Genomic surveillance of SARS-CoV-2 in Puerto Rico reveals emergence of an autochthonous lineage and early detection of variants

Gilberto A. Santiago (gsantiago@cdc.gov)
Centers for Disease Control and Prevention https://orcid.org/0000-0002-8864-9503
Betzabel Flores
Centers for Disease Control and Prevention
Glenda L. Gonzalez
Centers for Disease Control and Prevention
Keyla N. Charriez
Centers for Disease Control and Prevention
Limari Cora-Huertas
University of Puerto Rico, Rio Piedras
Hannah R. Volkman
Centers for Disease Control and Prevention https://orcid.org/0000-0001-7274-9641
Steven Van Belleghem
University of Puerto Rico, Rio Piedras
Vanessa Rivera-Amill
Ponce Health Sciences University
Laura E. Adams
Centers for Disease Control and Prevention
Melissa Marzan
Puerto Rico Department of Health
Lorena Hernandez
Puerto Rico Department of Health
Iris Cardona
Puerto Rico Department of Health
Eduardo O'Neill
Centers for Disease Control and Prevention
Gabriela Paz-Bailey
Centers for Disease Control and Prevention
Riccardo Papa
University of Puerto Rico, Rio Piedras
Jorge L. Munoz-Jordan
Centers for Disease Control and Prevention

Article

Keywords: SARS-CoV-2, genomic surveillance, NGS, Puerto Rico, variants, B.1.588

Posted Date: January 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1277781/v1

Version of Record: A version of this preprint was published at Communications Medicine on August 11th, 2022. See the published version at https://doi.org/10.1038/s43856-022-00168-7.

Abstract

Puerto Rico has experienced the full impact of the COVID-19 pandemic. Since SARS-CoV-2, the virus that causes COVID-19, was first detected on the island in March of 2020, it spread rapidly though the island's population and became a critical threat to public health. We conducted a genomic surveillance study through a partnership with health agencies and academic institutions to understand the emergence and molecular epidemiology of the virus on the island. We sampled COVID-19 cases monthly over 19 months and sequenced a total of 753 SARS-CoV-2 genomes between March 2020 and September 2021 to reconstruct the local epidemic in a regional context using phylogenetic inference. Our analyses revealed that multiple importation events propelled the emergence and spread of the virus throughout the study period, including the introduction and spread of most SARS-CoV-2 variants detected world-wide. Lineage turnover cycles through various phases of the local epidemic were observed, where the predominant lineage was replaced by the next competing lineage or variant after approximately 4 months of circulation locally. We also identified the emergence of lineage B.1.588, an autochthonous lineage that predominated circulation in Puerto Rico from September to December 2020 and subsequently spread to the United States. The results of this collaborative approach highlight the importance of timely collection and analysis of SARS-CoV-2 genomic surveillance data to inform public health responses.

Introduction

The current coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was initially declared a Public Health Emergency of International Concern in January 2020[1, 2]. Despite global efforts to interrupt transmission chains with quarantine, isolation and travel restrictions at the onset of the pandemic, SARS-CoV-2 spread rapidly across the globe, creating a global pandemic and threat to human health worldwide. By November 15th, 2021, the World Health Organization (WHO) reported 253 million confirmed COVID-19 cases in 222 countries and over 5 million deaths[3, 4]. SARS-CoV-2 reached all 50 states of the United States and associated territories, including Puerto Rico, by March 2020, after multiple introductions by travelers with infection[5, 6, 7]. The rapid spread across the United States was primarily propelled by interstate transmission chains and air travel to the associated territories[6, 8, 9].

SARS-CoV-2 is an enveloped virus with a single-stranded positive-sense RNA genome of approximately 30,000 base pairs. During replication, a virus-encoded exonuclease provides a proof-reading activity that contributes to the observed low mutation rate and stable genome[10, 11]. Nevertheless, the unprecedented spread of SARS-CoV-2 globally and the wealth of genomic sequence data available through the international initiative for genomic studies and surveillance has facilitated phylodynamic approaches to infer viral evolutionary rate, growth rate, and estimated time of origin for specific outbreaks[11]. Studies have revealed that the viral genome has been accumulating mutations of concern, especially in the spike protein region, which confer phenotypes with increased fitness and pathogenicity[12, 13, 14]. Increased infectivity, resistance to monoclonal antibody therapy and evasion of the immune response were among the most frequently observed phenotypes attributed to WHO-monitored variants; these phenotypes often dominated transmission and replacement of other lineages upon emergence[15, 16, 17, 18]. The Variant Being Monitored (VBM) B.1.1.7 (Alpha) was first identified in the United States in late December 2020 and was then characterized by a considerable increase in COVID-19 incidence associated with increased infectivity and occasionally more severe disease manifestations that increased hospitalization rates[19, 20]. Alpha became the dominant variant, especially in Europe and the United States, until the emergence of Variant of Concern (VOC) B.1.617.2/AY.x (Delta), first identified in the United States in May 2021, which developed into a prominent variant with an apparent higher virulence and pathogenic phenotype[21, 22, 23]. Because of the potential for increased transmissibility, morbidity mortality, and decreased efficacy of vaccines and other intervention strategies, monitoring the spread of variants (VBMs and VOCs) rapidly became a public health concern and priority[24, 25].

Puerto Rico, an unincorporated territory of the United States, is a densely populated island and a popular tourist destination located in the Caribbean basin. SARS-CoV-2 was first identified in Puerto Rico on March 13th, 2020, in two European travelers who arrived on a cruise ship and in one local resident who had close contact with family members with recent travel history. Additional travel-related and local cases were confirmed within the following weeks[26]. In response to the emerging threat, the government of Puerto Rico executed the most restrictive (compared to the United States) national stay-at-home order on March 15th, 2020, to mitigate transmission while preparing the public health infrastructure for the imminent impact[27, 28]. Travel restrictions imposed by the United States during the initial pandemic minimized international traffic to Puerto Rico, although domestic travel from the United States continued. Puerto Rico represents a unique epidemiologic setting in a geographically isolated location (an island), but with a regular influx of travelers mostly from the United States. This is an ideal setting to monitor introduction and spread of SARS-CoV-2 variants and answer questions to help inform SARS-CoV-2 spread and disease prevention strategies. Puerto Rico's public health response incorporated the increase of laboratory capacity extensive molecular surveillance, which presented a unique opportunity to study the impact of SARS-CoV-2 variant turnover, local dissemination, and evolution during a period of changing epidemiology and public health responses.

In response to the impending local epidemic, we established a partnership with the local health authorities and academia to conduct a genomic surveillance initiative to sample complete genomes of SARS-CoV-2 across the island through time, monitor lineage circulation and understand the genomic epidemiology of the COVID-19 pandemic in Puerto Rico. This report presents the results from 19 months of genomic surveillance and phylogenetic analyses, which identified multiple introduction events that propelled the rapid expansion and persistent transmission of the virus on the island and lead to the establishment of an autochthonous lineage between August 2020 and January 2021.

Results

Local epidemic and variant detection

During March–July 2020, the number of confirmed COVID-19 cases remained low, associated with the strict stay-at-home order. When the order was lifted in June 2020, the number of cases increased steeply during the summer of 2020, initiating the local epidemic (Figure 1A). Since then, we observed 3 epidemic

peaks in December 2020, April 2021, and August 2021. By September 30th, 2021, the Puerto Rico Department of Health (PRDH) and the Centers for Disease Control and Prevention (CDC) reported over 181,599 confirmed cases[26].

We conducted genomic surveillance for 19 months since March 2020, where, each month, we sequenced SARS-CoV-2 positive specimens from recently symptomatic and asymptomatic patients residing in 63 of the 78 municipalities, covering all 7 health regions of the island. The frequency of lineages detected was calculated periodically from all viral genomic sequences from Puerto Rico published in GISAID, including our sequence datasets, and reported to the PRDH to inform case investigations and surveillance (Figure 1B). Our data comprises the genomes from the initial SARS-CoV-2 confirmed infections detected in March 2020, which included 3 European tourists that arrived on the island in a cruise ship and 8 residents with no recent travel history declared. PANGO lineage assignment after sequencing identified lineage A.1, a lineage predominant in Europe at the time, in the 3 travelers with infection, while lineages B.1 and B.1.1 were identified in residents with infection. B.1x lineages predominated in the United States and the Americas. The initial phase of the epidemic was characterized by the detection of a wide diversity of B.1x lineages that circulated at low frequency for short periods of time, suggesting that the local epidemic was initiated by multiple introduction events. In August 2020, we detected the emergence of lineage B.1.588 in various municipalities of the island. Lineage B.1.588 rapidly became the predominant lineage in Puerto Rico, circulating at high frequency for approximately 4 months and causing the first epidemic peak in November 2020 (Figures 1A and 1B). Circulation of lineage B.1.588 declined during the first peak of the epidemic in the winter of 2020, a season of local holiday festivities and frequent travel. During this season, the diversity of B.1x lineages increased, and the first emergent variants were detected in the island, VBM B.1.427/429 (Epsilon) in December 2020 and Alpha in January 2021 concordant to variant emergence in the United States (Figure 1B). Concurrently, the first stage of the COVID-19 vaccination campaign in Puerto Rico started in mid-December 2020 for the elderly population and first responders. A steep reduction in confirmed cases was observed in the following months despite the predominant circulation of Alpha and the introduction of VBM B.1.526 (Iota) in February 2021 and P.1/1.1 (Gamma) in April 2021 (Figure 1B). The second peak of the epidemic was observed in April 2021 with Alpha predominating (Figures 1A and 1B). Though other emerging variants continued to be detected, the frequency of detection remained low, and Alpha predominated for approximately 3-4 months. The second stage of the COVID-19 vaccination campaign started in April 2021 for all adults and was immediately followed by a sharp decrease in confirmed cases, a period in which approximately 50% of the population had received at least one dose of the vaccine[26] (Figure 1A). VOC B.1.617.2/AY.x (Delta) was first detected in June 2021, concordant with the emergence in the United States, and rapidly dominated transmission. During the same period, we detected the emergence of VBM B.1.621 (Mu), which caused a small local outbreak in the western part of the island, as well as a modest increase in Gamma infections (Figure 1B). The third epidemic peak was observed in August 2021, coinciding with a summer of increased travel and the removal of local government-imposed restrictions on business indoor occupant capacity and public gatherings (Figures 1A and 1B). During this period, most COVID-19 cases in Puerto Rico were caused by Delta and approximately 18 Delta sub-lineages were detected in the island, with AY.3 as the most frequently sampled sub-lineage (Figure 1C). By September 30th, 2021, a steep decrease in confirmed cases was observed, a point in which more than 77% of the population had received at least one dose of the vaccine (Figure 1A).

Phylogenetic reconstruction of the local pandemic

This study generated 753 complete genomes from viruses sampled between March 2020 and September 2021. Our dataset was combined with 2,611 publicly available genomes in GISAID to understand the emergence and spread of the viruses circulating in Puerto Rico in a global context. We reconstructed the local and regional epidemic using a time-calibrated phylogenetic tree inferred with maximum likelihood (Figure 2). This global phylogenetic analysis estimated that the initial SARS-CoV-2 introductions occurred between February 19 and March 16, 2020. Most viral genomes from Puerto Rico descend or are closely related to genomes from the United States. However, we were unable to determine the precise origin at the state level due to the limited sampling during the emergence period and subsequent low circulation. The resulting tree topology inferred viral sequences from Puerto Rico scattered across the global tree, smaller short-lived monophyletic clusters, and larger monophyletic clusters that suggest sustained transmission of a particular genotype. Our analysis also showed the emergence and spread of the SARS-CoV-2 variants detected in Puerto Rico. Multiple monophyletic clusters of Puerto Rican sequences were inferred within the clades formed by each emergent variant and the size of the clades is proportional to the frequency of genomes sampled in the island (Figures 1B and 2). The observed clustering patterns indicate multiple virus introductions with rapid and explosive expansion across the island in a short period of time.

Detection and spread of autochthonous lineage B.1.588

During the initial phase of epidemic transmission, we detected the emergence of an autochthonous lineage, B.1.588, which rapidly spread across the island. Based on GISAID data and cov-lineages.org reports (https://cov-lineages.org/lineages), lineage B.1.588 was first detected in Puerto Rico on August 2nd, 2020: sequence EPI_ISL_1168693. Initially, lineage B.1.588 circulated only in Puerto Rico, accounting for approximately half of the viruses sampled in the island in September 2020. B.1.588 quickly became the predominant lineage in Puerto Rico during the first epidemic peak, circulating for 4 months until it was replaced by the emergence of Alpha in January 2021 (Figure 2). This study sequenced 97 out of the 115 B.1.588 genomes from Puerto Rico found in GISAID. To understand the emergence and spread of this lineage, we reconstructed a phylogenetic sub-tree using maximum likelihood and Bayesian inference with 103 B.1.588 sequences from Puerto Rico, 58 B.1.588 sequences from the United States and an additional set of 77 B.1 lineage sequences closely related to B.1.588 (Figure 3). Our analysis estimated that lineage B.1.588 diverged from its parental lineage B.1 between May 21st, 2020, and July 16th, 2020 in Puerto Rico, after the appearance of two non-conservative mutations: T20I in the spike protein and M234I in the nucleocapsid protein. Subsequently, lineage B.1.588 spread broadly to the United States, mainly in New York, Texas, Florida, and California, where it circulated until May 2021 concomitant with a diversity of other lineages and variants. More than 990 B.1.588 genomes have been reported in the United States.

Emergence of SARS-CoV-2 variants

VBM Alpha was first detected in Puerto Rico in January 2021, co-circulating with local predominant lineage B.1.588 and other B lineages at a lower rate (Figure 1B). Notably, this VBM replaced the well-established autochthonous lineage B.1.588. The emergence and epidemiology of Alpha in Puerto Rico

resembled the patterns observed in the United States, with rapid spread and a sharp increase in confirmed cases[19, 20]. To understand the emergence and spread of Alpha in Puerto Rico, we inferred a maximum likelihood phylogenetic sub-tree with all Alpha genomes obtained in our dataset in addition to a subset of other Alpha genomes from Puerto Rico, the United States, and a regional context backdrop (Figure 4). The resulting inference estimated that the emergence of Alpha in Puerto Rico may have occurred between November 6th, 2020, and December 31st, 2020. Tree topology showed multiple monophyletic clusters of Puerto Rican sequences diverging across a period of 4-5 months of circulation. The larger clusters of Puerto Rican sequences suggest that local transmission of specific Alpha genotypes was sustained, succeeding after in multiple introduction events. Most of these clusters were associated with sequences from the United States, suggesting that multiple introductions occurred over a period of 5-6 months, propelling the local transmission of this variant. We also found a subset of Puerto Rican sequences associated with sequences from the Caribbean and the Americas but low node support impaired resolution of transmission patterns.

VOC Delta was first detected in Puerto Rico in June 2021, during a period when SARS-CoV-2 transmission was declining, and the vaccination campaign was progressing rapidly (Figures 1A and 1B). After its initial detection, Delta spread rapidly across the island (Figure 1B). Over 30% of the COVID-19 cases sampled and sequenced in June 2021 were caused by Delta. This variant has been characterized broadly as the most dominant emerging variant, replacing most lineages, and causing most of COVID-19 cases in the United States and Puerto Rico from its emergence through November 2021. To understand the rapid emergence and spread of Delta in the island, we reconstructed a maximum likelihood phylogenetic sub-tree with all Delta Puerto Rican sequences obtained in our dataset supplemented with additional sequences from Puerto Rico and the United States retrieved from GISAID with collection dates between May 1st, 2021, and September 31st, 2021. According to our phylogenetic inference, the emergence of Delta in Puerto Rico may have occurred between April 15th and June 14th, 2021, potentially after one or multiple introductions. The precise origin of the introductions was challenging to resolve, considering that multiple sequences from Mexico, the United States, and the Caribbean cluster among the early sampled sequences from Puerto Rico with low node support (Figure 5). The first Delta lineage to be detected was B.1.617.2, which seems directly related to a small number of VBM B.1.617.1 (Kappa) that clustered basal to the sub-tree. Tree topology is similar to the patterns observed in the Alpha sub-tree, where more than 17 distinct clusters with sequences from Puerto Rico were observed diverging across 4 months of circulation in the island. Most of these clusters were associated with distinct Delta sub-lineages and seem closely related to similar sequences from the United States and the Caribbean. These clustering patterns and the diversity of Delta sub-lineages detected suggest that multiple introductions throughout 5-6 months propelled the emergence and transmission of this variant in the island.

Discussion

Despite initial containment of COVID-19 cases in the spring of 2020, SARS-CoV-2 subsequently spread rapidly across the island propelled by multiple introductions and extensive intra-island transmission. Our partnership responded rapidly with the launch of this collaborative study aimed at enhancing the local capacity for genomic surveillance, understanding the development of the epidemic, and tracking viral lineages of public health concern. Our efforts also led to the early identification and tracking of most of the SARS-CoV-2 variants of lineages circulating in the island; information that was reported monthly to the Puerto Rico Department of Health (PRDH) to inform case investigation and epidemic response. As a result, our study contributed more than 75% of the genomes from the initial phase of the local epidemic through December 2020, and overall, to over 23% of the genomes sampled in Puerto Rico published in GISAID by October 31st, 2021.

Our systematic genomic surveillance and phylogenetic analyses suggest that the local epidemic was initiated and frequently boosted by travel-associated cases. This hypothesis is supported by the diversity of B.x lineages detected in the island during a period when the local population movement was restricted by stay-at-home orders and other response measures enforced by the local government[27, 28]. In addition, the identification of multiple clusters of Puerto Rican sequences phylogenetically related to viruses found circulating primarily in the United States further supports this assessment. The increase in COVID-19 cases observed in the summer of 2020 coincided with lifting the stay-at-home order and increased traffic of tourists during a period of high SARS-CoV-2 transmission in the United States. Travel restrictions between countries with high incidence and the United States affected Puerto Rico by interrupting international travel but potentially increasing the volumes of domestic travel to and from the island. Consequently, we did not observe frequent clustering of Puerto Rican sequences with sequences from regions other than the Americas. The frequent mixing of viruses between Puerto Rico and the United States, the lineage diversity circulating in the island, and density of sampling at the location of origin generated phylogenetic uncertainty, limiting the capacity to resolve the precise origin of some local clusters and specific transmission chains[29].

Upon the emergence and rapid spread of SARS-CoV-2 during the first 10 months of the epidemic (March 2020-January 2021), we observed a shift in phylogenetic clustering patterns that most likely resulted from *in situ* viral evolution and adaptation to the local population or environment. This pattern was observed with the divergence of lineage B.1.588 from the persistent local transmission of B.1 lineage, possibly due to founder effect. The fixation of two non-conservative mutations in the spike and nucleocapsid proteins seem to have improved fitness for local circulation. Similar clustering patterns resembling adaptation, as well as the divergence of local lineages, have been observed in other regions of the world after a period of sustained local transmission leading to *in situ* evolution, including in the United Kingdom after the re-opening of a national lockdown [30, 31, 32, 33]. Interestingly, B.1.588 declined during a period of peak transmission of SARS-CoV-2 and increased travel in which Alpha was potentially introduced. These observations indicate the limited effectiveness of the initial efforts to prevent SARS-CoV-2 introductions and spread. The subsequent decline of the first epidemic peak could potentially be attributed to the combination of the local government response measures and the launch of the vaccination campaign.

The predominance of Alpha and the subsequent second epidemic peak could suggest that this variant presented a more virulent phenotype with a higher infectivity rate that outcompeted the autochthonous B.1.588 and other lineages in January 2021. This is consistent with the recurring observation that, once VBM/VOC emerged in the island, the frequency of other B.1x lineages was reduced substantially from that point forward. It is possible that this dynamic may have contributed to increased transmission during the initial phase of vaccination before the local population reached a 50% vaccination rate. Our findings are concordant with the epidemiological scenario in the United States.

As the vaccination campaign progressed during the spring of 2021, the number of cases decreased substantially, and the local government reduced most restrictions to public gatherings and indoor activities[27]. Puerto Rico then experienced another wave of increased travel during the summer of 2021 coinciding with the emergence of Delta. Though this variant was first detected in June 2021, our analyses estimated that the variant could have been introduced during the second epidemic peak in April 2021. If so, Delta could have faced a complex scenario with Alpha predominating circulation and almost half of the population vaccinated against the virus. However, the transmission of Delta displaced most of the circulating lineages and led to the third epidemic peak. Two different scenarios could be proposed for the emergence of Delta in Puerto Rico. First, Delta could have been introduced when Alpha dominated transmission, the two variants co-circulated, but Delta presented the phenotype that could outcompete Alpha. Key mutation patterns in Delta have been proposed to confer an advantage over other lineages[34]. Alternatively, the transmission of Alpha could have been already declining when Delta was introduced and emerged in a susceptible population. This second scenario proposed a gap between the dominance of Alpha and Delta. Recent reports from Madrid suggest that virus competition upon the emergence of Delta was not the exclusive factor driving the decline of Alpha but also a period of declined Alpha transmission facilitating the emergence of Delta in the region [35]. Though it is possible to speculate that Delta's phenotype exhibited some resistance to the vaccine, the decline of the third epidemic peak after approximately 4 months of Delta circulation, coincides with the population reaching a 77% vaccination rate with at least one dose.

The pattern of lineage turn-over observed in this study, where the predominant variants circulated for approximately 4 months and then were replaced by another variant in Puerto Rico, should be further compared with transmission in larger countries with larger populations and human movement. Similar patterns have been observed in the United States, where variant emergence and spread potentially affect Puerto Rico[50]. Travel restrictions could have blocked the introduction of additional variants directly to Puerto Rico through international travel. However, our results suggest that lineage turn-over was in part driven by domestic travel with the United States. We also speculate that these cycles could be related to the limitations of an island geography, the rapid response of the local government with restrictive measures to controls the spread of the virus and a vaccination campaign that reached over 84% of the population by December 1st, 2021[26].

Although our study sampled viral genomes from various cases and periods during the local pandemic, the universal ARTIC v3 tiled PCR-amplicon NGS workflow's sensitivity used in this study is limited to specimens with PCR Ct values below 28. Thus, we question if there are viral genetic differences from infections with lower viral loads or inconsistent variant-calling from samples with lower viral loads that might affect the accuracy of lineage assignment[36, 37, 38]. The sampling for this study was also limited due to the ability of our partnership to procure samples from every municipality of the island, especially during the first year of the local epidemic. Future national genomic surveillance programs could benefit from improved systematic sampling engaging with clinical laboratories to ensure timely reporting of results to the local public health authorities, proper sample storage, and transfer to sequencing laboratories. Regarding phylogenetic analyses, the slow mutation rate and the low genetic diversity of the virus frequently impair the resolution of internal nodes with statistical support affecting the interpretation of phylogenetic histories and geographic origins. These analyses could also be affected by selecting context genomes from the unprecedented abundance of genomic data published in GISAID.

This study provides an overview of the COVID-19 epidemic in Puerto Rico during March 2020–November 2021 from the molecular epidemiology perspective. The documentation of an autochthonous lineage and dynamics of virus movement between the United States and Puerto Rico is important to inform prevention and surveillance efforts in both regions. Our phylogenetic study offers the genomic framework to understand the genomic changes occurring through time in the Puerto Rican population and elucidate the mutational landscape within this region. Furthermore, our ongoing genomic surveillance initiative will facilitate the study of SARS-CoV-2 intra-island phylodynamics and compare pre- and post-vaccination populations. Finally, this report highlights the importance of government and academic partnerships to respond to public health threats and the potential of systematic genomic surveillance to improve disease prevention and control.

Declarations

DISCLAIMER

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the National Institute of Health.

ACKNOWLEDGEMENTS

We thank our partners from the Puerto Rico Department of Health, especially the staff from the Institute of Public Health Laboratories and the Biological and Chemical Emergencies Laboratory for their contribution to the genomic surveillance framework and sample procurement. We also thank Dr. Anderson Brito and Chaney Kalinich, former members of the Dr. Nathan Grubaugh laboratory at the Yale School of Public Health, New Haven, CT, for providing invaluable technical assistance with data analysis. We acknowledge all healthcare workers and authors submitting data to GISAID, Supplementary Table 2.

This project was partially funded by the Centers for Disease Control and Prevention's Advanced Molecular Detection Program and the COVID-19 Laboratory Task Force. Additional funding for this study was provided by the National Institute of General Medical Sciences of the National Institute of Health under award number U54GM133807 (NoA: 5U54GM133807-02) to Dr. Riccardo Papa. This project was also supported by UPR COVID-19 emergency funds #2020-2488 to Dr. Riccardo Papa.

References

1. World Health Organization (WHO). Novel Coronavirus (2019-nCoV). Situation Report-1. WHO. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed on 25-January-2021).

2. World Health Organization (WHO). Novel Coronavirus (2019-nCoV). Situation Report-51. WHO. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed on 25-January-2021).
3. World Health Organization (WHO). Weekly epidemiological update on COVID-19. Edition 66. 16 November 2021. WHO. Available online: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—16-november-2021 (accessed on 19-November-2021).
4. World Health Organization (WHO). WHO Coronavirus (COVID-19) Dashboard. Available online: https://covid19.who.int/ Accessed on 19 November 2021.) (2021).
5. Fauver JR, *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell **181**, 990-996.e995 (2020).
6. Gharpure R, *et al.* Multistate Outbreak of SARS-CoV-2 Infections, Including Vaccine Breakthrough Infections, Associated with Large Public Gatherings, United States. Emerg Infect Dis **28**, (2021).
7. Worobey M, *et al.* The emergence of SARS-CoV-2 in Europe and North America. Science **370**, 564–570 (2020).
8. Perkins TA, Cavany SM, Moore SM, Oidtman RJ, Lerch A, Poterek M. Estimating unobserved SARS-CoV-2 infections in the United States. Proc Natl Acad Sci U S A **117**, 22597–22602 (2020).
9. Zeller M, *et al.* Emergence of an early SARS-CoV-2 epidemic in the United States. Cell **184**, 4939-4952 e4915 (2021).
10. Minskaia E, *et al.* Discovery of an RNA virus 3'->5' exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci U S A **103**, 5108–5113 (2006).
11. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol **6**, veaa061 (2020).
12. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. J Med Virol **92**, 455–459 (2020).
13. Giovanetti M, *et al.* Evolution patterns of SARS-CoV-2: Snapshot on its genome variants. Biochem Biophys Res Commun **538**, 88–91 (2021).
14. Roy U. Comparative structural analyses of selected spike protein-RBD mutations in SARS-CoV-2 lineages. Immunol Res, (2021).
15. Bhattacharya M, Chatterjee S, Sharma AR, Agoramoorthy G, Chakraborty C. D614G mutation and SARS-CoV-2: impact on S-protein structure, function, infectivity, and immunity. Appl Microbiol Biotechnol, (2021).
16. Kepler L, Hamins-Puertolas M, Rasmussen DA. Decomposing the sources of SARS-CoV-2 fitness variation in the United States. Virus Evol **7**, veab073 (2021).
17. Korber B, *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell **182**, 812-827 e819 (2020).
18. Volz E, *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature **593**, 266–269 (2021).
19. Washington NL, *et al.* Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. Cell **184**, 2587-2594 e2587 (2021).
20. Alpert T, *et al.* Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. Cell **184**, 2595-2604 e2513 (2021).
21. Truelove S, *et al.* Projected resurgence of COVID-19 in the United States in July-December 2021 resulting from the increased transmissibility of the Delta variant and faltering vaccination. medRxiv, (2021).
22. Centers for Disease Control and Prevention (CDC). COVID-19: SARS-CoV-2 variant classifications and definitions. Atlanta, GA: US Department of Health and Human Services. 2021. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html.) (2021).
23. England Public Health. SARS-CoV-2 variants of concern and variants under investigation in England—technical briefing 17. London, United Kingdom: Public Health England; 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/997418/Variants_of_Concern_VOC_Technical_Briefi (2021).
24. The 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).
25. Brito AF, *et al.* Global disparities in SARS-CoV-2 genomic surveillance. medRxiv, (2021).
26. Puerto Rico Department of Health. COVID-19 in Puerto Rico Dashboard. Available online: https://covid19datos.salud.gov.pr/ (Accessed on 1 November 2021).) (2021).
27. Government of Puerto Rico. Department of State. Executive Orders-15 March 2020. Available online: https://www.estado.pr.gov/en/executive-orders/.
28. Moreland A, *et al.* Timing of State and Territorial COVID-19 Stay-at-Home Orders and Changes in Population Movement - United States, March 1-May 31, 2020. MMWR Morb Mortal Wkly Rep **69**, 1198–1203 (2020).
29. Turakhia Y, *et al.* Stability of SARS-CoV-2 phylogenies. PLoS Genet **16**, e1009175 (2020).
30. Munis AM, Andersson M, Mobbs A, Hyde SC, Gill DR. Genomic diversity of SARS-CoV-2 in Oxford during United Kingdom's first national lockdown. Sci Rep **11**, 21484 (2021).
31. Happi AN, Ugwu CA, Happi CT. Tracking the emergence of new SARS-CoV-2 variants in South Africa. Nat Med **27**, 372–373 (2021).
32. Molina-Mora JA, *et al.* SARS-CoV-2 genomic surveillance in Costa Rica: Evidence of a divergent population and an increased detection of a spike T1117I mutation. Infect Genet Evol **92**, 104872 (2021).
33. Resende PC, *et al.* Evolutionary Dynamics and Dissemination Pattern of the SARS-CoV-2 Lineage B.1.1.33 During the Early Pandemic Phase in Brazil. Front Microbiol **11**, 615280 (2020).
34. Ferreira I, *et al.* SARS-CoV-2 B.1.617 Mutations L452R and E484Q Are Not Synergistic for Antibody Evasion. J Infect Dis **224**, 989–994 (2021).
35. Rodriguez-Grande C, *et al.* SARS-CoV-2 B.1.1.7 Decline Is Not Driven by the Introduction of a More Successful Variant. Microbiol Spectr, e0112821 (2021).

36. Charre C, *et al.* Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. Virus Evol **6**, veaa075 (2020).
37. Mavian C, *et al.* Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. Proc Natl Acad Sci U S A **117**, 12522–12523 (2020).
38. Kubik S, *et al.* Recommendations for accurate genotyping of SARS-CoV-2 using amplicon-based sequencing of clinical samples. *Clin Microbiol Infect* **27**, 1036 e1031-1036 e1038 (2021).
39. Santiago GA, *et al.* Tracing the Origin, Spread, and Molecular Evolution of Zika Virus in Puerto Rico, 2016-2017. Emerg Infect Dis **27**, 2971–2973 (2021).
40. Rambaut A, *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol **5**, 1403–1407 (2020).
41. O'Toole A, *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol **7**, veab064 (2021).
42. Vilsker M, *et al.* Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics **35**, 871–873 (2019).
43. Grubaugh ND, *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol **20**, 8 (2019).
44. Hadfield J, *et al.* Nextstrain: real-time tracking of pathogen evolution. Bioinformatics **34**, 4121–4123 (2018).
45. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**, 3059–3066 (2002).
46. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **32**, 268–274 (2015).
47. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol **4**, vex042 (2018).
48. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23**, 127–128 (2007).
49. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol **29**, 1969–1973 (2012).
50. Centers for Disease Control and Prevention (CDC). COVID Data Tracker. Atlanta, GA: US Department of Health and Human Services. 2021. Available online: https://covid.cdc.gov/covid-data-tracker/#variant-proportions January (2022).

Online Methods

Epidemiological data

We retrieved the number COVID-19 cases reported by PRDH from March 2020 to September 30th, 2021 from the PRDH database dashboard on December 1st, 2021 available here

https://covid19datos.salud.gov.pr/. The collection of cases includes cases classified as confirmed (by molecular tests) or probable (by antigen tests) and plotted by date of sample collection.

Patient sample selection

Nasopharyngeal swab samples pre-selected for genomic surveillance were received from COVID-19 passive surveillance conducted by the Puerto Rico Department of Health (PRDH), the Ponce Health Sciences University (PHSU) and hospital-based acute febrile illness surveillance conducted by the Centers for Disease Control and Prevention (CDC) Dengue Branch. A total of 785 samples were collected from March 2020 to September 30th, 2020 from the 7 health regions of the island, including 63 out of the 78 municipalities, and selection criteria included all samples with SARS-CoV-2 detected by reverse-transcriptase polymerase chain reaction (RT-PCR), viral load (CT<28) and sufficient residual sample volume stored at -80°C[39]. All samples were de-linked from patient identifiable information and processed under the guidelines approved by the CDC and Ponce School of Medicine institutional review boards (IRB) protocol 6731.

Lineage frequency analysis

Frequency of SARS-CoV-2 lineage detection in Puerto Rico was calculated using the total number of SARS-CoV-2 genomes published in the Global Initiative on Sharing All Influenza Data (GISAID) (https://www.gisaid.org) with collection dates ranging between March 1st, 2020 and September 30th, 2021. All complete genome sequences and metadata were retrieved from the GISAID database as of October 31st, 2021. The dataset was filtered for complete genome data, high-coverage data, and complete collection date for a final dataset of 2,514 entries. Lineage assignment on GISAID was determined by the Phylogenetic Assignment of Named Global Outbreak Lineages (Pangolin)[40, 41]. R with ggplot package was used to calculate lineage frequency and plot the graph focusing on the following lineages of interest: B.1.1.7 (Alpha), P.1+P.1.1 (Gamma), B.1.588, Delta (B.1.617.2+AY.x), B.1.427+B.1.429 (Epsilon), B.1.526 (Iota), B.1.621/1 (Mu), and all other Pangolin-designated B lineages grouped as Other. No genomes collected in May 2020 have been published in GISAID by October 31st, 2021.

Complete genome sequencing and assembly

Complete SARS-CoV-2 genome sequences were generated directly from clinical nasopharyngeal samples. Viral RNA was extracted from viral transport media using the automated MagNA Pure 96 system (Roche) with the MagNA Pure 96 DNA and Viral Nucleic Acid Small Volume Kit (Roche) following manufacturer-recommended protocols for 0.2mL sample input volume and 0.1mL RNA elution volume. MP96 external lysis buffer was used to pre-treat the samples for neutralization and assist the lysis process. First strand cDNA was synthesized with random hexamers using SuperScript IV reverse transcriptase

(ThermoFisher), and tiling PCR amplicons were generated using Q5® high-fidelity DNA polymerase (New England Biolabs) and the ARTIC nCoV-2019 V3 primer scheme purchased from Integrated DNA Technologies (https://github.com/artic-network/artic-ncov2019/blob/master/primer_scheme/nCoV-2019/V3/nCov-2019.tsv). Candidate samples for sequencing presented clearly visible bands of target size (approximately 400bp) in DNA gel electrophoresis for both primer pools. PCR products were purified with AMPure XP magnetic beads (Beckman Coulter) and quantified using Qubit 4.0 fluorometer (ThermoFisher). DNA libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) reducing all reagents volumes to 25% from the manufacturer recommended protocol to increase throughput. The resulting products were screened for size and quality using the Bioanalyzer 2100 instrument (Agilent Technologies) and quantified with Qubit 4 fluorometer (ThermoFisher). Passing libraries were pooled and run in the MiSeq sequencer instrument (Illumina) using the MiSeq Reagent Kit v3 in 600-cycle program.

The resulting sequence reads were screened for quality, trimmed and assembled into complete consensus SARS-CoV-2 genomes using the Genome Detective Virus Tool v1.136[42] (https://www.genomedetective.com) and assembly confirmed with iVar[43]. The Pangolin COVID-19 Lineage Assigner tool was used for lineage assignment[41] (https://pangolin.cog-uk.io). A total of 753 samples were sequenced with more than 95% genome coverage at a minimum of 10x sequence depth. All sequence data obtained for this study was submitted to GISAID, accession numbers available in Supplementary Table 1.

Phylogenetic analysis

Our Puerto Rico SARS-CoV-2 genomes dataset was analyzed against a diverse panel of genomes from across the world which provide regional phylogenetic context. Initially, we downloaded the Genomic Epidemiology metadata package for all entries from GISAID on August 18th, 2021 to screen genomes for subsampling. However, due to the large number of genomes available in GISAID, we downloaded and combined the following pre-sampled datasets for regional studies: NextRegion-North America, NextRegion-South America, and NextRegion-Global. We then used the standard ncov augur/auspice multiple input workflow available in the Nextstrain platform[44] (https://github.com/nextstrain/ncov) to subsample contextual genomes and phylogenetic inference with time-stamped trees. The custom subsampling scheme program selected 2,611 contextual genomes from the United States, North America, the Caribbean, Central America, South America, Africa, Europe, Asia, and Oceania, with higher proportions on The Americas based on collection dates and genetic proximity to our Puerto Rico dataset. The combined dataset of 3,364 genomes was aligned using MAFFT[45] and a global maximum likelihood (ML) phylogenetic inference was reconstructed with IQ-TREE[46]. The ncov workflow then transferred the ML tree to TreeTime[47] for time calibration and ancestral state reconstruction of the tree topology at constate rate of $8x10^{-4}$ nucleotide substitutions per site per year. The resulting global ML tree was visualized with Nextstrain auspice[44] and annotated with iTol for region of origin and emerging variants[48]. Subsampling from the Genomic Epidemiology metadata package retrieved in August and from the concatenated NextRegions produced phylogenetic inference trees with similar topologies. A list of all the sequences used in this study, including sequence labels and authors can be found in Supplementary Table 2.

Significant lineages of interest were studied further by reconstruction of phylogenetic inference sub-trees. For the B.1.588 lineage sub-tree, we selected all B.1.588 genomes published in GISAID by October 31st, 2021. Contextual B.1 lineage genomes were selected based on phylogenetic clustering near the base of the B.1.588 clade in the global tree and by temporal proximity to the date range of B.1.588 circulation between June 2020 – January 2021. Maximum likelihood phylogenetic trees were reconstructed with the resulting dataset of 239 genomes under the GTR+G+I nucleotide substitution model and 1,000 bootstrap replicates using IQ-TREE v1.6.12[46]. The resulting tree topology and node support were compared to Bayesian maximum clade credibility (MCC) tree reconstruction using BEAST v1.10.4[49]. Briefly, we used time-calibrated genomes with sample collection dates and the nucleotide substitution model parametrized using Yang96 model under strict molecular clock, and Bayesian Skyline coalescent model. Markov chains were run for a total of 100 million steps with sampling every 10,000 steps in the chain. Run results were evaluated in Tracer (http://tree.bio.ed.ac.uk/software/tracer/) to ensure stationary parameters with statistical errors reflected in 95% highest probability density ranges with effective sample size (ESS) higher than 200 for each tree prior. MCC trees were generated in TreeAnnotator from BEAST package after discarding 10% of runs as burn-in. The resulting ML and MCC trees were visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

For the B.1.1.7 (Alpha VOC) lineage sub-tree, we selected all B.1.1.7 + Q.x designated genomes from the global tree and supplemented the dataset with additional B.1.1.7 + Q.x genomes from Puerto Rico and the United States to understand the lineage emergence and spread in the island. A custom subsampling scheme was selected on the Nextstrain ncov workflow to select genomes from samples collected between November 1st, 2020 and February 28th, 2021. Maximum likelihood phylogenetic trees were reconstructed with the resulting dataset of 729 genomes under the GTR+G+I nucleotide substitution model and 1,000 bootstrap replicates using IQ-TREE v1.6.12[46]. The resulting ML tree was visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

For the B.1.617.2 (Delta VOC) lineage sub-tree, we selected all B.1.617.2 + AY.x designated genomes from the global tree and supplemented the dataset with additional B.1.617.2 + AY.x genomes from Puerto Rico to understand lineage spread and sub-lineage clustering patterns across the island. The genome selector Python script designed by Anderson Brito (https://github.com/andersonbrito/ncov) was used to select additional Delta-designated genomes from Puerto Rico with collection dates range between June 1st, 2021 to September 30th, 2021. Maximum likelihood phylogenetic trees were reconstructed with the resulting dataset of 815 genomes under the GTR+G+I nucleotide substitution model and 1,000 bootstrap replicates using IQ-TREE v1.6.12[46]. The resulting ML tree was visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree).

The date of the most recent common ancestor (tMRCA) determined by the Nextstrain ML phylogenetic inference was confirmed with Bayesian coalescent analyses for B.1.588, Alpha VOC and Delta VOC lineage sub-trees. Due to the large size of the datasets, each sub-tree was reduced by tree-pruning to datasets with less than 150 genomes. tMRCA analyses were performed with BEAST v1.10.4 under a strict molecular clock fixed at $8x10^{-4}$ substitutions per site per year and 150 million Markov chains sampling every 10,000 steps. Results were evaluated in Tracer (http://tree.bio.ed.ac.uk/software/tracer/) for convergence and ESS greater than 200.

Figures

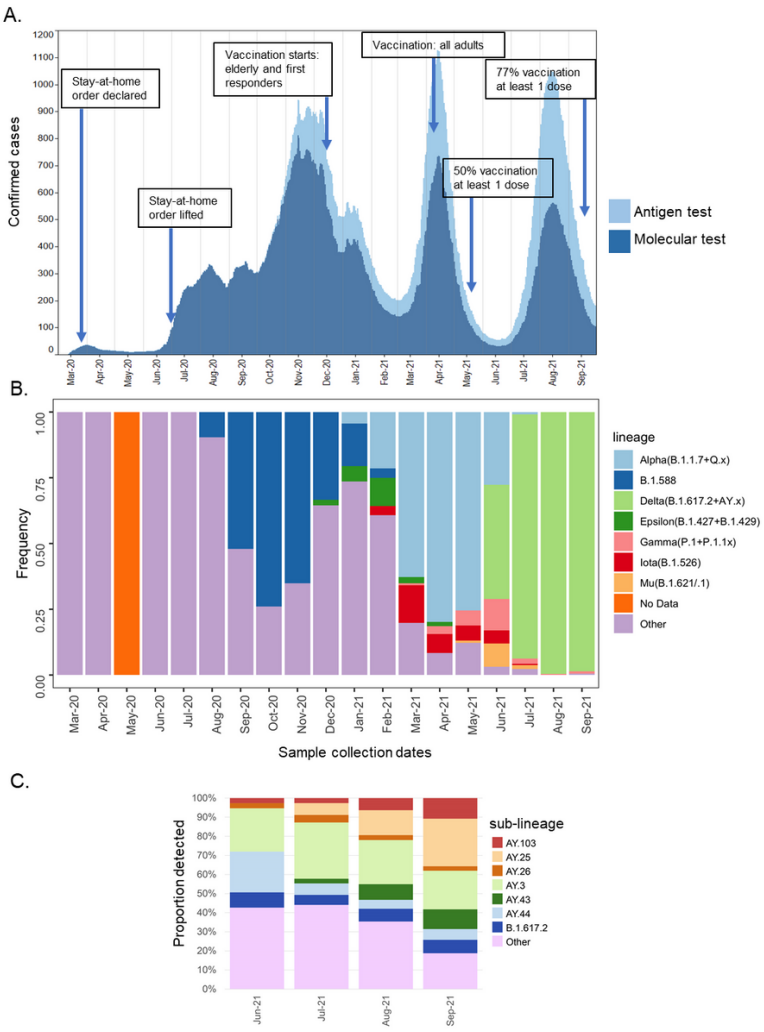

Figure 1

Epidemiology and dynamics of SARS-CoV-2 lineage turn-over in Puerto Rico. (A) Graphical representation of the number of daily SARS-CoV-2 cases confirmed by antigen tests (light blue) and molecular tests (dark blue) reported by the PRDH from March 2020 to September 30th, 2021, shown as a 21-day rolling average. Arrows indicate the timeline of government responses and vaccination milestones. (B) Proportion of all SARS-CoV-2 lineages and emerging variants detected in Puerto Rico and published in GISAID from March 2020 to September 30th, 2021 (n=2,514 sequences after filtering for high-coverage genomes with complete sampling dates). Non-VBM/VOC lineages are categorized as a collective labeled “Other”, except for local lineage B.1.588 due to high frequency. No genomes were obtained during May 2020. (C) Proportion of all Delta sub-lineages detected in Puerto Rico published in GISAID until September 30th, 2021 (n=1,360 genomes). Sub-lineages with more than 5% detection are represented by individual color, whereas sub-lineages with less than 5% detection are categorized as a collective labeled “Other”.

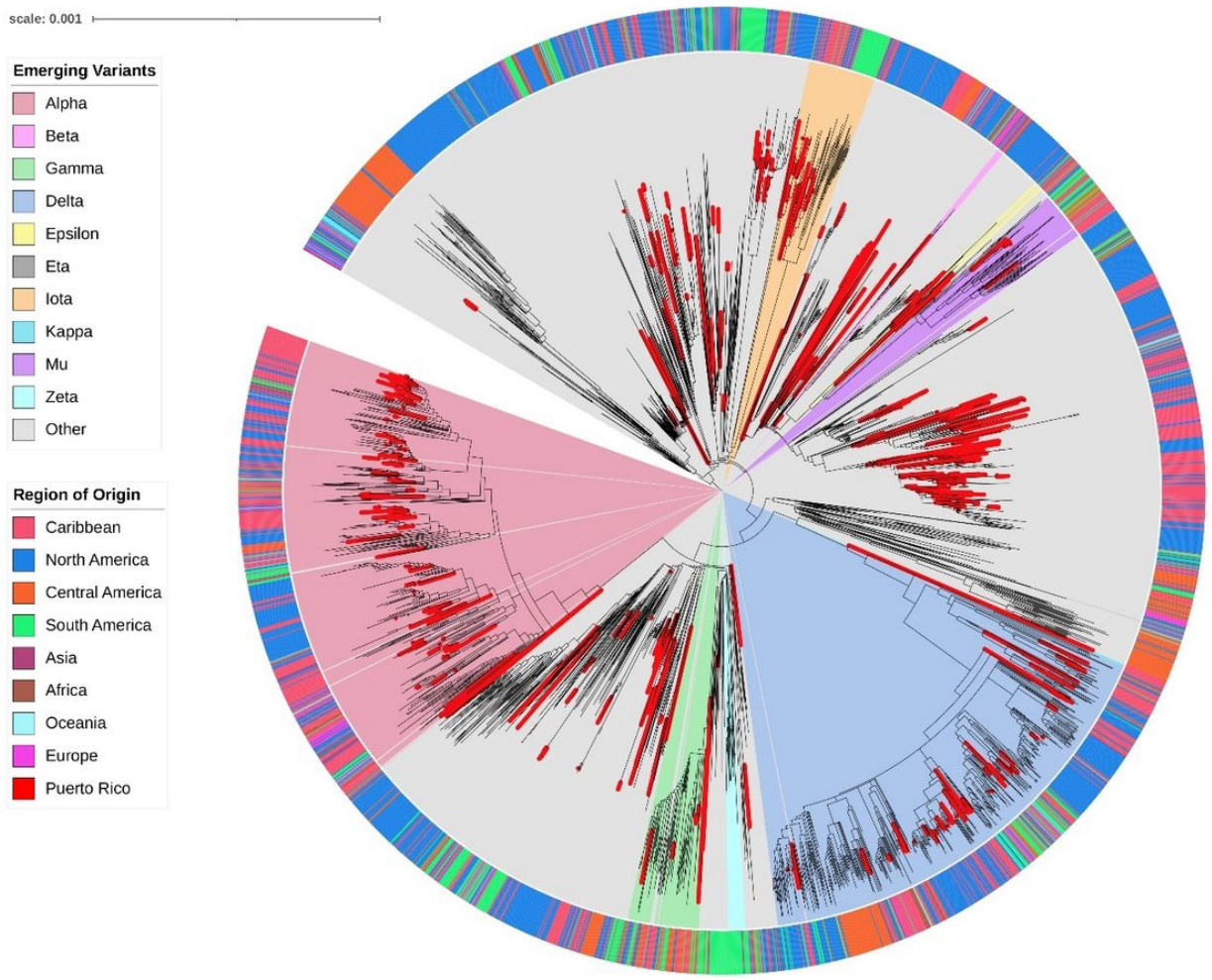

Figure 2

Phylogenetic reconstruction of local SARS-CoV-2 epidemic in Puerto Rico in a global context. Maximum likelihood tree inferred with 3,364 complete genomes including 753 viral genomes from Puerto Rico sampled between March 23rd, 2020 and September 30th, 2021 (red branches) combined with 2,611 complete genomes retrieved from GISAID during the same period to provide a global backdrop with a higher focus on the Americas region. Node structure is supported by 1,000 bootstrap replicates. Branches marked in red represent taxa from Puerto Rico. The outer ring is color-coded by region of origin. The inner wedges are color-coded to represent emerging variants of interest or concern. The phylogenetic tree is rooted in Wuhan/WH01/2019 and Wuhan/Hu-1/2019 reference genomes.

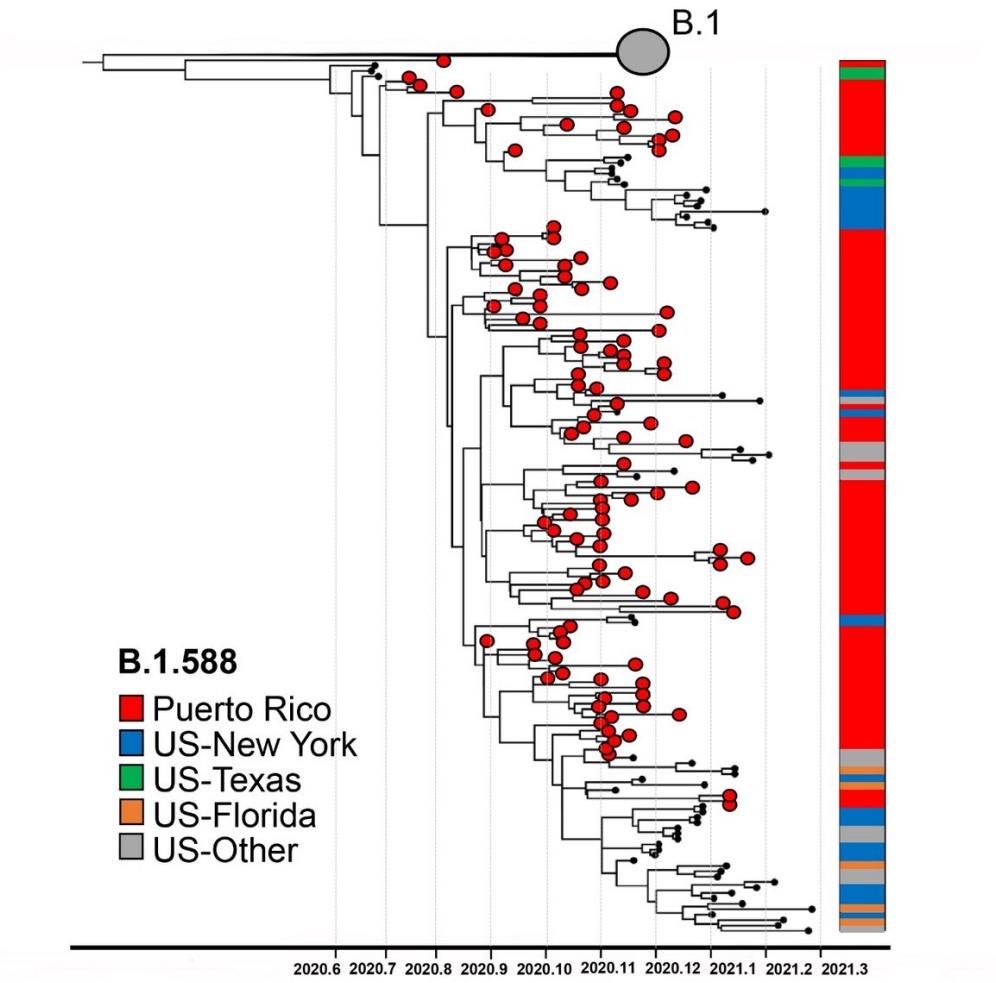

Figure 3

Emergence of autochthonous lineage B.1.588. Phylogenetic reconstruction of monophyletic lineage B.1.588 using Bayesian maximum clade credibility tree inferred with 239 complete genomes including 130 genomes from Puerto Rico (103 B.1.588 genomes) sampled between July 2020 and March 2021. Node support was tested by posterior probability. The gray circle represents B.1 viral genomes from Puerto Rico and the United States that cluster basal to the B.1.588 monophyletic lineage. Red circle taxa tips represent viral genomes from Puerto Rico. Colored shade bar on the right of the tree indicates the taxa region of origin. The phylogenetic tree is rooted in Wuhan/WH01/2019 and Wuhan/Hu-1/2019 reference genomes.

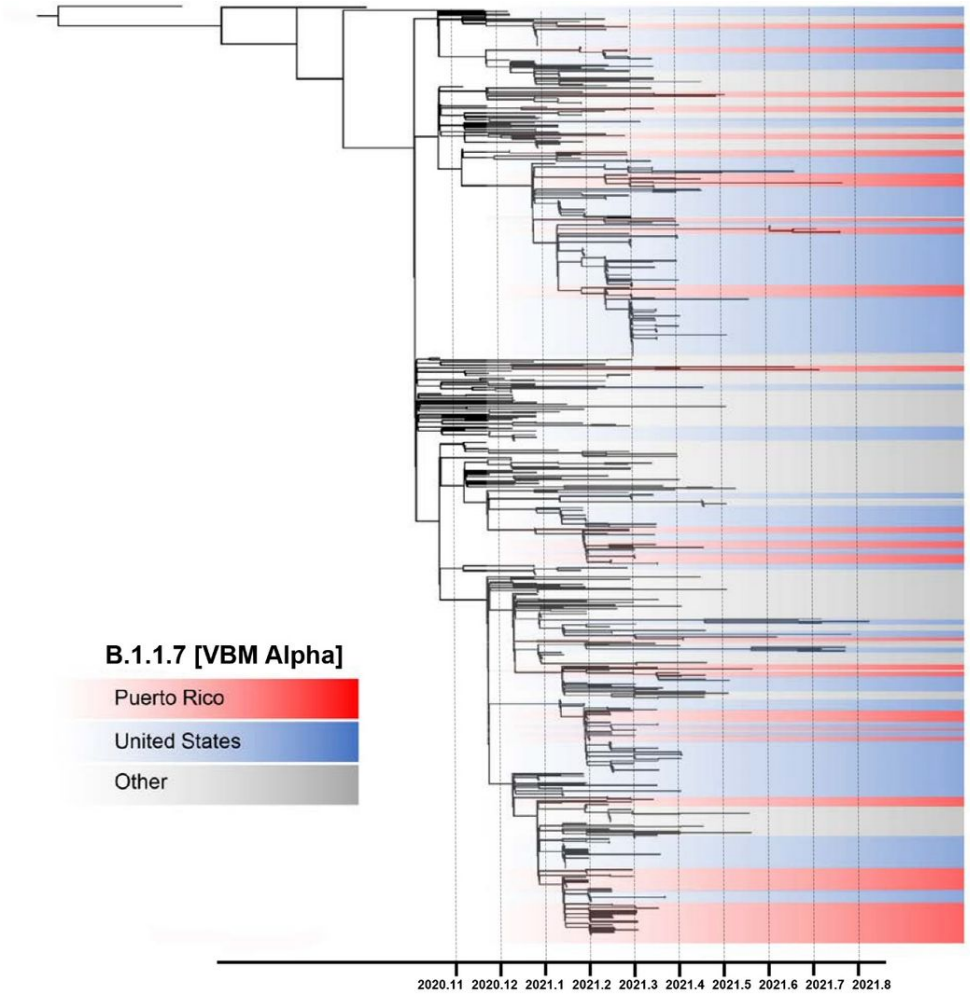

Figure 4

Emergence and spread of VBM Alpha in Puerto Rico driven by multiple introductions. Phylogenetic reconstruction using a maximum likelihood tree inferred with 730 time-calibrated complete genomes, including 160 viral genomes from Puerto Rico and 570 contextual viral genomes from the United States and the Americas to provide a regional backdrop. Node structure supported by 1,000 bootstrap replicates. Tree topology shaded in red represents clusters of viral genomes from Puerto Rico, blue shades represent clusters of genomes from the United States, and the gray shades represent clusters from other countries.

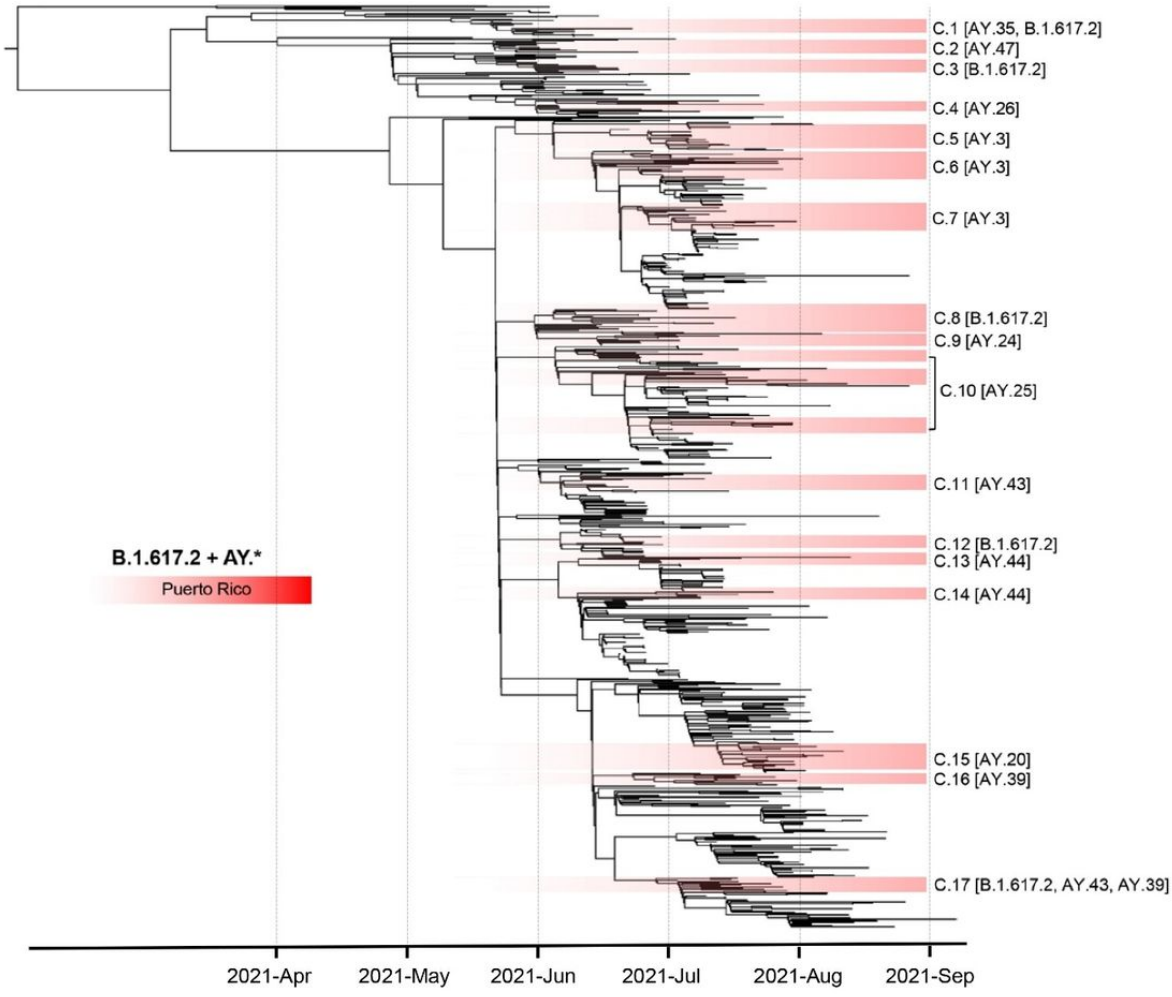

Figure 5

Emergence and spread of VOC Delta in Puerto Rico driven by introduction of multiple sub-lineages. Phylogenetic reconstruction using a maximum likelihood tree inferred with 815 time-calibrated complete genomes including 324 viral genomes from Puerto Rico and 491 contextual genomes from the United States and the Americas to provide a regional backdrop. Node structure supported by 1,000 bootstrap replicates. Tree topology sections shaded in red represent clusters of viral genomes from Puerto Rico. Each cluster from Puerto Rico is labeled with cluster number C.x and Delta sub-lineage PANGO assignment [AY.x].

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.docx
- SupplementaryTable2.docx