orsini E, Marcello A. 2021. Whole-genome sequences of SARS-CoV-2 isolates from the Dominican Republic. Microbiol Resour Announc 10:e00952-21. https://doi.org/10.1128/MRA.00952-21
14. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, Colquhoun R, Ruis C, Abu-Dahab K, Taylor B, Yeats C, du Plessis L, Maloney D, Medd N, Attwood SW, Aanensen DM, Holmes EC, Pybus OG, Rambaut A. 2021. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol 7:veab064. https://doi.org/10.1093/ve/veab064
15. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 5:1403–1407. https://doi.org/10.1038/s41564-020-0770-5
16. Aksamentov I, Roemer C, Hodcroft EB, Neher RA. 2021. Nextclade: clade assignment, mutation calling and quality control for viral genomes. JOSS 6:3773. https://doi.org/10.21105/joss.03773
17. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123. https://doi.org/10.1093/bioinformatics/bty407
18. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300
19. Wickham H. 2016. Ggplot2: elegant graphics for data analysis. Springer Verlag, New York.
20. Yu G. 2020. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics 69:e96. https://doi.org/10.1002/cpbi.96
21. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:vey016. https://doi.org/10.1093/ve/vey016
22. Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA. 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. Syst Biol 68:1052–1061. https://doi.org/10.1093/sysbio/syz020
23. Umakanthan S, Chauhan A, Gupta MM, Sahu PK, Bukelo MM, Chattu VK. 2021. COVID-19 pandemic containment in the Caribbean region: a review of case-management and public health strategies. AIMS Public Health 8:665–681. https://doi.org/10.3934/publichealth.2021053
24. Singh J, Pandit P, McArthur AG, Banerjee A, Mossman K. 2021. Evolutionary trajectory of SARS-CoV-2 and emerging variants. Virol J 18:166. https://doi.org/10.1186/s12985-021-01633-w
25. Moslimani M, Noe-Bustamante L, Shah S. 2023. Facts on Hispanics of Dominican origin in the United States. Available from: www.pewresearch.org/hispanic/fact-sheet/us-hispanics-607 facts-on-dominican-origin-latinos
26. Paulino-Ramírez R, López P, Mueses S, Cuevas P, Jabier M, Rivera-Amill V. 2023. Genomic surveillance of SARS-CoV-2 variants in the Dominican Republic and emergence of a local lineage. Int J Environ Res Public Health 20:5503. https://doi.org/10.3390/ijerph20085503
27. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. 2021. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. Emerg Microbes Infect 10:885–893. https://doi.org/10.1080/22221751.2021.1922097
28. Trobajo-Sanmartín C, Miqueleiz A, Portillo ME, Fernández-Huerta M, Navascués A, Sola Sara P, López Moreno P, Ordoñez GR, Castilla J, Ezpeleta C. 2021. Emergence of SARS-CoV-2 variant B.1.575.2, containing the E484K mutation in the spike protein, in Pamplona, Spain, May to June 2021. J Clin Microbiol 59:e01736-21. https://doi.org/10.1128/JCM.01736-21
29. Mushegian A, Kreitman A, Nelson MI, Chung M, Mederos C, Roder A, Banakis S, Desormeaux AM, Jean Charles NL, Grant-Greene Y, Marseille S, Pierre K, Lafontant D, Boncy J, Journel I, Buteau J, Juin S, Ghedin E. 2024. Genomic analysis of the early COVID-19 pandemic in Haiti reveals Caribbean-specific variant dynamics. PLoS Glob Public Health 4:e0003536. https://doi.org/10.1371/journal.pgph.0003536