VIRUS EVOLUTION

*Virus Evolution*, 2022, **8(1)**, 1–13

DOI: https://doi.org/10.1093/ve/veac011

Advance access publication date: 15 February 2022

**Research Article**

# Unrecognized introductions of SARS-CoV-2 into the US state of Georgia shaped the early epidemic

Ahmed Babiker,[1,2,†,‡] Michael A. Martin,[3,4,†,§] Charles Marvil,[2] Stephanie Bellman,[5] Robert A. Petit III,[1] Heath L. Bradley,[2] Victoria D. Stittleburg,[1] Jessica Ingersoll,[2] Colleen S. Kraft,[1,2] Yan Li,[6] Jing Zhang,[6] Clinton R. Paden,[6] Timothy D. Read,[1] Jesse J. Waggoner,[1] Katia Koelle,[3,††] and Anne Piantadosi[1,2,*,‡‡]

[1]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, 201 Dowman Drive, Atlanta, GA 30322, USA, [2]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, 201 Dowman Drive, Atlanta, GA 30322, USA, [3]Department of Biology, Emory University, 201 Dowman Drive, Atlanta, GA 30322, USA, [4]Population Biology, Ecology, and Evolution Graduate Program, Laney Graduate School, Emory University, 201 Dowman Drive, Atlanta, GA 30322, USA, [5]Environmental Health Sciences PhD Program, Laney Graduate School, Emory University, 201 Dowman Drive, Atlanta, GA 30322, USA and [6]Division of Viral Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, USA

[†]These authors contributed equally.

[‡]https://orcid.org/0000-0003-0578-4871

[§]https://orcid.org/0000-0002-9689-4066

[††]https://orcid.org/0000-0002-0254-6141

[‡‡]https://orcid.org/0000-0002-5942-1534

*Corresponding author: E-mail: anne.piantadosi@emory.edu

## Abstract

In early 2020, as diagnostic and surveillance responses for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) ramped up, attention focused primarily on returning international travelers. Here, we build on existing studies characterizing early patterns of SARS-CoV-2 spread within the USA by analyzing detailed clinical, molecular, and viral genomic data from the state of Georgia through March 2020. We find evidence for multiple early introductions into Georgia, despite relatively sparse sampling. Most sampled sequences likely stemmed from a single or small number of introductions from Asia three weeks prior to the state's first detected infection. Our analysis of sequences from domestic travelers demonstrates widespread circulation of closely related viruses in multiple US states by the end of March 2020. Our findings indicate that the exclusive focus on identifying SARS-CoV-2 in returning international travelers early in the pandemic may have led to a failure to recognize locally circulating infections for several weeks and point toward a critical need for implementing rapid, broadly targeted surveillance efforts for future pandemics.

**Key words:** SARS-CoV-2; introductions; travel; Georgia.

## 1. Introduction

Phylogenetic studies have been critical to investigating the introduction and spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) throughout the USA and globally. Understanding the source of viral introductions and the subsequent dynamics of viral spread is essential for evaluating the efficacy of public health interventions and informing the response to future outbreaks. For example, a phylogenetic analysis indicated that the first identified case of SARS-CoV-2 in the USA, in mid-January 2020, did not directly lead to the initial wave of infections in Washington State; instead, its transmission was stopped by public health interventions. By contrast, undetected introductions into Washington State, likely in early February, sparked significant downstream transmission, despite federal policies to limit travel from China beginning on 2nd February (Worobey et al. 2020).

On a broader scale, relatively uninterrupted travel in early 2020 allowed multiple introductions of SARS-CoV-2 into specific regions of the USA. For example, in New York City, the epicenter of the US outbreak in Spring 2020, multiple undetected introductions of viral lineages, likely from Europe, sparked local transmission chains (Gonzalez-Reiche et al. 2020). These undetected introductions into the USA are thought to have resulted in a significant level of unobserved infection in early 2020 (Perkins et al. 2020).

Few studies have attempted to characterize early patterns of SARS-CoV-2 introduction and circulation in the southeastern USA, and none to date have focused on the state of Georgia, a major national and international travel hub due to Hartsfield-Jackson Atlanta International Airport. The first reported SARS-CoV-2 case in Georgia was on 2 March 2020 in Fulton County (Georgia Department of Public Health 2020) and reported cases rose slowly throughout the month, topping 100 per day for the first time on 20 March 2020. By the end of March, a total of 3,929 cases had been reported in the state (Dong, Du, and Gardner 2020).

Reported cases are a function of both underlying epidemiological dynamics and detection by the public health system: healthcare-seeking behavior, frequency of diagnostic testing, and completeness of reporting to public health. From January through March 2020, there were rapid shifts in the availability of and recommendations for SARS-CoV-2 diagnostic testing in Georgia and the USA as a whole. Due to limited availability of SARS-CoV-2 diagnostic tests and the assumption that viral transmission was largely restricted to China, testing was initially limited to individuals with recent travel history to mainland China or those who had contact with a known traveler or a diagnosed case of SARS-CoV-2 (Patel et al. 2020). As large outbreaks were identified outside of China and testing through clinical laboratories became possible, testing was expanded to include high-risk individuals with compatible illness and potential community exposure (Health Alert Network 2020; Schuchat 2020). Reflecting these national trends, SARS-CoV-2 testing for patients within the Emory Healthcare (EHC) system prior to 15 March 2020 required physician request, public health agency approval, and testing via the Georgia Department of Public Health (GADPH) or the Centers for Disease Control and Prevention (CDC) (Babiker et al. 2020b). Only 176 patients were tested for SARS-CoV-2 in the EHC system between 26 January 2020 and 16 March 2020. On 29 February 2020, guidelines from the Food and Drug Administration (CDC 2020) allowed certified labs to validate testing for SARS-CoV-2, and testing volumes nationwide increased considerably. On 15 March 2020, local testing at EHC began, and in the second half of March 2020 over 2,700 tests were performed at the Emory University Hospital Molecular and Microbiology Laboratories.

Changing test volumes can obfuscate underlying epidemiological dynamics, and case data alone cannot be used to evaluate the relative importance of viral introductions versus local transmission in sustaining viral spread within a region. To better understand the early epidemic in Georgia, we analyzed SARS-CoV-2 whole genome sequences sampled in Georgia from 29 February 2020 (the first available sequence) through 31 March 2020. We assessed the changing frequencies of viral clades and, by incorporating globally sampled sequences, estimated the number and timing of viral introductions into the state. Where available, we interrogated travel history to identify the contribution of international and domestic travel to SARS-CoV-2 spread within Georgia. Finally, we combined sequence data with detailed clinical metadata to evaluate associations between viral genotype and clinical parameters. These results add to the growing body of work characterizing the spread of SARS-CoV-2 into and within the USA and provide insight into early events that shaped the outbreak in Georgia.

## 2. Results

### 2.1 One Hundred Eight (108) SARS-CoV-2 genomes from the first month of the pandemic in Georgia were sequenced

To understand the diversity and spread of SARS-CoV-2 in Georgia during early 2020, we sequenced forty-seven complete SARS-CoV-2 genomes from patients seen within the EHC system through 31 March 2020 (Supplementary Tables S1 and S2, Supplementary Figures S1 and S2) and combined them with the sixty-one publicly available SARS-CoV-2 sequences generated by other groups within this time frame (Supplementary Tables S3–S5). These 108 sequences represent 2.7 per cent of the 3,929 reported cases in Georgia through 31 March 2020 (Fig. 1A). They include two specimens from 29 February 2020, before the first officially reported case, and at least ten samples per week throughout the month of March, with the exception of the week ending 29 March 2020. Thus, this dataset provides a temporally comprehensive sampling of the circulating viruses within the state at the time. County-level sampling location data were available for fifty-six sequences (Fig. 1B, C), which were largely sampled from the Atlanta metro area, the most densely populated region of the state, in which 46 per cent of the reported cases in this period occurred. Another significant portion of the reported cases occurred in Dougherty County and were associated with a funeral (Willis 2020). Sequences from this outbreak are not known to be included in our analysis; however, half of the included sequences did not have available county-level data. Nine sampled individuals are known to have traveled within two weeks prior to symptom onset (Supplementary Table S2).

### 2.2 Four major SARS-CoV-2 clades were present in the state of Georgia during early 2020

To assess the genetic diversity of the SARS-CoV-2 sequences circulating within Georgia during early 2020, we assigned each of them to a phylogenetic clade (Bedford, Hodcroft, and Neher 2020) (Fig. 2A). Among these sequences, the first identified clade was 20B, which was observed in two sequences sampled on 29 February 2020. Sequences in this clade harbor the canonical substitutions C14408T, A23403G (responsible for the widely reported D614G amino acid substitution in the spike protein (Volz et al. 2020)), G28881A, and G28882A relative to Wuhan/Hu-1 (EPI_ISL_402125 (Wu et al. 2020)). Clade 20B was prominent throughout Europe (Alm et al. 2020) and a number of US states (Zeller et al. 2021) throughout early 2020. Despite being the first identified clade in Georgia, local transmission of 20B appears to have been limited, and it was only sporadically ($N = 6/108$) identified throughout March 2020.

By contrast, clade 19B, a more ancestral clade, rapidly became dominant in Georgia throughout the spring of 2020. Sequences in this clade harbor the canonical substitutions 8782T and 28144C relative to Wuhan/Hu-1. This clade was first identified in Georgia on 3 March 2020, and nearly three-quarters ($N = 77/108$) of the analyzed Georgia sequences fell within clade 19B.

The remaining twenty-three available Georgia sequences from March 2020 were assigned to clades 20A ($N = 7/108$) and 20C ($N = 16/108$). Given the genetic diversity delimiting these clades and the global diversity of the clades at the time, these findings imply that there were multiple introductions of SARS-CoV-2 into GA, likely from multiple global sources. The temporal distribution of Pango lineages (Rambaut et al. 2020) mirrored these clade distributions (Supplementary Table S1, Supplementary Figure S3).

### 2.3 Multiple SARS-CoV-2 introductions into Georgia occurred by the end of March 2020

We reconstructed a maximum likelihood phylogenetic tree containing the 108 Georgia sequences, along with 4,514 global sequences, which were downsampled from all available sequences over the same time period to be geographically representative of case counts and to maximize phylogenetic resolution around the Georgia sequences (weighted downsampling strategy, Section 4). Sampling dates were used to exclude eleven sequences (one from Georgia) that did not follow the expected molecular clock and to estimate the dates of internal nodes based on a time-resolved tree. A significant temporal signal in these data was confirmed using root-to-tip regression (Supplementary Figure S4). The sequences from Georgia were distributed heterogeneously throughout the tree (Fig. 2B, Supplementary Figure S5,

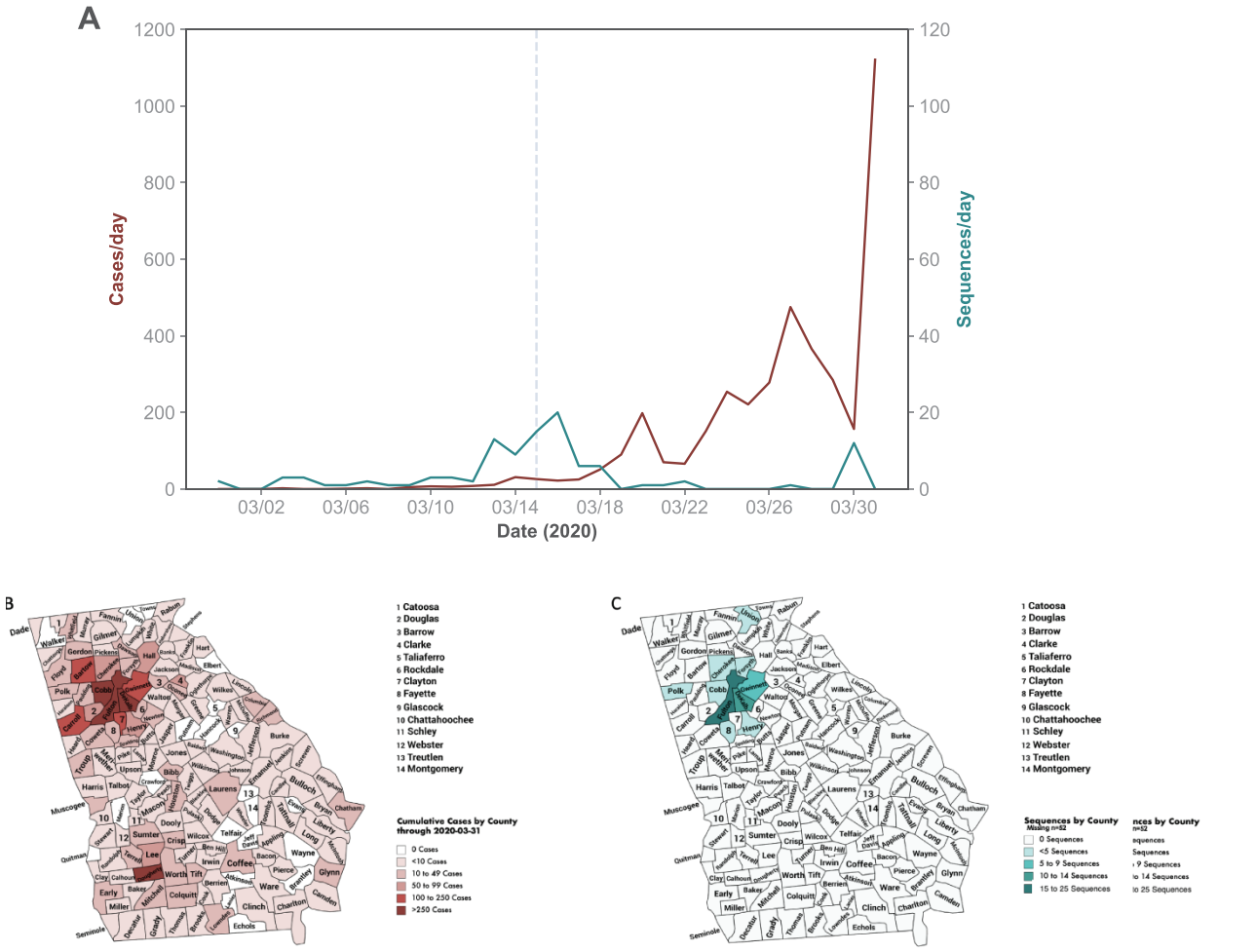


**Figure 1.** Temporal and spatial distribution of SARS-CoV-2 cases and sequences in the state of Georgia. (A) Daily numbers of reported cases within the state of Georgia (red) and daily number of available sequences (GISAID). The dashed line indicates 15 March 2020, the date that EHC received Emergency Use Authorization for diagnostic testing. (B) Cumulative number of reported cases as of 31 March 2020 by county. (C) County of residence for the patients from which viral sequences were sampled, where available ($N = 56$). County-level location data are unavailable for the remaining sequences. The Atlanta metro region comprises 10 counties within the Atlanta Regional Commission: Cherokee, Clayton, Cobb, DeKalb, Douglas, Fayette, Fulton, Gwinnett, Henry, and Rockdale.

Supplementary Table S5). The majority of Georgia sequences ($N = 69$) were closely related and clustered together within clade 19B (Pango lineage A), while the rest either did not cluster together or descended from highly polytomous nodes along with many other sequences.

To quantify the number of introductions into Georgia represented by this dataset, we inferred the discrete location of internal nodes using maximum likelihood ancestral state reconstruction. As undersampling of Georgia sequences can only bias the number of introductions downwards, this represents the lower limit of the number of true introductions through 31 March 2020. We conservatively estimate that there were at least 19 [95 per cent CI 17–21] introductions into Georgia in this time range (Fig. 2C). The earliest was estimated to have occurred in early to mid-February 2020 and gave rise to the sixty-nine closely related 19B sequences. Most introductions occurred in late February or early March 2020 and appear as singletons or doubletons in the tree. Highly consistent results were obtained using an alternative downsampling strategy designed to be temporally and geographically homogeneous (up to twenty sequences per country per week) (Supplement, Supplementary Figure S6).

## 2.4 The earliest lineage in Georgia was most likely introduced directly from Asia several weeks prior to SARS-CoV-2 detection in the state

To provide a more robust analysis of the evolutionary history of the sixty-nine closely related Georgia 19B sequences, we employed Bayesian phylogenetic reconstruction, which simultaneously estimates tree structure and discrete states of internal nodes and provides a posterior distribution of possible reconstructions. To provide context for the ancestral origins of the 19B subclade, we identified the shared mutations between it and its closest relatives (Fig. 2B, clade marked with red '+'): T26729C and G28077C (which are subclade-defining) and T28144C (which is clade 19B-defining). We then selected all available high-quality sequences that contained these mutations, which included 67 sequences from Georgia (two were removed due to the presence of ambiguous nucleotides at clade defining genome positions), 370 from other US states, and 91 from other countries (Supplementary Table S6). One sequence was excluded as it did not follow the expected molecular clock. A significant temporal signal in the remaining data was confirmed using root-to-tip regression (Supplementary Figure S7). For computational efficiency we excluded five US sequences

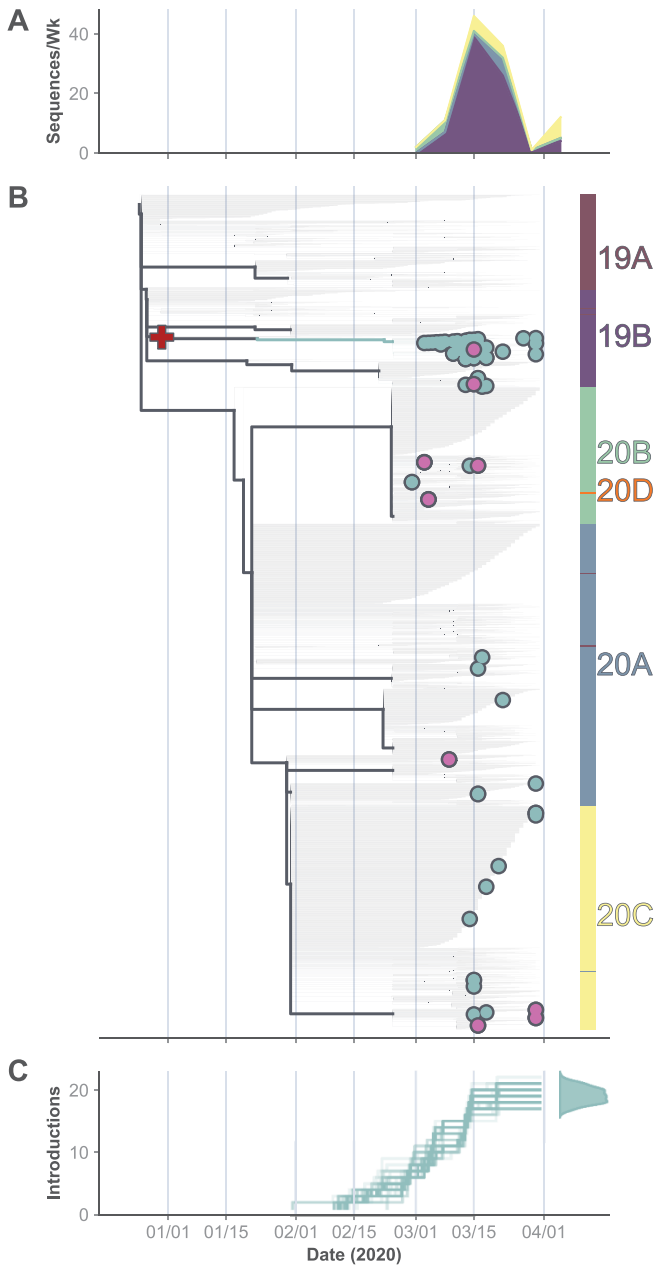


**Figure 2.** Presence of multiple clades and maximum likelihood phylogenetic analysis indicate multiple introductions of SARS-CoV-2 into Georgia. (A) Number of sequences from Georgia per clade, per week included in the phylogenetic analysis. (B) Time-resolved maximum likelihood tree of 4,611 globally sampled sequences rooted at Wuhan/Hu-1, downsampled based on the cumulative number of cases in a given country as of 31 March 2020 and genetic distance to Georgia sequences (weighted downsampling strategy). Internal nodes are colored based on their estimated location either inside (green) or outside (gray) of Georgia. Georgia tips are colored in green except for those with known travel history, which are shown in pink. The color bar at right shows the clade identity of each sequence in the tree. Branch widths are weighted for visual clarity. Red + indicates the phylogenetic clade used to select sequences for downstream Bayesian phylogenetic analyses. (C) Estimated cumulative number of introductions into Georgia (transition from a non-Georgia node to a Georgia-node/tip) based on the ancestral state reconstruction of internal nodes. Estimation was repeated on 100 bootstrap replicate trees and the timing of introduction events for each replicate is shown as an individual line. The Gaussian kernel density plot at right shows the estimated cumulative number of introduction events as of 31 March 2020.

without state metadata, eighteen sequences from US states with fewer than four sequences in this subclade and seventy-two international sequences sampled after 29 February 2020. The excluded international sequences either did not cluster with the US sequences ($N = 1$, sampled from China) or were evolutionarily descendant from the US sequences and are thus likely exportations from the USA to these international regions (Supplementary Figures S8 and S9). Therefore, 432 sequences were included in our Bayesian phylogenetic analysis including the 67 from Georgia, 346 from other US states, 12 from China (including Hong Kong), 5 from South Korea, and 2 from Vietnam (Supplementary Table S6).

Our analysis revealed that the US sequences in the 19B subclade were phylogenetically distinct from the ancestral sequences, consistent with a single or small number of introductions of this lineage into the USA (Fig. 3, Supplementary Table S7). To evaluate the source of introduction, we inferred the location of the most recent common ancestor (MRCA) of all US sequences in the 19B subclade. The MRCA was assigned to Georgia in 65 per cent of sampled trees (posterior probability of 0.65). The next most likely discrete state of the MRCA node was South Carolina, with posterior probability of 0.13. Thus, Georgia is the most likely site of introduction of the 19B subclade, with the important caveat that undersampling at the beginning of the epidemic makes it difficult to draw firm conclusions. Although we maximized our power to detect multiple introductions of this subclade into the state of Georgia by including nearly all available phylogenetically related domestic sequences, it is possible that the 19B subclade was originally introduced into a state with minimal genomic surveillance. Furthermore, it is possible that over-sampling of sequences within Georgia relative to other regions biased these results. To assess the impact of this possibility, we performed Bayesian phylogenetic (BEAST) analyses in which sequences were downsampled in a temporally and geographically homogeneous manner (Section 4). In each of these downsampled replicates, the MRCA of all US sequences was assigned to Georgia with the highest probability (Supplementary Figure S10A, Supplementary Table S7), consistent with our non-downsampled analysis. The next most probable ancestral state in these downsampled analyses was South Carolina, which indicates that if this clade was not first introduced into Georgia, it was likely introduced elsewhere in the southeastern USA. While the distribution of estimated number of introductions into Georgia in the downsampled alignments was slightly higher than the full alignment (Supplementary Figure S10B), these analyses consistently support a limited number of introductions into Georgia. Much of the genetic diversity of non-Georgia sequences appears nested within the genetic diversity of sequences from Georgia (Fig. 3, Supplementary Figure S8), consistent with one, or a small number of, introduction(s) into Georgia.

Importantly, there was a gap of approximately three weeks between the estimated time to most common ancestor (tMRCA) of the US sequences in this analysis (8 February 2020 [1 February 2020, 14 February 2020]) and the earliest sampled US sequence (1 March 2020), highlighting a relative lack of dense sampling of SARS-CoV-2 genomes throughout the USA during the spring of 2020. Although the earliest US sequence in this analysis was sampled in Mississippi on 1 March 2020 (EPI_ISL_648018), it had one additional single-nucleotide polymorphism (SNP) (G922A) relative to Wuhan/Hu-1 compared to the earliest Georgia sequence (EPI_ISL_420786, sampled 3 March 2020), supporting the

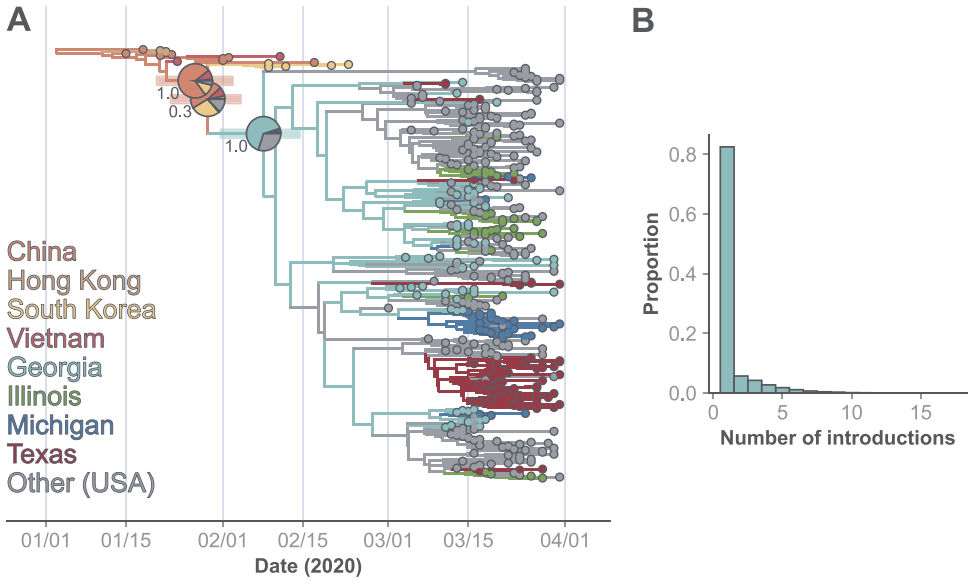


**Figure 3.** Bayesian phylogenetic analysis of genetically related Georgia 19B sequences and their phylogenetic neighbors reveals undetected circulation in February 2020. (A) Maximum clade credibility tree (median node heights) of select 19B sequences. Tips are colored by their state (USA) or country (intl.) of origin. Less abundant states are colored as 'Other (USA)' for visualization purposes only. Internal nodes are colored by their most probable location based on the set of estimated trees and ancestral state reconstruction. Select nodes annotated with their 95 per cent highest posterior density of estimated date (horizontal bar), location probabilities (pie chart), and posterior support (text). Negative branch lengths are manually set to 0 for visualization purposes. (B) Estimated number of introductions of the 19B subclade shown in (A) into the state of Georgia for each sampled tree in the Bayesian phylogenetic reconstruction.

hypothesis that this 19B subclade was likely first introduced into Georgia. As the first case in the state of Georgia was not reported until 2 March 2020, this analysis indicates that SARS-CoV-2 was likely spreading within the state for approximately three weeks prior to detection in either diagnostic or sequencing data.

While the source of the 19B subclade introduction was ambiguous in our phylogenetic reconstruction (posterior probabilities for China: 0.47, South Korea: 0.28, Vietnam: 0.09), the branching structure of this subclade relative to related sequences from China and South Korea was well-resolved by the data with posterior probability of 1. Overall, these results indicate that this subclade was most likely introduced from Asia in late January or early February and spread undetected throughout the USA for three to four weeks.

## 2.5 Analysis of sequence metadata identified a small number of travel-associated introductions

To assess the contribution of domestic and international travel to the introduction of SARS-CoV-2 into Georgia, we leveraged the extensive clinical and epidemiological data available for EHC patients in this study. Clinical data were available for forty-six of the forty-seven EHC patients from whom complete SARS-CoV-2 sequences were obtained, as well as an additional 8 patients without complete SARS-CoV-2 sequences (Table 1, Supplementary Table S2). Twenty-five of these 54 patients (46 per cent) were female and 29 (54 per cent) were male, and ages ranged from 21 to 92 years. Thirty (56 per cent) of these patients were African American, a larger proportion than the demographics of Georgia in general (United States Census Bureau 2019), likely owing to both the disproportionate representation of the Atlanta metro in these data and the disproportionate impact of the SARS-CoV-2 pandemic on people of color (Subbaraman 2020). The duration of symptoms prior to sample collection ranged from 1 to 28 days (median 6 days). Clinical severity ranged from mild (outpatient or ED visit only, $N = 19$) to moderate (inpatient without intensive care unit (ICU) admission, $N = 23$) to severe (inpatient with ICU admission, $N = 12$), and four patients died. Although we did not specifically select samples from returning travelers, we found that nine patients (17 per cent) had traveled outside of Georgia within the two weeks prior to diagnosis. Four had traveled internationally, including three of the four patients with the earliest

**Table 1.** Demographic and clinical data from fifty-four EHC patients. One patient with no available data was excluded.

| | N (%) |
|---|---|
| Age (mean [standard deviation]) | 53.0 [17.1] |
| Female sex | 25 (46.3) |
| Race | |
| African American | 30 (55.6) |
| Asian | 4 (7.4) |
| White | 17 (31.5) |
| Hispanic/Latino | 3 (5.6) |
| Travel in preceding two weeks | 9 (16.6) |
| Diabetes mellitus | 13 (24.1) |
| Hypertension | 25 (46.3) |
| Obesity | 25 (46.3) |
| Lung disease | 9 (16.7) |
| Immunosuppression | 14 (25.9) |
| Days from symptom onset to sample collection (median [aq1–aq3]) | 5.5 [4,8] |
| SARS-CoV-2 disease severity[a] | |
| Mild | 19 (35.2) |
| Moderate | 23 (42.6) |
| Severe | 12 (22.2) |
| In-hospital death | 4 (7.4) |

[a]SARS-CoV-2 disease severity was classified as severe if the patient was admitted to an ICU, moderate if the patient was hospitalized without ICU admission, and mild if the patient had an outpatient or emergency department visit only.

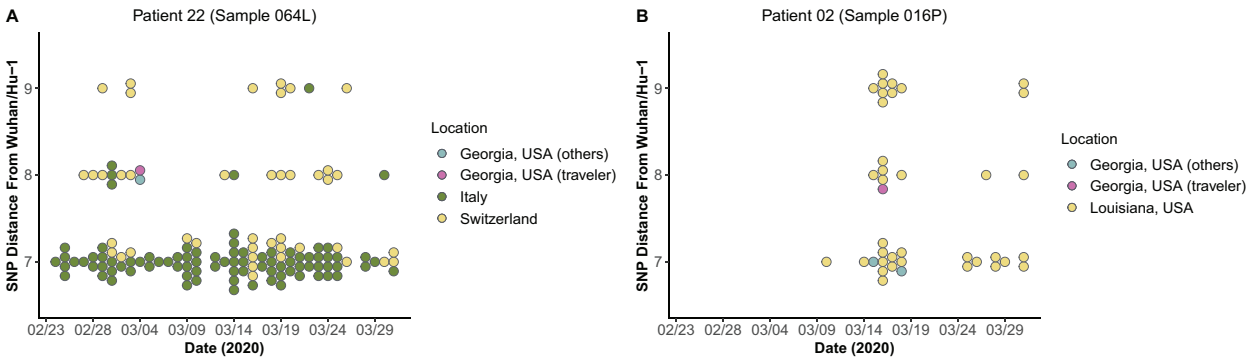


**Figure 4.** Analysis of SARS-CoV-2 whole genome sequences from EHC patients with recent travel provides examples of travel-associated infections of SARS-CoV-2 coming into Georgia. (A) The sequence from P22 (traveler) was compared to related sequences from Georgia (others) and the regions of travel. Sequences in this analysis were within the same ancestral lineage as the P22 sequence and differed from it by 0 or 1 SNPs compared to Wuhan/Hu-1 (*y*-axis). (B) The sequence from P02 (traveler) was compared to related sequences from Georgia (others) and the region of travel. As in (A), sequences in this analysis were within the same ancestral lineage as the P02 sequence and differed from it by 0 or 1 SNPs compared to Wuhan/Hu-1 (*y*-axis).

dates of testing, consistent with restrictions in place to prioritize SARS-CoV-2 testing from returning travelers in early March 2020.

In one patient (P22), there was strong SARS-CoV-2 genomic evidence that the infection had been acquired in the location of travel (Italy and Switzerland); the SARS-CoV-2 sequence from P22 was identical to 18 of the 1,657 publicly available sequences from Italy and Switzerland sampled within the same time frame. It matched one other sequence from Georgia, from a sample that had matching metadata (date of sample collection, patient age, and patient gender). We therefore presumed that these samples were from the same individual, with independent sequencing performed by our group and the GADPH. Further analysis of SARS-CoV-2 sequences within the same lineage demonstrated that there were many sequences ancestral to the P22 sequence by one SNP from Italy and Switzerland, but none from Georgia (Fig. 4A, Supplementary Table S8), consistent with travel-associated infection. Supporting this, the patient had been traveling in Europe for a month prior to symptom onset, encompassing the entire plausible incubation period for SARS-CoV-2. We were unable to draw definitive conclusions about where infection was acquired for the remaining patients with international travel, not only due to insufficient epidemiological and viral genomic data but also due to the limited diversity of circulating SARS-CoV-2 at the time (Supplementary Results, Supplementary Table S8, Supplementary Figure S11A).

Domestic travel to states with ongoing community transmission could also have introduced SARS-CoV-2 lineages into Georgia. For example, there is considerable genomic evidence that one patient in our study (P02) was infected while traveling to New Orleans. The sequence from P02 was distinct to all samples from Georgia but identical to seven SARS-CoV-2 sequences from Louisiana (Fig. 4B). This finding is consistent with a recently published study on the spread of SARS-CoV-2 into and within Louisiana (Zeller et al. 2021). By contrast, another patient in our study had also recently traveled to Louisiana (P39), yet the most closely related sequences were found in both Georgia ($N = 2$) and Louisiana ($N = 19$). These sequences were not identical to P39 but were three SNPs more ancestral to it. Thus, it is not clear based on viral genomic data whether P39 was infected in Georgia or Louisiana. This uncertainty could be resolved with detailed epidemiological data, e.g. if the patient had traveled to Louisiana outside of the plausible incubation period for SARS-CoV-2. However, travel dates were incompletely recorded in the medical record for this patient (Supplementary Table S8).

Inferring the location of infection for other domestic travelers was also challenging due to the circulation of highly similar viruses in multiple states and ambiguities in travel history. For example, the SARS-CoV-2 sequence from P14, who had traveled to Mississippi, was identical to not only one sequence from Mississippi but also six from Georgia (Supplementary Figure S11B), and the patient was in both locations during the potential incubation period. The SARS-CoV-2 sequence from P05, who had traveled to Colorado, was identical to not only two SARS-CoV-2 sequences from Colorado but also one from Georgia, and the dates of travel were incompletely documented in the medical record (Supplementary Figure S11C). The SARS-CoV-2 sequence from P27, who had traveled to North Carolina, had no identical matches but harbored an additional mutation to sequences from both North Carolina and Georgia (Supplementary Figure S11D), and the patient was in both locations during the potential incubation period. Overall, given higher rates of domestic travel as compared to international travel and the short tMRCA of all circulating SARS-CoV-2 sequences, it is unsurprising that the viral lineages circulating within US states in early 2020 were highly similar. This similarity prevented us from conclusively inferring the location of infection for domestic travelers.

An additional challenge to these analyses is that in all cases, highly similar SARS-CoV-2 sequences were present in widespread locations outside of Georgia and the region of travel (Supplementary Table S8), making it difficult to exclude the possibility that patients were infected through alternative mechanisms such as contact with another traveler or unreported travel themselves.

## 2.6 The 19B subclade disappeared by the end of April 2020

Given the genetic relationship of many Georgia sequences within clade 19B, we wanted to know to what extent this subclade seeded outbreaks beyond the timeframe of our phylogenetic analyses. We first identified the shared substitutions between these Georgia sequences to generate a subclade-defining mutational profile (Fig. 5A). These SNPs include T490A, C3177T, T18736C, C24034T, T26729C, G28077C, and A29700G as well as the two 19B defining SNPs C8782T and T28144C. Of these nine substitutions, five were non-synonymous: T490A (ORF1ab Asp75Glu), C3177T (ORF1ab Pro971Leu), T18736C (ORF1ab Phe6158Leu), G28077C (ORF8 Val62Leu), and T28144C (ORF8 Leu84Ser).

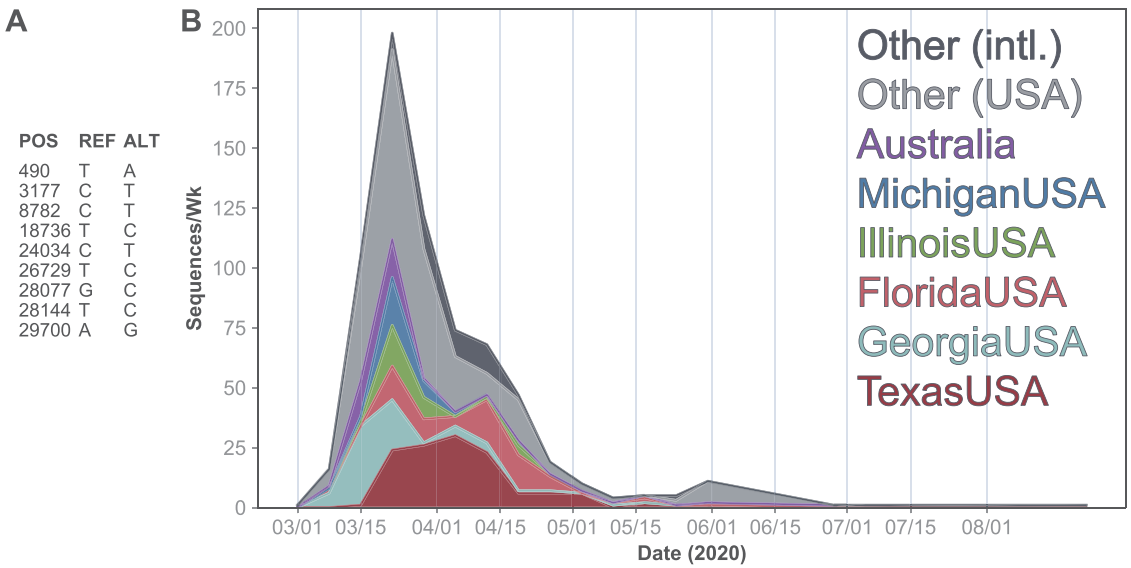


**Figure 5.** Shared mutations between related Georgia 19B sequences and global sequences harboring this mutational profile indicate its decline in early 2020. (A) Shared mutations (relative to Wuhan/Hu-1) of all the sixty-nine closely related Georgia 19B sequences, which define the 19B subclade. (B) Number of sequences per week with the mutational profile of the 19B subclade shown in (A). Sequences are colored by the region they were sampled from.

The total number of global high-quality sequences per week sharing the above subclade-defining mutational profile peaked in mid-March, within the timeframe of our phylogenetic analyses (Fig. 5B, Supplementary Table S9). The 19B subclade appeared to go extinct by the end of April; consequently, the number of Georgia sequences belonging to this subclade also dropped to zero shortly after its peak. The apparent extinction of the 19B subclade was consistent with the widely reported increase of sequences and clades harboring the D614G substitution in the spike protein (Volz et al. 2020). This subclade was most frequently detected domestically, as opposed to internationally. It was most prominently identified in Texas, where it was first identified on 11 March 2020 and was consistently observed from then until the end of April. Internationally, this 19B subclade was most frequently observed in Australia. Due to limited sequences, particularly from the state of Georgia in May 2020, we did not attempt to estimate the number of reported infections attributable to this subclade over time.

### 2.7 Samples with D614G substitution did not differ in CT value, subgenomic RNA level, or clinical severity

Due to the rapid decline of the 19B subclade, we wondered whether samples from that lineage displayed clinical or molecular features that could be associated with lower transmissibility. We focused our analysis on spike amino acid position 614. The D614G substitution is a defining substitution between 20X and 19X phylogenetic clades that has been associated with increased transmissibility (Volz et al. 2020), potentially due to higher viral loads (Plante et al. 2021). Sequence data at position 614 were available for 48 EHC patients in this study, 19 (40 per cent) of which carried the 614G amino acid residue and 29 (60 per cent) of which carried the D614 residue. There was no statistically significant difference in the SARS-CoV-2 $C_T$ (cycle threshold) value between nasopharyngeal (NP) samples with the G residue ($n = 18$) and D residue ($n = 28$) (median: 24.0 vs. 24.2, $P = 0.547$) (Supplementary Figure S12), including after adjustment for day of symptom onset and disease severity ($P = 0.84$).

We also assessed whether there were differences in subgenomic RNA (sgRNA), which is a marker for active viral replication (Wölfel et al. 2020). sgRNA was detected in a similar proportion of samples with the D and G residues [30/32 (93.8 per cent) and 15/17 (88.2 per cent), respectively; $P = 0.60$] **(Supplementary Table S1)**, and the level of sgRNA was similar for samples with the D residue (mean $C_T$ 28.6, SD 5.6) and G residue (mean $C_T$ 27.3, SD 6.3; $P = 0.50$). This did not differ when adjusted for the SARS-CoV-2 genomic RNA $C_T$ value in these samples (D residue, mean $C_T$ difference 4.7 cycles, SD 2.5; G residue, mean $C_T$ difference 4.2 cycles, SD 1.7; $P = 0.49$).

Finally, disease severity was similar between patients with the D residue (mild: 42.9 per cent, moderate: 35.7 per cent, severe: 21.4 per cent) and G residue (mild: 27.8 per cent, moderate: 44.4 per cent severe: 27.8 per cent, $P = 0.585$). More broadly, across all EHC patients, disease severity was not associated with age; other comorbidities were not assessed. Disease severity was associated with time since symptom onset; patients with severe disease had experienced a longer duration of symptoms prior to diagnosis (mean of 8.1 days) than those with mild disease (5.1 days, $P = 0.01$) (Supplementary Figure S13A). Disease severity was not associated with the SARS-CoV-2 $C_T$ value, as the mean $C_T$ was 26.6 for patients with severe disease vs. 24.8 for moderate disease vs. 24.3 for mild disease (Supplementary Figure S13B), and there was no significant association after adjustment for symptom duration, age, and the $C_T$ value ($P = 0.42$).

## 3. Discussion

Despite its high domestic and international connectivity, Georgia was spared a large SARS-CoV-2 epidemic in early 2020. However, little is known about the viral evolutionary dynamics in the state during this time. Here, we detected at least 19 introductions of SARS-CoV-2 into Georgia from phylogenetic analysis of 108 sequences obtained through the end of March 2020. As this estimate includes only those lineages represented in the available sequencing data, the true number of introductions is certainly higher. Furthermore, observing roughly 19 introductions

among only 108 sequences implies that a large proportion of the sequences in this analysis were attributed to a novel introduction compared to local transmission. While phylogenetic studies focused specifically on the spread of SARS-CoV-2 in the southeastern USA in early 2020 are relatively limited, studies from the region (broadly defined) consistently found multiple circulating lineages (Louisiana (Zeller et al. 2021), Maryland (Thielen et al. 2021), and North Carolina (McNamara et al. 2020)), indicative of multiple introductions, and this pattern is mirrored in US states in other regions (California (Deng et al. 2020), Connecticut (Fauver et al. 2020), Illinois (Lorenzo-Redondo et al. 2020), Massachusetts (Lemieux et al. 2021), New York (Gonzalez-Reiche et al. 2020), Washington (Bedford et al. 2020; Worobey et al. 2020), and Wisconsin (Moreno et al. 2020)). Existing studies that leverage genomic and travel data have shown that both international and domestic travel fueled the early domestic spread of SARS-CoV-2 (Fauver et al. 2020).

Notably, nearly 65 per cent of the SARS-CoV-2 sequences sampled from Georgia through March 2020 were highly genetically related and fell within a single 19B subclade. Bayesian phylogenetic reconstruction of these Georgia sequences, as well as globally sampled sequences within the same subclade and ancestral relatives, demonstrates that they were likely the result of a single or small number of introductions into the USA in early February. Based on our analysis, Georgia was the most likely site of introduction of this lineage into the USA, but because international SARS-CoV-2 genomic surveillance was fairly limited in early 2020, it is difficult to know this for certain. Importantly, the time to MRCA of the sequences in this subclade is estimated to have been two to four weeks before the first detected SARS-CoV-2 infection in Georgia, which was reported on 2 March 2020 (Georgia Department of Public Health 2020). Due to stochasticity in transmission dynamics at low infectious population sizes (Pekar et al. 2021), this lineage was likely introduced even earlier. The estimated detection lag of this lineage is therefore one to two weeks longer than was observed in the UK (du Plessis et al. 2021). Thus, SARS-CoV-2 was circulating within Georgia for a substantial period of time before being identified by clinical or genomic surveillance.

Finding a large number of sequences from a single or small number of introductions is consistent with the substantial transmission heterogeneity of SARS-CoV-2 that has been reported both within Georgia (Lau et al. 2020) and elsewhere (Miller et al. 2020; Popa et al. 2020; Lemieux et al. 2021). For example, a recent phylogenetic study of SARS-CoV-2 sequences from Louisiana estimated that a single introduction into Louisiana was responsible for the majority of transmission within the state following superspreading events associated with Mardi Gras (Zeller et al. 2021). Additionally, a study of genomes collected in the Boston area have identified large superspreading events associated with nursing facilities and an international business conference (Lemieux et al. 2021).

Our analysis of SARS-CoV-2 infection in domestic travelers returning to Georgia also underscores the fact that there was widespread unrecognized transmission in early 2020. In fact, due to the presence of identical viruses in multiple states, it was difficult to infer from viral genomic data alone whether returning travelers in this study were infected in Georgia or travel locations such as North Carolina, Mississippi, and Colorado. Taken together, these results emphasize that the early focus of diagnostic testing on returning international travelers, rather than more broad testing of patients with COVID-19 symptoms, led to under-recognition of existing infections in early 2020.

In addition, while our analysis of returning travelers highlights the need for more comprehensive genome sequencing of emerging pathogens, it also emphasizes the limited resolution of genomic epidemiology when the genetic diversity of a pathogen is low. Viral genomic analyses can be enhanced by the collection of finely resolved metadata. Our study benefited from linked clinical and epidemiological data for nearly half of the SARS-CoV-2 samples sequenced, but despite extensive chart review, we encountered limitations, e.g. in reporting specific dates of travel and symptom onset. Thus, there is a need for a dedicated infrastructure for data collection in the setting of outbreak analysis, beyond routinely collected clinical data.

Our study also provides information regarding the dynamics of early SARS-CoV-2 lineages in the USA. The 19B subclade, which caused most of the infections described in this study, appears to have spread from Georgia both domestically and internationally (e.g. to Australia) before dying out in April/May of 2020. The apparent extinction of this D614-containing 19B subclade occurred concurrently with the widely reported sweep of SARS-CoV-2 clades harboring the 614G mutation (Volz et al. 2020). The increased transmission of 614G-containing viruses may be due to their ability to cause infection with higher viral loads (Plante et al. 2021). We did not observe a difference in either viral load or subgenomic RNA in patients with D614 or 614G-containing viruses in this study, which may be due to small sample size.

While the 19B subclade reported here was associated with limited forward transmission, we did not find strong evidence for ongoing transmission from the other observed introductions of SARS-CoV-2 into Georgia. However, we primarily analyzed genomes collected through the end of March 2020, so it is possible that other observed introductions, particularly those that occurred later in the time frame of this analysis, seeded downstream transmission chains that are not described here.

Overall, our findings provide several key take-home messages about the early SARS-CoV-2 pandemic that may be applicable to future outbreaks. First, our study recognizes that, despite intensive effort, diagnostic testing capabilities lagged well behind SARS-CoV-2 transmission early in the pandemic. In addition, the focused effort on diagnosis in returning international travelers meant that substantial local and domestic transmission was occurring, but was missed. In a broader context, our findings highlight that highly transmissible pathogens may potentially spread faster than can be detected by the current surveillance infrastructure around the world. This lesson also applies to emerging variants of SARS-CoV-2 (Centers for Disease Control and Prevention 2021). When new variants with likely enhanced transmission are reported to be circulating widely in other countries, it is highly likely that community transmission is already occurring within the USA, given the mobility of the population. Therefore, success of public health policies and interventions countering these variants depends on early planning and implementation, prior to detection in the USA. Given the inevitable challenges in developing and rolling out diagnostic tests for a novel pathogen, these findings underscore the importance of early, empiric public health interventions to attenuate transmission while diagnostic and sequencing efforts 'catch up'.

Future pandemic responses will benefit from public health measures that presume early unrecognized transmission and act to mitigate it, while also implementing aggressive population-based surveillance, including prioritizing testing of asymptomatic contacts. These activities will be synergistic with the much needed, and now expanding, infrastructure for pathogen genomic

surveillance and enhanced collection of detailed clinical and epidemiological data.

## 4. Methods

### 4.1 Collection of clinical data and samples

This study was approved by the Emory University Institutional Review Board. Clinical data including demographics, comorbid conditions, duration of symptoms prior to testing, travel history, and severity of illness were extracted by chart review. Disease severity was classified as mild (ED or outpatient visit only), moderate (inpatient without ICU admission), or severe (inpatient with ICU admission).

Residual clinical samples (nasopharyngeal, oropharyngeal, swab samples, and bronchoalveolar lavage samples) were collected from EHC patients between 3 March 2020 and 31 March 2020, including from inpatient and outpatient sites across 8 hospitals and multiple clinics. Total nucleic acids were extracted and underwent testing in a SARS-CoV-2 triplex real-time reverse-transcriptase polymerase chain reaction (rRT-PCR), as described (Waggoner et al. 2020). Testing for subgenomic RNA was performed using a modified forward primer (5′-CGATCTCTTGTAGATCTGTTCTC-3′) and the reverse primer and probe for the N2 target used in the triplex SARS-CoV-2 rRT-PCR.

### 4.2 SARS-CoV-2 genome sequencing

Samples underwent DNAse treatment (ArcticZymes), cDNA synthesis with random primers and Superscript III (Invitrogen), Nextera XT tagmentation (Illumina), and Illumina sequencing (Babiker et al. 2020a). A median of 36.4 million reads were obtained per sample, and individual results are listed in Supplementary Table S1. Reference-based SARS-CoV-2 genome assembly was performed using viral-ngs v.2.0.21 (Broad Institute 2020) with reference NC_045512 (Wu et al. 2020). Reads per million (RPM) was calculated by dividing the number of mapped reads by the total number of reads and multiplying by 1 million.

Lower titer viruses were sequenced using a multiplex PCR amplification strategy followed by amplicon sequencing as described (Paden et al. 2020). Briefly, RNA was reverse transcribed using random hexamers. The resulting cDNA was used as a template for four pools of SARS-CoV-2-specific multiplex PCR. PCR amplicons were purified and used to prepare sequencing libraries using the Illumina Nextera FLEX kit and sequenced on an Illumina NovaSeq instrument. Reads were trimmed for quality and primers were removed using BBDuk (Bushnel 2022) and assembled using the BETACORONAVIRUS module of IRMA v.1.0.2 (Shepard et al. 2016).

### 4.3 Clade assignment

We assigned all sequences in our dataset to a given clade using Nextclade v.0.13.0 (Hadfield et al. 2018; Bedford, Hodcroft, and Neher 2020). Pango lineages were assigned using the Pangolin COVID-19 Lineage Assigner with pangoLEARN v.2021-08-09 (Rambaut et al. 2020).

### 4.4 Statistical analysis

Comparison of categorical variables was performed by the Chi square test (or Fisher's when expected frequencies $< 5$). Comparison of continuous variables was performed by the Wilcoxon rank sum test or Kruskal–Wallis test when appropriate. Correlation of $C_T$ values to log RPM was assessed by Poisson regression. Statistical analysis was performed using R v.4.0.2 (Vienna, Austria) (R Core Team 2020) and the RStudio interface v.1.3.1073 (Boston, MA, USA) (RStudio Team 2020). Maps showing the number of cases and number of sequences per Georgia county were generated with https://mapchart.net accessed on 30 August 2021.

### 4.5 Global sequence data

To place the sequences from Georgia in a global context, we downloaded all sequences sampled through 31 March 2020 and labeled as 'complete', 'high coverage', and 'collection date complete' from the Global Initiative for Sharing All Influenza Data (GISAID) database (Shu and McCauley 2017) as of 27 March 2021. We excluded any sequences from non-human hosts, any sequences related to a cruise ship, and any sequences with known travel history (to avoid biasing the ancestral state reconstruction), as annotated in the NextMeta file. These sequences, as well as the new EHC sequences presented in this analysis (when multiple samples from the same subject were available, we only included the NP swab sample), were aligned to Wuhan/Hu-1 (EPI_ISL_402125) using MAFFT v7.464 (Katoh et al. 2002) and removing any insertions relative to Wuhan/Hu-1 (Wu et al. 2020). To account for the potential sequencing error, we masked the first and last 100 nucleotides of the genome as well as sites 11,083, 15,324, and 21,575, which were identified as 'highly homoplasic' in early SARS-CoV-2 sequencing data (De Maio et al. 2020). Sequences with less than 28,000 A, C, T, and G nucleotides after aligning were removed. The GISAID Acknowledgement Table is provided in Supplementary Table S3.

### 4.6 Maximum likelihood phylogenetic analysis

For our phylogenetic analyses, we first downsampled the available global sequence data to maintain a representative geographical distribution of sequences (weighted downsampling strategy). We downsampled the available sequences from each country based on the cumulative number of reported SARS-CoV-2 cases by 31 March 2020 (Dong, Du, and Gardner 2020) and a target alignment size of 6,000 sequences. For countries where the number of available sequences was greater than the product of the target alignment size and the relative number of cumulative cases in that country, we sampled sequences with weight $1/(1+D)$ where $D$ is the minimum SNP distance of a given sequence to all available Georgia sequences. Only A, C,T, and G nucleotides were considered when calculating pairwise distances. NumPy v.1.19 (Harris et al. 2020) in Python v.3.9.4 (Python Software Foundation 2020) was used to calculate the pairwise distances. We manually included all Georgia sequences and Wuhan/Hu-1 in the final alignment. The alignment included 4,622 sequences, including 108 from Georgia (Supplementary Table S5). An alternative downsampling procedure, including a maximum of twenty sequences per country per week, was investigated to assess the robustness of our results (Supplementary Methods, Supplementary Figure S6A). Downsampling was conducted in Python using BioPython (Cock et al. 2009) and Pandas v.1.1 (Pandas Development Team 2020).

IQ-TREE v.2.1.3 (Nguyen et al. 2015) was used to generate maximum likelihood phylogenies with 1,000 ultrafast bootstrap replicates (Hoang et al. 2018), collapsing small branches, and using ModelFinder to identify the best fit nucleotide substitution model (Kalyaanamoorthy et al. 2017). A GTR + F + I + G4 model was chosen. TreeTime v.0.8.2 (Sagulenko, Puller, and Neher 2018) was used to remove any sequences falling outside four interquartile ranges of the expected molecular clock rate, rooting at Wuhan/Hu-1.

The date of internal nodes was estimated using TreeTime with a fixed clock rate of 0.001 (Duchene et al. 2020) and a coalescent skyline. TreeTime was run for a maximum of three iterations and polytomies were not resolved. Root-to-tip regression, conducted using SciPy v.1.5.4 (Virtanen et al. 2020), confirmed a significant clock rate ($P < 0.0001$) in the set of included sequences (Supplementary Figure S4).

TreeTime was also used to reconstruct the ancestral states of internal nodes (Georgia/Non-Georgia) with a sampling bias correction of 2.5. We used the reconstructed traits of internal nodes to estimate the number of introductions into Georgia (transition from a non-Georgia node to a Georgia node along a given lineage). To provide a conservative estimate, we attributed multiple Georgia nodes descending from a non-Georgia polytomous internal node to be the result of a single introduction. Furthermore, we only considered the earliest (in time) introduction into Georgia for each lineage giving rise to a Georgia sequence. In other words, we did not account for the reintroduction of a given lineage into Georgia when counting the number of introductions. This procedure was repeated on 100 bootstrap replica trees to account for phylogenetic uncertainty.

## 4.7 Bayesian phylogenetic analysis

For a more robust reconstruction of the timing and source of introduction for the highly related sequences belonging to clade 19B, we conducted a Bayesian discrete phylogeographic reconstruction. We identified the set of highly related Georgia sequences by calculating the pairwise phylogenetic distance between all Georgia sequences in the time-resolved maximum likelihood phylogeny using BioPython. SciPy was used to identify clusters in this distance matrix with a cutoff of 0.3 years. We identified sixty-nine Georgia sequences in the largest cluster.

As we wished to include the ancestral relatives to these sixty-nine sequences in our Bayesian phylogenetic analysis, we first identified their great-grandparent in the time-resolved maximum likelihood phylogeny. Next, we identified the set of nucleotide substitutions shared between all sequences that descended from that great-grandparent. We allowed for the presence of ambiguous nucleotides when identifying shared SNPs (e.g. an R nucleotide was assumed to match both A and G nucleotides). We identified three nucleotide substitutions shared between these sequences: T26729C, G28077C, and the 19B clade defining SNP T28144C. The other 19B clade defining SNP C8782T was identified in all sequences descending from this node except one, EPI_ISL_454974. Finally, we identified all 'complete', 'high coverage', and 'sampling date complete' sequences sampled through 31 March 2020 in GISAID as of 27 March 2021 that matched this mutational profile (excluding any with ambiguous nucleotides at any sites in the mutational profile) after aligning to Wuhan/Hu-1 as described above. Again, we excluded any sequences from non-human samples, related to cruise ships, or with travel history. IQ-Tree was used to generate a maximum likelihood phylogeny of these sequences with the same parameters as described above and TreeTime was used to remove any samples falling outside four interquartile ranges of the expected molecular clock rate, rooted at the best fit root as identified by least-squares regression. The final alignment included 527 sequences, of which 67 were from Georgia. Root-to-tip regression confirmed a significant clock rate ($P < 0.0001$) in the set of included sequences of sequences (Supplementary Figure S7).

To improve computational efficiency, we removed sequences from any states with fewer than four sequences in the data set or US sequences without a specified state. Furthermore, international sequences sampled after 29 February 2020 were excluded as they were evolutionary descendant from the MRCA of all US sequences and therefore likely represent exportations of this subclade from the USA (Supplementary Figures S8 and S9). Furthermore, we generated five downsampled alignments in which at most five sequences per country/US state per week were randomly sampled, including Georgia. Each downsampled alignment included 226 sequences (Supplementary Table S6).

Bayesian phylogenetic inference was conducted using BEAST2 v2.6.6 (Bouckaert et al. 2019) with Beagle v3.1.2 (Ayres et al. 2012) and discrete trait estimation implemented in BEAST_CLASSIC v.1.50 (Lemey et al. 2009). We assumed an exponential population coalescent using a Laplace distribution for the growth rate prior ($\mu = 0.0$, scale $= 10.0$) and a Lognormal ($\mu = 1.0$, $\sigma = 2.0$) prior on the population size. We used an HKY + $\Gamma$4 substitution model with a Lognormal ($\mu = 1.0$, $\sigma = 1.25$) prior on $K$. We used a relaxed molecular clock (Drummond et al. 2006) with a normal ($\mu = 1E-3$, $\sigma = 1E-4$) prior on the mean clock rate (Duchene et al. 2020), and an exponential ($\mu = 0.33$) prior on the standard deviation of the clock rate. Uniform priors were used for nucleotide frequencies and the proportion of invariant sites. We parameterized the discrete ancestral state reconstruction with a Poisson ($\lambda = (N_{traits}{}^{*}N_{traits} - 1)/8$, offset $= N_{traits} - 1$) distribution for the number of non-zero rates, a $\Gamma$ ( $\alpha = 1.0$, $\beta = 1.0$) prior for the relative rates, and a $\Gamma$ ($\alpha = 0.001$, $\beta = 1000$) prior on the rate of discrete trait changes. Rates were assumed to be symmetric. Included sequences were assigned to their country (international sequences) or state (US sequences) of origin. BEAST XML files were generated using a custom Python script and XML templates originally generated using Beauti v.2.6.3 and edited by hand. The MCMC chain was run for 285 M steps, saving every 5,000 steps. The first 10 per cent of MCMC steps were discarded as burn in. The ESS value of all parameters was >100 and >200 for parameters relevant to our conclusions as annotated by Tracer (Rambaut et al. 2018). The maximum clade credibility summary tree (with median node heights) was reconstructed using TreeAnnotator v.2.6.3. When tabulating the number of introductions of the 19B subclade into Georgia, we considered only the earliest (in time) introduction along a given lineage. In other words, we did not account for the reintroduction of a given lineage into Georgia when counting the number of introductions.

Downstream analysis of the TreeTime and BEAST output was conducted in Python using BioPython, Pandas, and NumPy. Results were visualized using Baltic v.0.1.6 (Dudas 2020), Matplotlib v.3.3.356 (Hunter 2007), and Seaborn v.0.11.157 (Waskom et al. 2020).

## 4.8 Georgia travel history

To assess the probability that patients with recent travel history were infected during travel, we compared the sequence from each traveler to sequences circulating in the region they were traveling to, sequences circulating in the state of Georgia, and sequences circulating globally. To ensure that our inferences were not biased by homoplastic artifacts in phylogenetic reconstruction, we generated a mutational profile for each traveler's sequence by identifying the SNPs relative to Wuhan/Hu-1. Insertions and deletions were not considered in this analysis. Next, we identified all sequences that matched either Wuhan/Hu-1 or the traveler's sequence at all positions in the mutational profile, not allowing for Ns or ambiguous nucleotides. We calculated the genetic distance from Wuhan/Hu-1 for the traveler's sequence and

all sequences from a given region, considering only A, C, T, and G characters. Sequences with a smaller genetic distance than the traveler sequence harbored a subset of the mutations in the traveler sequence, while those with a larger genetic distance harbored all of the mutations in the traveler sequence plus additional mutations. This analysis was conducted in Python using NumPy and Pandas. Figures for this analysis were generated in R v.4.0.4 using RStudio v.1.4.1106 with GGplot2 v.3.3 (Wickham et al. 2021).

### 4.9 Mutational profile of closely related Georgia 19B sequences

First, the shared SNPs (relative to Wuhan/Hu-1) between the sixty-nine closely related Georgia 19B sequences were identified from the sequence alignment described above using BioPython. We allowed for the presence of ambiguous nucleotides when identifying shared SNPs. We refer to sequences harboring this mutational profile as belonging to the 'clade 19B subclade'. The variants in the mutational profile were annotated using snpEff v. 5.0 (Cingolani et al. 2012).

Next, we downloaded all 'complete', 'high coverage', and 'collection date complete' from GISAID sampled and uploaded through 27 March 2020 that shared the L84S amino acid (T28144C nucleotide) substitution, a clade defining mutation of 19B. We removed non-human samples, those related to cruise ships, and samples with travel history and aligned them to Wuhan/Hu-1 with MAFFT with the same parameters described above. We identified all sequences that non-ambiguously matched the Georgia 19B subclade mutational profile (Supplementary Table S9) using NumPy in Python. We summed the number of identified sequences per week for each US state as well as the total number of other countries using Pandas. Results were visualized using Matplotlib.

## Data availability

All consensus sequence data used in this analysis are available from the GISAID (https://www.gisaid.org). Accession numbers are available in Supplementary Tables S1, S3, S5, S6, and S9. Sequence data newly generated for this project is available on NCBI under BioProject PRJNA634356, including both consensus sequences and raw reads (cleaned of human reads). Metadata for the Georgia, USA, sequences needed to replicate the analysis is available in Supplementary Tables S1, S2, S4, and S8 as well as at https://zenodo.org/record/6038869. Metadata for non-Georgia sequences is available via GISAID. Code necessary to replicate this analysis is available at https://zenodo.org/record/6038869. Output files from the BEAST analysis can be found at https://figshare.com/articles/dataset/Unrecognized_introductions_of_SARS-CoV-2_into_the_state_of_Georgia_shaped_the_early_epidemic_v_1_0/14935380.

## Supplementary data

Supplementary data is available at *Virus Evolution* online.

## Acknowledgements

We would like to acknowledge our laboratory colleagues at the Emory University Healthcare Microbiology, Molecular and Referral Laboratories, who have worked tirelessly to provide necessary care to our patients during this time. We thank the Emory Clinical Virology Research Laboratory, the Georgia Clinical Research Centers, and the Yerkes NHP Genomics Core for support in sample collection and sequencing. We thank Audrey Kunkes and the Georgia Department of Public Health for their support. Publicly available SARS-CoV-2 sequences from Georgia were generously contributed by Mayo Clinic Laboratories, Quest Diagnostics, and the U.S. Air Force School of Aerospace Medicine. We gratefully acknowledge the authors from the originating laboratories responsible for obtaining the specimens, as well as the submitting laboratories where the genome data were generated and shared via GISAID, on which this research is based (Supplementary Table S3). All submitters of data may be contacted directly via www.gisaid.org.

## Funding

This study was supported by the CDC contract 75D30121C10084 under BAA ERR 20-15-2997, the Pediatric Research Alliance Center for Childhood Infections and Vaccines and Children's Healthcare of Atlanta, and the Emory WHSC COVID-19 Urgent Research Engagement (CURE) Center, made possible by generous philanthropic support from the O. Wayne Rollins Foundation and the William Randolph Hearst Foundation. The Yerkes NHP Genomics Core is supported in part by NIH P51 OD011132, and sequence data were acquired on an Illumina NovaSeaq6000 funded by NIH S10 OD 026799. Sample collection was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378. Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number K08AI139348 (A.P.) and F31 AI154738 (M.A.M.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Conflict of interest:** None declared.

## Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention, the Public Health Service, or the US Department of Health and Human Services.

## References

Alm, E. et al. (2020) 'Geographical and Temporal Distribution of SARS-CoV-2 Clades in the WHO European Region, January to June 2020', *Eurosurveillance*, 25: pii=2001410.

Ayres, D. L. et al. (2012) 'BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics', *Systematic Biology*, 61: 170–3.

Babiker, A. et al. (2020a) 'Metagenomic Sequencing to Detect Respiratory Viruses in Persons under Investigation for COVID-19', *Journal of Clinical Microbiology*, 59: 1. (M. J. Loeffelholz, Ed.).

Babiker, A. et al. (2020b) 'SARS-CoV-2 Testing', *American Journal of Clinical Pathology*, 153: 706–8.

Bedford, T. et al. (2020) 'Cryptic Transmission of SARS-CoV-2 in Washington State', *Science*, 370: 571–5.

Bedford, T., Hodcroft, E. B., and Neher, R. A. (2020), *Updated Nextstrain SARS-CoV-2 Clade Naming Strategy*. Nextstrain. <https://nextstrain.org/blog/2021-01-06-updated-SARS-CoV-2-clade-naming> accessed 8 Feb 2022.

Bouckaert, R. et al. (2019) 'BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 15: e1006650.

Broad Institute. (2020), *Viral-Pipelines*. <https://github.com/broadinstitute/viral-pipelines> accessed 8 Feb 2022.

Bushnel, B. (2022), *BBDuk. Joint Genome Institute*. <https://sourceforge.net/projects/bbmap/> accessed Feb 8, 2022.

CDC. (2020), *CDC 2019-Novel Coronavirus (2019-ncov) Real-Time RT-PCR Diagnostic Panel*. Centers for Disease Control and Prevention. <https://www.fda.gov/media/134922/download> accessed 8 Feb 2022.

Centers for Disease Control and Prevention. (2021), *SARS-CoV-2 Variant Classifications and Definitions*. <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fcases-updates%2Fvariant-surveillance%2Fvariant-info.html> accessed 8 Feb 2022.

Cingolani, P. et al. (2012) 'A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff', *Fly*, 6: 80–92.

Cock, P. J. A. et al. (2009) 'Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics', *Bioinformatics*, 25: 1422–3.

De Maio, N. et al. (2020), *Masking Strategies for SARS-CoV-2 Alignments*. Virological. <https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480> accessed 8 Feb 2022.

Deng, X. et al. (2020) 'Genomic Surveillance Reveals Multiple Introductions of SARS-CoV-2 into Northern California', *Science*, 369: 582–7.

Dong, E., Du, H., and Gardner, L. (2020) 'An Interactive Web-based Dashboard to Track COVID-19 in Real Time', *The Lancet Infectious Diseases*, 20: 533–4.

Drummond, A. J. et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4: e88. (D. Penny, Ed.).

du Plessis, L. et al. (2021) 'Establishment and Lineage Dynamics of the SARS-CoV-2 Epidemic in the UK', *Science*, 371: 708–12.

Duchene, S. et al. (2020) 'Temporal Signal and the Phylodynamic Threshold of SARS-CoV-2', *Virus Evolution*, 6: veaa061.

Dudas, G. (2020), *Baltic*. <https://github.com/evogytis/baltic> accessed 7 Mar 2022.

Fauver, J. R. et al. (2020) 'Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States', *Cell*, 181: 990–6.e5.

Georgia Department of Public Health. (2020), *Gov. Kemp, Officials Confirm Two Cases of COVID-19 in Georgia*. Government of Georgia. <https://dph.georgia.gov/press-releases/2020-03-02/gov-kemp-officials-confirm-two-cases-covid-19-georgia> accessed 8 Feb 2022.

Gonzalez-Reiche, A. S. et al. (2020) 'Introductions and Early Spread of SARS-CoV-2 in the New York City Area', *Science*, 369: 297–301.

Hadfield, J. et al. (2018) 'Nextstrain: Real-time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3. (J. Kelso, Ed.).

Harris, C. R. et al. (2020) 'Array Programming with NumPy', *Nature*, 585: 357–62.

Health Alert Network. (2020), *Updated Guidance on Evaluating and Testing Persons for Coronavirus Disease 2019 (COVID-19)*. Centers for Disease Control and Prevention. <https://emergency.cdc.gov/han/2020/HAN00429.asp> accessed 8 Mar 2022.

Hoang, D. T. et al. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.

Hunter, J. D. (2007) 'Matplotlib: A 2D Graphics Environment', *Computing in Science & Engineering*, 9: 99–104.

Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.

Katoh, K. et al. (2002) 'MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.', *Nucleic Acids Research*, 30: 3059–66.

Lau, M. S. Y. et al. (2020) 'Characterizing Superspreading Events and Age-specific Infectiousness of SARS-CoV-2 Transmission in Georgia, USA', *Proceedings of the National Academy of Sciences*, 117: 22430–5.

Lemey, P. et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520. (C. Fraser, Ed.).

Lemieux, J. E. et al. (2021) 'Phylogenetic Analysis of SARS-CoV-2 in Boston Highlights the Impact of Superspreading Events', *Science*, 371: eabe3261.

Lorenzo-Redondo, R. et al. (2020) 'A Clade of SARS-CoV-2 Viruses Associated with Lower Viral Loads in Patient Upper Airways', *EBioMedicine*, 62: 103112.

McNamara, R. P. et al. (2020) 'High-Density Amplicon Sequencing Identifies Community Spread and Ongoing Evolution of SARS-CoV-2 in the Southern United States', *Cell Reports*, 33: 108352.

Miller, D. et al. (2020) 'Full Genome Viral Sequences Inform Patterns of SARS-CoV-2 Spread into and within Israel', *Nature Communications*, 11: 5518.

Moreno, G. K. et al. (2020) 'Revealing Fine-scale Spatiotemporal Differences in SARS-CoV-2 Introduction and Spread', *Nature Communications*, 11: 5558.

Nguyen, L. T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Paden, C. R. et al. (2020) 'Rapid, Sensitive, Full-Genome Sequencing of Severe Acute Respiratory Syndrome Coronavirus 2', *Emerging Infectious Diseases*, 26: 2401–5.

Pandas Development Team. (2020), *Pandas*. NumFocus. <https://pandas.pydata.org> accessed 14 Nov 2020.

Patel, A. et al. (2020) 'Initial Public Health Response and Interim Clinical Guidance for the 2019 Novel Coronavirus Outbreak — United States, December 31, 2019–February 4, 2020', *MMWR. Morbidity and Mortality Weekly Report*, 69: 140–6.

Pekar, J. et al. (2021) 'Timing the SARS-CoV-2 Index Case in Hubei Province', *Science*, 372: 412–7.

Perkins, T. A. et al. (2020) 'Estimating Unobserved SARS-CoV-2 Infections in the United States', *Proceedings of the National Academy of Sciences*, 117: 22597–602.

Plante, J. A. et al. (2021) 'Spike Mutation D614G Alters SARS-CoV-2 Fitness', *Nature*, 592: 116–21.

Popa, A. et al. (2020) 'Genomic Epidemiology of Superspreading Events in Austria Reveals Mutational Dynamics and Transmission Properties of SARS-CoV-2', *Science Translational Medicine*, 12: eabe2555.

Python Software Foundation. (2020), *Python Language Reference*. <http://www.python.org> accessed 29 Mar 2022.

R Core Team. (2020), *R: A Language and Environment for Statistical Computing*. Vienna, Australia. <https://www.R-project.org/> accessed 11 Dec 2020.

Rambaut, A. et al. (2018) 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901–4. (E. Susko, Ed.).

Rambaut, A. et al. (2020) 'A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 5: 1403–7.

RStudio Team. (2020) *RStudio: Integrated Development for R*. Boston, MA. <https://www.rstudio.com> accessed 11 Aug 2020.

Sagulenko, P., Puller, V., and Neher, R. A. (2018) 'TreeTime: Maximum-likelihood Phylodynamic Analysis', *Virus Evolution*, 4: 1–9.

Schuchat, A., CDC COVID-19 Response Team. (2020) 'Public Health Response to the Initiation and Spread of Pandemic COVID-19 in the United States, February 24–April 21, 2020', *MMWR. Morbidity and Mortality Weekly Report*, 69: 551–6.

Shepard, S. S. et al. (2016) 'Viral Deep Sequencing Needs an Adaptive Approach: IRMA, the Iterative Refinement Meta-assembler', *BMC Genomics*, 17: 708.

Shu, Y., and McCauley, J. (2017) 'GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality', *Eurosurveillance*, 22: 2–4.

Subbaraman, N. (2020) 'How to Address the Coronavirus's Outsized Toll on People of Colour', *Nature*, 581: 366–7.

Thielen, P. M. et al. (2021) 'Genomic Diversity of SARS-CoV-2 during Early Introduction into the Baltimore–Washington Metropolitan Area', *JCI Insight*, 6: e144350.

United States Census Bureau. (2019), *Georgia*. United States Government. <https://data.census.gov/cedsci/profile?g=0400000US13> accessed 8 Feb 2022.

SciPy 1.0 Contributors, Virtanen, P. et al. (2020) 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python', *Nature Methods*, 17: 261–72.

Volz, E. M. et al. (2020) 'Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity', *Cell*, 184: 64–75.e11.

Waggoner, J. J. et al. (2020) 'Triplex Real-Time RT-PCR for Severe Acute Respiratory Syndrome Coronavirus 2', *Emerging Infectious Diseases*, 26: 1633–35.

Waskom, M. et al. (2020), *Seaborn*. <https://seaborn.pydata.org/index.html> accessed 21 Jan 2022.

Wickham, H., et al. (2021), *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. New York. <http://ggplot2.org> accessed 8 Feb 2022.

Willis, H., and Williams, V. (2020). *A Funeral Is Thought to Have Sparked A Covid-19 Outbreak in Albany*, Ga. — *And Led to Many More Funerals*. Washington Post. <https://www.washingtonpost.com/politics/a-funeral-sparked-a-covid-19-outbreak—and-led-to-many-more-funerals/2020/04/03/546fa0cc-74e6-11ea-87da-77a8136c1a6d_story.html> accessed 8 Feb 2022.

Wölfel, R. et al. (2020) 'Virological Assessment of Hospitalized Patients with COVID-2019', *Nature*, 581: 465–9.

Worobey, M. et al. (2020) 'The Emergence of SARS-CoV-2 in Europe and North America', *Science*, 370: 564–70.

Wu, F. et al. (2020) 'A New Coronavirus Associated with Human Respiratory Disease in China', *Nature*, 579: 265–9.

Zeller, M., et al. (2021), *Emergence of an Early SARS-CoV-2 Epidemic in the United States* (Preprint). Epidemiology. <http://medrxiv.org/lookup/doi/10.1101/2021.02.05.21251235> accessed 26 Mar 2021.