medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Early introductions and community transmission of SARS-CoV-2 variant B.1.1.7 in the United States

Tara Alpert[1]*, Erica Lasek-Nesselquist[2,3]*, Anderson F. Brito[1]*, Andrew L. Valesano[4]*, Jessica Rothman[1]*, Matthew J. MacKay[5]*, Mary E. Petrone[1], Mallery I. Breban[1], Anne E. Watkins[1], Chantal B.F. Vogels[1], Alexis Russell[2], John P. Kelly[2], Matthew Shudt[2,3], Jonathan Plitnick[2,3], Erasmus Schneider[2,3], William J. Fitzsimmons[4], Gaurav Khullar[5], Jessica Metti[5], Joel T. Dudley[5], Megan Nash[5], Jianhui Wang[6], Chen Liu[6], Pei Hui[6], Anthony Muyombwe[7], Randy Downing[7], Jafar Razeq[7], Stephen M. Bart[7,8], Steven Murphy[9], Caleb Neal[9], Eva Laszlo[9], Marie L. Landry[10], Peter W. Cook[8], Joseph R. Fauver[1#], Christopher E. Mason[5#], Adam S. Lauring[4#], Kirsten St. George[2,3,#,@], Duncan R. MacCannell[8,#,@], Nathan D. Grubaugh[1,11,#,@]

[1] Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT 06510, USA
[2] Wadsworth Center, New York State Department of Health, Albany, NY 12208, USA
[3] Department of Biomedical Sciences, University at Albany, SUNY, Albany, NY 12222, USA
[4] Department of Internal Medicine, Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109, USA
[5] Tempus Labs, Chicago, IL 60654, USA
[6] Department of Pathology, Yale University School of Medicine, New Haven, CT 06510, USA
[7] Connecticut State Department of Public Health, Rocky Hill, CT 06067, USA
[8] Centers for Disease Control and Prevention, Atlanta, GA 30329, USA
[9] Murphy Medical Associates, Greenwich, CT 06830, USA
[10] Departments of Laboratory Medicine and Medicine, Yale School of Medicine, New Haven, CT 06510, USA
[11] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06510, USA
* These authors contributed equally
# Senior author
@Correspondence: kirsten.st.george@health.ny.gov; fms2@cdc.gov; nathan.grubaugh@yale.edu

Summary

The emergence and spread of SARS-CoV-2 lineage B.1.1.7, first detected in the United Kingdom, has become a national public health concern in the United States because of its increased transmissibility. Over 500 COVID-19 cases associated with this variant have been detected since December 2020, but its local establishment and pathways of spread are relatively unknown. Using travel, genomic, and diagnostic testing data, we highlight the primary ports of entry for B.1.1.7 in the US and locations of possible underreporting of B.1.1.7 cases. New York, which receives the most international travel from the UK, is likely one of the key hubs for introductions and domestic spread. Finally, we provide evidence for increased community transmission in several states. Thus, genomic surveillance for B.1.1.7 and other variants urgently needs to be enhanced to better inform the public health response.

Introduction

The rise of SARS-CoV-2 infections to unprecedented levels in the final months of 2020 has led to the evolution of several variants with concerning mutations or traits (Lauring and Hodcroft, 2021). One such variant designated B.1.1.7 (also 501Y.V1) was identified in the United Kingdom (UK) during late 2020 and proceeded to predominate circulation in the region within three months. Subsequent analyses suggested that B.1.1.7 quickly rose in frequency because it was approximately 50% more transmissible than other

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

circulating lineages (Davies et al., 2020), resulting in a 0.4-0.7 increase in the effective reproductive rate in the UK (Volz et al., 2021). Further analysis of contact tracing data suggested that infections with the B.1.1.7 variant resulted in a 30-50% higher secondary attack rate (PHE, 2020). The B.1.1.7 SARS-CoV-2 variant is defined by 17 amino acid changes, including 8 changes in the spike protein (Rambaut et al., 2020). Of particular note is the N501Y mutation in the receptor-binding domain of the spike protein, which is predicted to increase binding to human angiotensin-converting enzyme 2 (ACE2) receptors (Starr et al., 2020) and is also a defining feature of other variants of concern such as B.1.351 and P.1 (Faria et al., 2021; Lauring and Hodcroft, 2021; Tegally et al., 2020). In addition, B.1.1.7 variants have a deletion in the spike gene (Δ69/70 HV) that may increase cell infectivity (Kemp et al., 2020) and has provided a serendipitous tracking method for B.1.1.7 by causing a "spike gene target failure" (SGTF) in the commonly used Thermo Fisher TaqPath COVID-19 Combo Kit (Borges et al., 2021; Volz et al., 2021). As of February 2nd, 2021, B.1.1.7 has been detected in 73 countries (cov-lineages.org/global_report.html), which raises concerns that B.1.1.7 will follow the trajectory that it took in the UK in other places around the world.

Current virus genomic surveillance across the United States (US) is uneven, creating uncertainty about the extent of international introductions, domestic spread, and community transmission of the SARS-CoV-2 variant. There have been 541 reported COVID-19 cases associated with B.1.1.7 from 33 states (as of February 2nd, 2021)(CDC, 2021a), but these numbers are likely substantial underestimates. Approaches to enhance depth in certain populations and geographies may be required to address such uncertainty. Regardless, the existing framework limits the implementation of effective public health actions, such as targeted public health messaging and enhanced mitigation (Grubaugh et al., 2021), and allows B.1.1.7 to spread unimpeded (Galloway et al., 2021). The continued rise in B.1.1.7 cases may increase the burden on the US healthcare system and enable further evolution of mutations of public health concern (Wise, 2021).

Here, to investigate the locations of B.1.1.7 introductions into the US, identify significant surveillance gaps, and provide evidence for community transmission and domestic spread, we combined data from UK air travel into US airports, SARS-CoV-2 genomic sequencing, and clinical diagnostics. We found that SARS-CoV-2 genomic sequencing is inconsistent across the US, and we identified where surveillance for new B.1.1.7 introductions should be immediately supported. Furthermore, we found that New York likely serves as a hub for international introductions and domestic spread. Combining our work with other analyses (Washington et al., 2021), we found that community transmission of B.1.1.7 has been established and is likely increasing across the US, with primary evidence in Florida, California, New York, Connecticut, New Jersey, Michigan, Illinois, and Georgia.

Results

Locations of potential B.1.1.7 introductions from international travelers

During the early phases of the COVID-19 pandemic, international travel can seed the local establishment of new SARS-CoV-2 variants. Thus, to obtain a relative estimate of where B.1.1.7 introductions were most likely to occur, we analyzed air passenger travel volumes coming into all US airports from the initial primary source of the variant - the UK (**Figure 1A**; showing top 20 airports). The air passenger volumes represent the total number of passengers entering each airport as a final destination for the month of October, 2020. This was the most recent data available at the time of analysis, but (***1***) it represents a period in which B.1.1.7 was rapidly expanding in the UK and (***2***) we assume that the relative passenger volumes between airports was unlikely to significantly change during December 2020 to January 2021. We found that the highest travel volumes from the UK to the US were into the New York-New Jersey area (14,317

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

passengers), California (7,157), Florida (5,703), Illinois (4,216), Texas (3,724), Massachusetts (3,006), and the District of Columbia (2,966; **Data S1**).

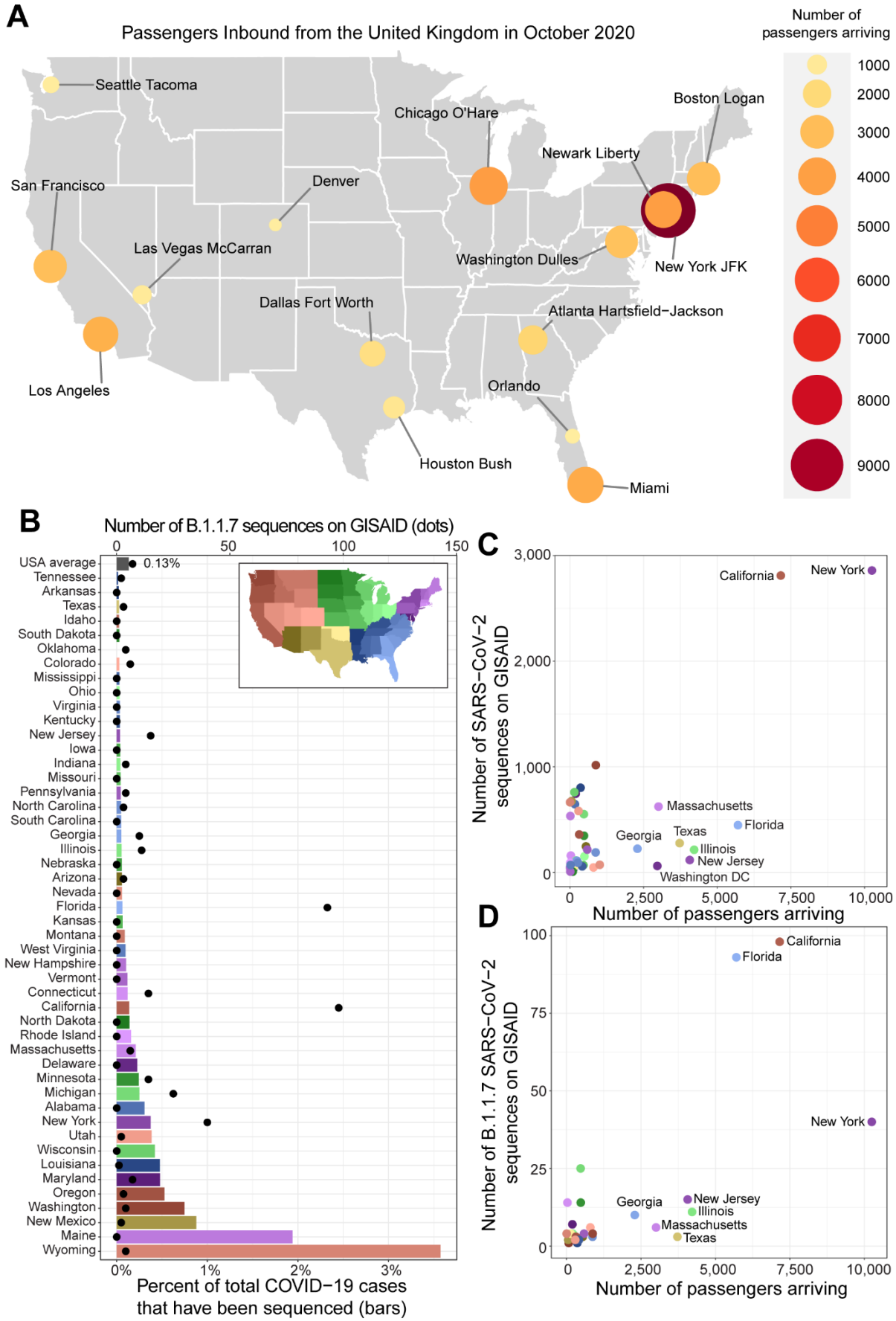
Genomic surveillance gaps and potential underreporting of B.1.1.7 cases

As of February 2, 2021, 541 B.1.1.7 cases from 33 states have been identified in the US (CDC, 2021a). These numbers, however, are likely significantly underreported, much more than overall COVID-19 cases, as whole genome sequencing is needed for B.1.1.7 confirmation. To identify where B.1.1.7 cases may be disproportionately underreported in the US, we evaluated the intensity of SARS-CoV-2 genomic surveillance for each state and compared that to our estimates of where B.1.1.7 introductions are likely to occur (**Figure 1A**).

We started by evaluating the percent of sequenced SARS-CoV-2 clinical samples relative to the number of reported COVID-19 cases (**Figure 1B**). For this, we downloaded (***1***) all SARS-CoV-2 genomes available on GISAID (gisaid.org; accessed on 2021-02-04) with "USA" listed as a location and (***2***) the total number of new COVID-19 cases for each state from December 2020 to January 2021 (covidtracking.com; accessed on 2021-02-04; **Data S1**). For this two month period, which was crucial for B.1.1.7 introductions, we found that only 0.1% of the COVID-19 cases in the US were sequenced and posted on GISAID. As sample testing, transport, sequencing, analysis, and data submission can take multiple days to weeks, more data from this time period will likely be available in the near future. Still, there are 26 states with less than 0.1% of the COVID-19 cases with available SARS-CoV-2 sequences during December 2020 to January 2021, and 14 of those states have not submitted B.1.1.7 sequences (**Figure 1B**, **Data S1**).

While a low fraction of sequenced COVID-19 cases will certainly hinder the detection of B.1.1.7 and other variants of concern, these data alone may not indicate where B.1.1.7 may be disproportionately underreported. Therefore we compared our estimates of B.1.1.7 introductions using air passenger volumes from the UK to all SARS-CoV-2 (**Figure 1C**, **Data S1**) and B.1.1.7 (**Figure 1D**, **Data S1**) genomes sequenced per state. Of the states receiving more than 2,500 air passengers from the UK, we found that New Jersey (118 sequenced cases, 0.04% of total), Illinois (215, 0.05%), Texas (268, 0.03%), and Florida (448, 0.06%) have sequenced a low proportion of samples from COVID-19 cases (**Figure 1B**). In Florida, many of the available SARS-CoV-2 sequences were targeted using TaqPath SGTF results, and thus they have sequenced 93 B.1.1.7 genomes to date, the second most in the country (Washington et al., 2021). In places like New Jersey, Illinois, and Texas, however, if SARS-CoV-2 genomic surveillance could be increased it might determine if B.1.1.7 cases are disproportionately underreported as compared to New York and California.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 1. Identification of regions at risk for importation of B.1.1.7 and at risk for underreporting B.1.1.7. **A.** Number of passengers inbound from the United Kingdom for the top 15 airports in October 2020 (labeled dots; size and color scaling with population, see legend) are shown on a map of the continental United States (grey). **B.** Bar plot represents the percentage of cases in each state during Dec 2020 and Jan 2021 (bottom x-axis; sourced from https://covidtracking.com/data) that have sequences uploaded to GISAID.org (accessed on Feb 4, 2021). Bars are colored according to region (legend, top right). The number of B.1.1.7 sequences for each state (top x-axis; black dots) was determined by the Pangolin lineage assignment in the GSIAID.org metadata. **C-D.** Total number of passengers arriving from the UK in October 2020 to each state in the continental US is plotted against the number of SARS-CoV-2 **(C)** or B.1.1.7 SARS-CoV-2 **(D)** sequences available on GISAID.org. Points are colored according to region (legend from panel B). The data used to create this figure are listed in **Data S1**.

Phylogenetic evidence for multiple B.1.1.7 introductions

Our travel data indicated that we should have observed many separate B.1.1.7 introductions within the US from the UK and perhaps other international locations where the variant may be circulating. To investigate this further and to evaluate if domestic spread had already occurred within the US, we combined our SARS-CoV-2 sequencing data from the Centers for Disease Control and Prevention, New York State Department of Health, University of Michigan, and Yale University. Our phylogenetic analysis of these sequences suggests that there were many separate introductions into the US and that some introductions within states were likely from domestic spread (**Figure 2**).

We generated or received permission to use 289 B.1.1.7 genomes collected from December 19, 2020 to January 28, 2021 from the following states: California (90), Florida (86), New York (28), Michigan (19), Connecticut (15), New Jersey (14), Illinois (10), Georgia (10), Pennsylvania (4), Indiana (3), North Carolina (3), Texas (2), Minnesota (2), Tennessee (2), and Massachusetts (1). We combined our data with 2,947 B.1.1.7 genomes available from the UK and other countries, and constructed divergence and time-informed maximum-likelihood phylogenetic trees (**Figure 2**). From our ancestral state reconstructions, we identified up to 73 potential separate B.1.1.7 introductions from Europe into the US, assuming singletons as independent introductions (**Figure 2B**). Our estimate, however, can be significantly influenced by sampling biases and gaps, sequencing and processing errors, and the imperfectness of estimating transitions between locations among sequences with low genetic diversity. Moreover, based on the low rate of sequencing of COVID-19 cases in the US (**Figure 1B**), the true number of introductions is significantly higher than 73. Importantly, though, our results inform us that (***1***) many separate B.1.1.7 introductions occurred within the US, even with decreased travel and testing requirements from the UK (Grubaugh et al., 2021), and (***2***) the majority of the sequenced B.1.1.7 cases already resulted from domestic spread or community transmission.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

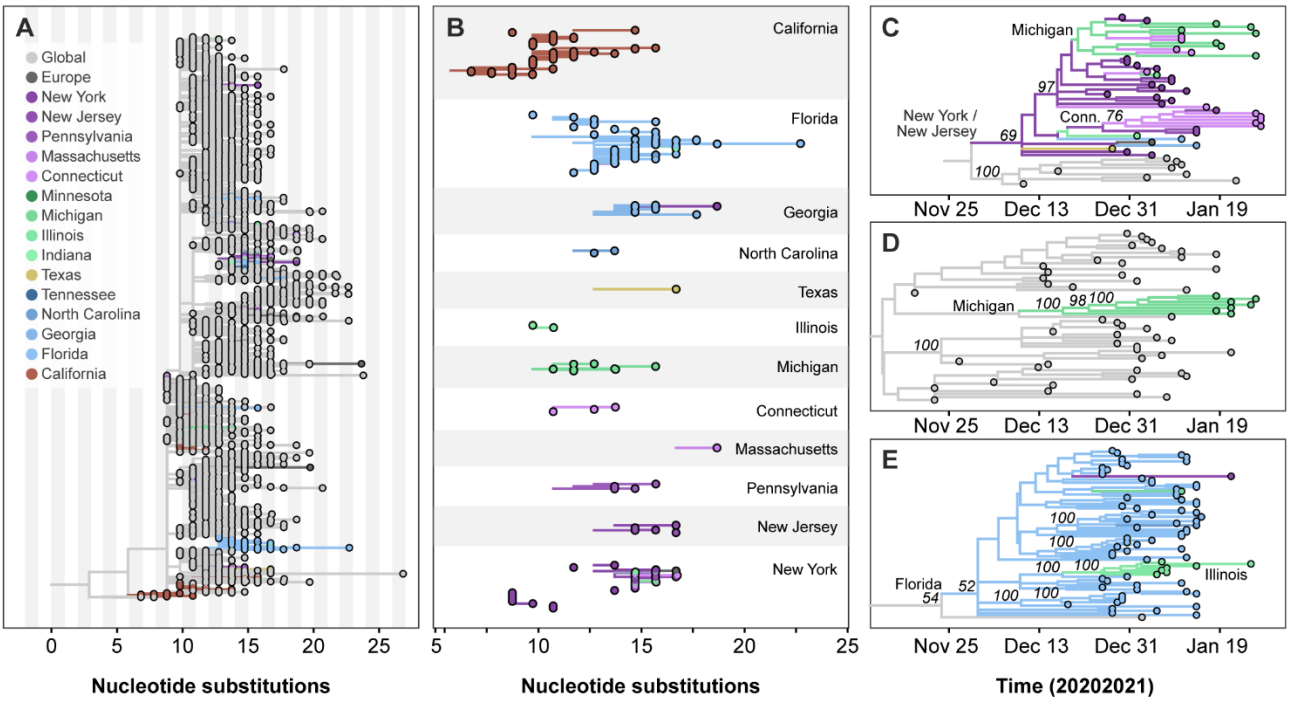


Figure 2. Multiple introductions, domestic spread, and community transmission of B.1.1.7 SARS-CoV-2 in the United States. A. Maximum likelihood phylogeny of B.1.1.7, including representative genomes from the US, Europe, other global locations. **B**. Highlights of singletons and clades representing direct introductions of B.1.1.7 from Europe, into distinct regions of the US, based on the same phylogenetic tree shown in (**A**). A list of the Europe to US transitions can be found in **Data S1**. **C-E**. Time-informed maximum likelihood phylogeny of distinct B.1.1.7 clades showing instances of domestic spread (**C**,**E**) and/or community transmission within New York (**C**), Connecticut (**C**), Michigan (**C**,**D**), and Illinois (**E**). The list of SARS-CoV-2 sequences used in this study and author acknowledgements can be found in **Data S2**.

New York as a hub for introductions and domestic spread

Based on our travel data (**Figure 1**) and evidence during the early phases of the COVID-19 pandemic (Gonzalez-Reiche et al., 2020; Maurano et al., 2020), we hypothesized that New York would be an important early location for B.1.1.7 establishment and further spread within the US. From the 28 B.1.1.7 genomes sequenced from New York in our dataset, we estimate that they represent at least 12 distinct introductions, assuming singletons as independent introductions (**Figure 2B**; **Data S1**). Of these introductions, one led to a cluster of at least 8 sequenced B.1.1.7 cases in New York (**Figure 2B**) and another led to a cluster of 43 sequenced B.1.1.7 cases across the US (13 from New York and 30 from 7 other states), indicating sustained community transmission in New York and domestic spread (**Figure 2C**).

The first case of B.1.1.7-associated COVID-19 in New York was retrospectively discovered from a sample collected in Saratoga County on December 24th, 2020 (Wadsworth-291673-01). At least 7 additional B.1.1.7-related COVID-19 cases were confirmed in Saratoga County, neighboring Warren County, and Westchester County, from samples collected between late December 2020 and mid-January 2021, with the majority having confirmed epidemiological links to the first detected case in Saratoga County. While SARS-CoV-2 genomes sequenced from these cases are identical to >200 other B.1.1.7 genomes primarily from the UK, their geographic proximity and established epidemiological links support a single introduction into the area rather than multiple seeding events.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Another cluster of B.1.1.7 cases in New York involved at least 13 individuals from New York City, Long Island, and Mid-Hudson, with the earliest retrospectively identified on December 27th, 2020 (**Figure 2C**). This cluster creates a larger monophyletic clade with sequenced B.1.1.7 cases from 7 other states: Connecticut (11), Michigan (8), New Jersey (7), Illinois (1), Texas (1), Georgia (1), and Florida (1). Our ancestral state reconstruction and time-informed phylogenetic analysis suggests that these cases likely resulted from a single introduction into the US, likely into New York, by early December (90% CI 2020-12-05 to 2020-12-19). While it is possible that this clade represents multiple introductions that occurred later in December, some of these cases have shared epidemiological links or travel histories that would indicate undetected transmission and interstate spread of B.1.1.7 was occurring throughout December 2020 and January 2021. One of the first two COVID-19 cases in Connecticut associated with B.1.1.7 (Yale-S019, collected on 2021-01-04) was associated with travel to New York, which led to transmission to household contacts in Connecticut (Yale-S105, Yale-S106, Yale-1006). Thus, New York represents both an international sink and a domestic source for B.1.1.7 transmission in the US.

Evidence for community transmission of B.1.1.7 across the United States

Our phylogenetic analysis from New York and other analysis from California and Florida (Washington et al., 2021) indicate that the B.1.1.7 introductions that led to community transmission began between late November and early December 2020. To investigate if there was evidence for community transmission of the B.1.1.7 SARS-CoV-2 variant in other states, we used two forms of B.1.1.7 surveillance data: virus genomic (**Figure 2**) and TaqPath SGTF results (**Figure 3**). The virus genomic data can reveal if clusters of B.1.1.7 cases are genetically similar, suggesting that the infections were locally acquired (Grubaugh et al., 2019). As the spike gene Δ69/70 HV deletion in B.1.1.7 genomes causes SGTF results when using the TaqPath assay, tracking the occurrence of SGTF can provide an indirect measure of changes in B.1.1.7 population frequency (Borges et al., 2021; Volz et al., 2021). Our combined data show evidence for genetically related and increased frequency of suspected B.1.1.7 cases in Connecticut, New Jersey, and Illinois during January 2021, indicative of community transmission.

We first used phylogenetic analyses of the sequence data to identify genetically related B.1.1.7 sequences within states. Based on the criteria of: (***1***) having sequenced identical B.1.1.7 genomes from multiple individuals within a state, and; (***2***) limiting our analysis to states with at least 10 B.1.1.7 genomes available for comparison, we found phylogenetic evidence for community B.1.1.7 transmission in Connecticut (**Figure 2C**), New Jersey (**Figure 2C**), Michigan (**Figure 2C, 2D**), Illinois (**Figure 2E**), and Georgia (Washington et al., 2021). In addition to likely domestic spread from New York-New Jersey to Connecticut and Michigan (**Figure 2C**), separate international introductions also likely contributed to local establishment in Michigan (**Figure 2D**).

Our collective clinical diagnostic laboratories performed 639,013 TaqPath COVID-19 RT-PCR tests from November 2020 through January 2021 across 5 states: Illinois (336,869 tests), Connecticut (192,760), New Jersey (52,725), New York (35,515), and Texas (11,144). The SARS-CoV-2 test positivity rate varied across the states and months, but all were high (range = 6-23%; **Figure 3A**; **Data S1**). Among the positive tests, the frequency of SGTF results increased during January in all states except Texas, where we had the least amount of TaqPath data (**Figure 3B**). While rates of SGTF increased to 0.6-2% during January 2021 in Illinois, Connecticut, New Jersey, and New York, this should be seen as a maximum potential frequency of B.1.1.7 within the tested populations, and not an absolute frequency. In the US, a second SARS-CoV-2 lineage, B.1.375, also has the spike gene Δ69/70 HV deletion that causes SGTF, and has a higher frequency than B.1.1.7 in many states from December 2020 to January 2021 (Larsen and Worobey,

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

2020; Moreno et al., 2021). In fact, from 254 SGTF samples tested at Yale University (from Connecticut, New Jersey, Illinois, and New York) as of January 31st, 2021, only 37 (15%) were identified as B.1.1.7 through sequencing. In contrast, the percentage SGTF identified as B.1.1.7 in Florida and California were significantly higher (Washington et al., 2021). Thus, SGTF data may be currently more helpful for monitoring trends in some states than others. Overall, while such indirect measures should be interpreted carefully, the combination of recent SGTF increases and the genetic relatedness of B.1.1.7 sequences suggests that community transmission of B.1.1.7 has become established and is on the rise in several US states.

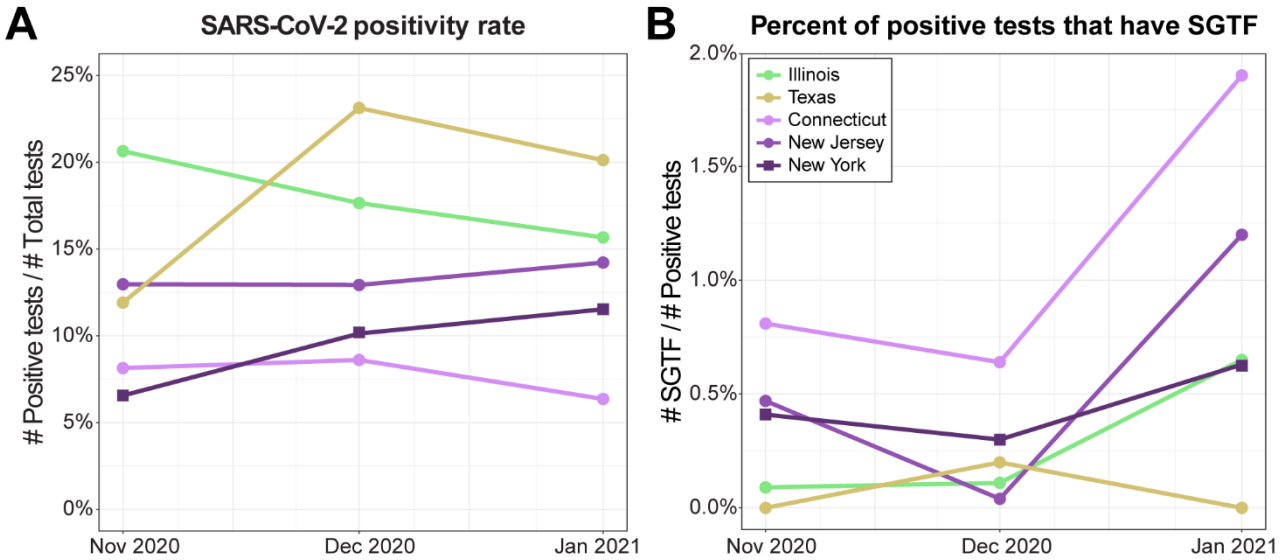


Figure 3. Increasing frequency of spike gene target failure (SGTF) results across five US states. A. The positivity rate of SARS-CoV-2 testing for five states (legend, panel B) is shown for a three-month window, calculated as the number of positive tests (including SGTF) divided by total tests. **B.** The percentage of positive tests which have SGTF are shown for the same time-period and states (legend, top left) from A. The data used to create this figure are listed in **Data S1**.

Discussion

Despite travel restrictions and increased testing requirements, we found evidence for a large number of independent B.1.1.7 introductions into the US, many of which have led to sustained community transmission. Incoming air passenger volumes from the UK predict that New York, California, and Florida would be at highest risk for importation, and high numbers of B.1.1.7 sequences from these states agree with that hypothesis. Indeed, our phylogenetic analyses suggest that in addition to separate introduction events, New York acted as a hub for B.1.1.7 importation and spread to other states. Moreover, we find evidence for the establishment of community transmission of B.1.1.7 in New York, New Jersey, Connecticut, and Illinois during January, 2021. Overall, our data highlight the relative ease with which SARS-CoV-2 variants can spread undetected throughout the US, particularly in areas where genomic surveillance efforts are minimal.

COVID-19 cases associated with the B.1.1.7 variant are likely significantly underreported across the US. This is because, as a whole, only about 0.13% of the COVID-19 cases in the US were sequenced during the months of December 2020 and January 2021, the period with the highest case rates in the country. The sequencing capacity is highly variable across the country, and our travel data helps to identify regions

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

which may be disproportionately underreporting cases of B.1.1.7 and where it would be prudent to immediately prioritize variant surveillance. States such as Texas, Illinois, and New Jersey received moderately high levels of UK travel, yet reported low numbers of SARS-CoV-2 genomic sequences from recent months, presenting the likelihood that B.1.1.7 is significantly underdetected in these regions.

It is estimated that 5% of the COVID-19 cases should be sequenced to detect emerging variants when they exist at a prevalence of 0.1% to 1.0% (Vavrek et al., 2021). With the nation-wide decrease in COVID-19 cases since reaching a new peak in early January 2021 and the initiatives to increase the capacity of SARS-CoV-2 sequencing within state public health laboratories (CDC, 2021b), there should be a considerable increase in the proportion of COVID-19 cases sequenced in the US during coming months. However, we should not expect an immediate jump to that 5% threshold, and we will need to supplement sequencing with other more conventional laboratory approaches. In the UK, the frequency of B.1.1.7 was primarily tracked across locations using SGTF results from the TaqPath COVID-19 PCR assay (Borges et al., 2021; Volz et al., 2021). Here we also demonstrate how SGTF may provide an initial assessment of B.1.1.7 population frequency changes, and how these results can help to prioritize samples for sequencing. In fact, Florida also received a high volume of UK travel (ranked #3 in the US) and has a low overall rate of virus sequencing (0.06% of cases sequenced, ranked #25), but targeted sequencing of SGTF samples has helped identify a large number of B.1.1.7 cases in this state (93, ranked #2)(Washington et al., 2021). The primary issue with using SGTF data is that the results are not definitive for B.1.1.7 in the US (Larsen and Worobey, 2020; Moreno et al., 2021). Thus, SGTF should be used with caution as a screening tool, not to measure the absolute frequency of B.1.1.7 within a population. Other PCR assays are being developed that are more specific for B.1.1.7 detection and that can also detect other current variants of concern (B.1.351 and P.1) (Vogels et al., 2021), and their widespread use for surveillance in public health labs could provide immediate data to guide public health decision-making, especially in areas where B.1.1.7 cases may be disproportionately underestimated.

Our study has several important limitations. First, our travel data are limited to incoming passengers from the UK, which will miss the potential for introductions from other countries where B.1.1.7 may be circulating. We limited our analysis to the UK for flight originations as this was the only known region where B.1.1.7 was dominant at the time of our analysis. Furthermore, our analysis did not account for the likelihood of transmission among the different regions in the US. As COVID-19 cases were at or near their peak across the country, our assumption was that transmission potential was high everywhere, and that the numbers of potentially infected travelers was a more significant factor. In reality, local conditions and behaviors play an important role for B.1.1.7 establishment, and could explain why cases are low in some states. Finally, our SARS-CoV-2 genomic data suggest that domestic spread and community transmission have been established in many places across the US, but the significant undersampling that we discuss throughout this manuscript highlights that we are missing several important pieces to this story. This suggests that some of our estimated transitions and estimated timings of introductions will likely need to be revisited when more data are available.

The SARS-CoV-2 B.1.1.7 variant of concern has become established in many states within the US. We must use this opportunity to reinforce communications and messaging surrounding the importance of mitigation measures to prevent this variant from exacerbating an already crippling pandemic (Grubaugh et al., 2021). Our estimates of introductions and community transmission of B.1.1.7 are significantly biased and underrepresented due the inadequacies of genomic surveillance in many places across the country. Public health authorities should be operating under the assumption that community transmission of B.1.1.7 is happening throughout the US, and that it is expanding at an exponential rate (Washington et al., 2021). In reality, it is difficult to implement new measures without supporting data, especially if they impact schools

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

or businesses. Thus, increasing surveillance for B.1.1.7 and other variants through sequencing and more conventional methods must be made a high priority.

Experimental Model and Subject Details

Ethics Statement

The Institutional Review Board from the Yale University Human Research Protection Program determined that the RT-qPCR testing and sequencing of de-identified remnant COVID-19 clinical samples obtained from clinical partners conducted in this study is not research involving human subjects (IRB Protocol ID: 2000028599).

Residual nasopharyngeal and saliva specimens from individuals who tested positive for SARS-CoV-2 by RT-PCR were obtained from the Michigan Medicine Clinical Microbiology Laboratory, University (of Michigan) Health Services, and LynxDx (Ann Arbor, MI). This work was approved by the University of Michigan Institutional Review Board (IRB Protocol ID: HUM185966), Expanded sequencing in January 2020 was performed as part of a public health investigation.

Residual portions of respiratory specimens from individuals who tested positive for SARS-CoV-2 by RT-PCR were obtained from the Wadsworth Center and partnering clinical laboratories. This work was approved by the New York State Department of Health Institutional Review Board, under study numbers 02-054 and 07-022.

Methods

Flight volumes and maps

The flight travel volume data were provided by OAG Aviation Worldwide Ltd. OAG Traffic Analyser, Version 2.5.11 2020 (https://analytics.oag.com/analyser-client/home [Accessed 16 December 2020]). Travel volume numbers are modeled estimates based on ticket sales and reporting from airline carriers. Travel volume represents the aggregate number of passenger journeys, not necessarily unique individuals. A subset of the available data was used only to capture flights whose origin was the UK and whose final destination was an airport in the US for flights that occurred in October 2020.The map presented in Figure 1, which shows the flight volume data per airport on flights inbound from the UK in October 2020 and overlaid onto a map of the US, was generated using R, with the maps and ggplot2 packages (Becker et al., 2018; Wickham, 2016).

Sample selection, screening, and SARS-CoV-2 sequencing

Yale University

Sample selection and RNA extraction

Samples were received in partnership with various clinical laboratories as either purified RNA or original nasal swab in viral transport media. Samples were screened for S-gene target failure (SGFT) using the Thermo Fisher TaqPath COVID-19 Combo Kit diagnostic assay prior to receipt at Yale. Nucleic acid was extracted from original samples (300 µL) using the MagMAX viral/pathogen nucleic acid isolation kit (Thermo Fisher) and eluted into 75 µL. All RNA was then screened again using an assay developed by our laboratory that is specific for variants of concern (Vogels et al., 2021). Samples identified by the screen as potential variants were then prioritized for sequencing.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Oxford Nanopore library preparation and sequencing

RNA extracted from positive samples served as the input for an amplicon-based approach for sequencing on the Oxford Nanopore Technologies (ONT; Oxford, United Kingdom) MinION (Quick et al., 2017). Sequencing libraries were prepared using the ONT Ligation Sequencing Kit (SQK-LSK109) and the ONT Native Barcoding Expansion pack as described in the ARTIC Network's protocol with V3 primers (IDT) (Quick, 2020) with the following modifications: cDNA was generated with SuperScriptIV VILO Master Mix (Thermo Fisher Scientific, Waltham, MA, USA), all amplicons were generated using 35 cycles of amplification, amplicons were then normalized to 15 ng for each sample, end repair incubation time was increased to 25 min followed by an additional bead-based clean up, and all clean up steps used a ratio of 1:1 beads:sample. No template controls were introduced for each run at the cDNA synthesis and amplicon synthesis steps and were taken through the entire library preparation and sequencing protocol to detect any cross-contamination. None was observed. 25 ng of the final library was loaded on a MinION R9.4.1 flow cell and sequenced for approximately 8-10 hours.

Bioinformatics processing

The RAMPART application from the ARTIC Network was used to monitor approximate genome coverage for each sample and control in real time during the sequencing run (https://github.com/artic-network/rampart). Fast5 files were basecalled using the the Guppy basecaller 4.4.0 fast model and consensus genomes were generated according to the ARTIC bioinformatic pipeline (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html) which uses Nanopolish to call variants (Loman et al., 2015). A threshold of 20x coverage was required for each amplicon to be included in the consensus genome.

University of Michigan

Sample selection, RNA extraction, reverse transcription, and genome amplification

All available specimens in the month of January were prepared for sequencing. Residual transport media or saliva was centrifuged at 1200 x g. and aliquoted. For nasopharyngeal and sputum specimens, RNA was extracted with the Invitrogen PureLink Pro 96 Viral RNA/DNA Purification Kit (200 µL of input sample eluted in 100 µL) or the QIAamp Viral RNA Mini kit (140 µL of input sample eluted in 50 µL). For saliva specimens, RNA was extracted with the Thermo Fisher MagMAX Viral RNA Isolation Kit (200 µL of input sample eluted in 50 µL). Extracted RNA was reverse transcribed with SuperScript IV (Thermo Fisher). For each sample, 1 µL of random hexamers and 1 µL of 10 mM dNTP were added to 11 µL of RNA, heated at 65 deg˚C for 5 min, and placed on ice for 1 min. Then a reverse transcription master mix was added (4 µL of SuperScript IV buffer, 1 µL of 0.1M DTT, 1 µL of RNaseOUT RNase inhibitor, and 1 µL of SSIV reverse transcriptase) and incubated at 42 °C for 50 min, 70 °C for 10 min, and held at 4 °C. SARS-CoV-2 cDNA was amplified in two multiplex PCR reactions with the ARTIC Network version 3 primer pools and protocol. Viral cDNA was amplified with the Q5 Hot Start High-Fidelity DNA Polymerase (NEB) with the following thermocycler protocol: 98 °C for 30 s, then 35 cycles of 98 °C for 15 s, 63 °C for 5 min, and final hold at 4 °C. Reaction products for a given sample were pooled together in equal volumes.

Illumina library preparation and sequencing

Pooled PCR product was purified with 1X volume of AMPure beads (Beckman-Coulter). Sequencing libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit (NEB) according to the manufacturer's protocol. Barcoded libraries were pooled in equal volume and extracted with a 1% agarose gel to remove adapter dimers. Pooled libraries were quantified with the Qubit 1X dsDNA HS Assay Kit (Thermo Fisher). Libraries were sequenced on an Illumina MiSeq (v2 chemistry, 2x250 cycles) at the University of Michigan Microbiome Core facility. Reads were aligned to the Wuhan-Hu-1 reference genome

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(GenBank MN908947.3) with BWA-MEM version 0.7.15. Sequencing adaptors and amplification primer sequences were trimmed with iVar 1.2.1. Consensus sequences were called with iVar 1.2.1 by simple majority at each position (>50% frequency), placing an ambiguous N at positions with fewer than 10 reads.

Oxford Nanopore library preparation and sequencing

After multiplex PCR amplification, libraries were prepared for sequencing with the Oxford Nanopore Technologies MinION using the ARTIC Network version 3 protocol (Quick, 2020). Samples were prepared in batches of 24 with one-pot native barcoding. Pooled PCR products were diluted in nuclease-free water with a dilution factor of 10. Amplicon ends were prepared for ligation with the NEBNext Ultra II End Repair/dA-Tailing Module (NEB). Unique barcodes (Oxford Nanopore Native Barcoding Expansion kits) were ligated per sample with the NEB Blunt/TA Ligase Master Mix. After barcoding, reactions were pooled together in equal volumes and purified barcoded amplicons with 0.4X volume of AMPure beads. Oxford Nanopore sequencing adapters were ligated with the NEBNext Quick Ligation Module (NEB) and the library was purified with 1X volume of AMPure beads. Final libraries were quantified with the Qubit 1X dsDNA HS Assay Kit (Thermo Fisher). Each library (15-20 ng) was loaded onto a flow cell (FLO-MIN106) and sequenced with the MinION.

Bioinformatics processing

Sequencing progress was monitored with RAMPART. Basecalling was performed with Guppy v4.0.14 and consensus genomes were called using the ARTIC Network bioinformatics pipeline (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html).

New York State Department of Health, Wadsworth Center

Sample selection and RNA extraction

Respiratory swabs in viral transport medium previously identified as SARS-CoV-2 positive by real-time RT-PCR were selected for sequencing, and included specimens received and tested in the Wadsworth Virology Laboratory and those submitted by clinical laboratories. An enhanced surveillance program was initiated in December 2020 and included retrospective sequencing of positive samples dating back to September 2020. Samples were generally required to have real-time Ct values less than 30 and minimal residual volumes of 100 µL. Most nucleic acid extractions were performed on a Roche MagNAPure 96 with the Viral NA Small Volume Kit (Roche, Indianapolis, IN) with 100µL sample input and 100µL eluate. Samples of special concern with Ct values in the low 30s were extracted on a NUCLISENS easyMAG instrument (bioMerieux, Durham, NC) with 1,000 µL sample input and 25 µL eluate.

Illumina library preparation and sequencing

Extracted RNA was processed for whole genome sequencing with a modified ARTIC protocol (https://artic.network/ncov-2019). Briefly, cDNA was synthesized with SuperScript™ IV reverse transcriptase (Invitrogen, Carlsbad, CA, USA) and random hexamers. Amplicons were generated by pooled PCR with two premixed ARTIC V3 primer pools (Integrated DNA Technologies, Coralville, IA, USA). Additional primers to supplement those showing poor amplification efficiency (https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019) were added separately to the pooled stocks. PCR conditions were 98℃ for 30 seconds, 24 cycles of 98℃ for 15 seconds/63℃ for 5 minutes, and a final 65℃ extension for 5 minutes. Amplicons from pool 1 and pool 2 reactions were combined and purified by AMPure XP beads (Beckman Coulter, Brea, CA, USA) with a 1X bead-to-sample ratio and eluted in 10mM Tris-HCl (pH 8.0). The amplicons were quantified using Quant-IT™ dsDNA Assay Kit on an ARVO™ X3 Multimode Plate Reader (Perkin Elmer, Waltham, MA, USA). Illumina sequencing libraries

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

were generated using the Nextera DNA Flex Library Prep Kit with Illumina Index Adaptors and sequenced on a MiSeq instrument (Illumina, San Diego, CA, USA).

Oxford Nanopore library preparation and sequencing

RNA was processed using the same ARTIC V3 protocol as described for Illumina library preparation. MinION libraries for up to 24 samples were generated according to the COVID-19 PCR tiling protocol (ONT). Native barcodes (ONT EXP-NBD104 and EXP-NBD114) were ligated to each DNA sample with NEB Blunt/TA Ligase Master Mix. Amplicons were pooled and purified using 0.4X AMPure XP beads and short-fragment buffer (ONT EXP-SFB001). Oxford Nanopore sequencing adapters were ligated with NEBNext Quick Ligation Module (NEB) and libraries were purified with 0.4X AMPure XP beads. About 15-25ng of each library was loaded on a FLO-MIN106 flowcell and sequenced with the MinION. Basecalling was performed by Guppy v4.2.3.

Bioinformatics processing

Illumina libraries were processed with ARTIC nextflow pipeline (https://github.com/connor-lab/ncov2019-artic-nf/tree/illumina, last updated April 2020). Briefly, reads were trimmed with TrimGalore (https://github.com/FelixKrueger/TrimGalore) and aligned to the reference assembly MN908947.3 (strain Wuhan-Hu-1) by BWA (Li & Durbin 2010). Primers were trimmed with iVar (Grubaugh et al. 2018) and variants were called with samtools mpileup function (Li et al. 2009), the output of which was used by iVar to generate consensus sequences. Positions were required to be covered by a minimum depth of 50 reads and variants were required to be present at a frequency ≥0.75. Consensus sequences were generated by the ARTIC bioinformatic pipeline v1.1.3 with Medaka variant calling (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html) for Oxford Nanopore libraries.

Clinical TaqPath COVID-19 RT-PCR testing and SGTF determination

Routine clinical COVID-19 diagnostic testing was performed in Clinical Laboratory Improvement Amendments of 1988 (CLIA) regulated laboratories at Yale University, Yale New Haven Hospital, and Tempus Labs following Emergency Use Authorization protocols submitted to the US Food and Drug Administration for use of the Applied Biosystems TaqPath COVID-19 Combo Kit (catalog number A47814). SGTF results were defined as any SARS-CoV-2 positive sample with N or ORF1AB Ct < 30 and S gene undetermined. The data were aggregated on a month and state level for surveillance purposes.

Quantification and statistical analysis

Phylogenetic analysis

Our dataset was composed of 289 SARS-CoV-2 B.1.1.7 genomes, collected from 15 US states from 2020-12-19 to 2021-01-27, provided by the Centers for Disease Control and Prevention, New York State Department of Health, University of Michigan, and Yale University. It was complemented with genomes submitted to GISAID (gisaid.org): 2,947 international B.1.1.7 genomes, mostly from Europe (2,923), collected between 2020-09-21 and 2021-01-31. This generated a dataset of 3,237 genomes, with author acknowledgements listed in **Data S2**. We performed multiple sequence alignment (MSA) using MAFFT (Katoh and Standley, 2013) implemented in augur (Hadfield et al., 2018). For phylogenetic analysis, the 5' and 3' ends of the MSA were masked alongside other problematic sites (Maio et al., 2020) using a script provided with the augur pipeline. Maximum likelihood analysis was performed using IQTree (Minh et al., 2020), under a GTR nucleotide substitution model, with 1,000 UFBoot replicates (Minh et al., 2013). Inference of divergence times and reconstruction of ancestral states for discrete phylogeographic analysis

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

were performed using TreeTime (Sagulenko et al., 2018). Phylogenetic trees were visualized using the Python package baltic (https://github.com/evogytis/baltic) and auspice (Hadfield et al., 2018).

Resource Availability

Lead contact

Further information and requests for data, resources, and reagents should be directed to and will be fulfilled by the Lead Contact, Nathan D. Grubaugh (nathan.grubaugh@yale.edu).

Materials availability

Data and code availability

Data used to produce all of the figures are included in **Data S1** and **Data S2**. Genomic data are available on GISAID (see **Data S2** for accession numbers). The air passenger data used in this study are proprietary and were purchased from OAG Aviation Worldwide Ltd. These data were used under the United States Centers for Disease Control and Prevention license for the current study and so are not publicly available. The authors are available to share the air passenger data upon reasonable request and with the permission of OAG Aviation Worldwide Ltd.

Supplemental Information Description

Data S1. Data for Figures 1-3.
Data S2. List of SARS-CoV-2 sequences used in this study and author acknowledgements.

Acknowledgments

We thank A. Grills and S. Morrison from the Centers for Disease Control and Prevention (CDC) for providing the OAG flight data, all of the frontline and essential workers for their continued service during the pandemic, and our friends and family - particularly V. Parsons, P. Jack, and S. Taylor - for their support. This work was funded by CTSA Grant Number TL1 TR001864 (T.A. and M.E.P.), Fast Grant from Emergent Ventures at the Mercatus Center at George Mason University (N.D.G.), CDC Contract # 75D30120C09570 (N.D.G.), CDC Contract # 75D30120C09870 (A.S.L.). Initial funding for sequencing at the Wadsworth Center was provided by the New York Community Trust. 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.

Author Contributions

Conceptualization, T.A., E. L-N., A.F.B., A.L.V., J.R., M.J.M., J.R.F, C.E.M, A.S.L, K.S.G, D.R.M., N.D.G.; Investigation, T.A., E.L-N., A.F.B., A.L.V., J.R., M.J.M., M.E.P., M.B., A.E.W., G.K., C.E.M., J.W., M.L.L.; Writing - Original Draft, T.A., E.L-N., A.F.B., J.F., N.D.G; Writing - Review & Editing, T.A., E. L-N., A.F.B., A.L.V., J.R., M.J.M., M.E.P., M.B., A.E.W., C.B.F.V., A.R., J.P.K.,M.S., J.P., E.S., W.J.F., G.K., J.M., J.T.D., M.N., J.W., C.L., P.H., A.M., R.D., J.R., S.M.B., S.M., C.N., E.L., M.L.L., P.C., J.R.F, C.E.M, A.S.L, K.S.G, D.R.M., N.D.G.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Declarations of Interests

M.J.M, G.K., J.M., J.T.D., M.N., and C.E.M. work for Tempus Labs. K.S.G. receives research support from ThermoFisher for the development of assays for the detection and characterization of viruses. All other authors declare no competing interests.

References

Becker, R.A., Wilks, A.R., Brownrigg, R., Minka, T.P., and Deckmyn, A. (2018). maps: Draw Geographical Maps. R package version 3.3. 0.

Borges, V., Sousa, C., Menezes, L., A, M.G., Picão, M., Almeida, J.P., Vieita, M., Santos, R., Silva, A.R., Costa, M., et al. (2021). Tracking SARS-CoV-2 VOC 202012/01 (lineage B.1.1.7) dissemination in Portugal: insights from nationwide RT-PCR Spike gene drop out data.

CDC (2021a). US COVID-19 Cases Caused by Variants.

CDC (2021b). Cases, Data, and Surveillance.

Davies, N.G., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J., Pearson, C.A.B., Russell, T.W., Tully, D.C., Abbott, S., Gimma, A., et al. (2020). Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. medRxiv 2020.12.24.20248822.

Faria, N.R., Claro, I.M., Candido, D., Moyses Franco, L.A., Andrade, P.S., Coletti, T.M., Silva, C.A.M., Sales, F.C., Manuli, E.R., Aguiar, R.S., et al. (2021). Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. Virological.

Galloway, S.E., Paul, P., MacCannell, D.R., Johansson, M.A., Brooks, J.T., MacNeil, A., Slayton, R.B., Tong, S., Silk, B.J., Armstrong, G.L., et al. (2021). Emergence of SARS-CoV-2 B.1.1.7 Lineage - United States, December 29, 2020-January 12, 2021. MMWR Morb. Mortal. Wkly. Rep. *70*, 95–99.

Gonzalez-Reiche, A.S., Hernandez, M.M., Sullivan, M.J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J., Khan, Z., et al. (2020). Introductions and early spread of SARS-CoV-2 in the New York City area. Science *369*, 297–301.

Grubaugh, N.D., Ladner, J.T., Lemey, P., Pybus, O.G., Rambaut, A., Holmes, E.C., and Andersen, K.G. (2019). Tracking virus outbreaks in the twenty-first century. Nature Microbiology *4*, 10–19.

Grubaugh, N.D., Hodcroft, E.B., Fauver, J.R., Phelan, A.L., and Cevik, M. (2021). Public health actions to control new SARS-CoV-2 variants. Cell *0*.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kemp, S.A., Collier, D.A., Datir, R., Ferreira, I.A., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., et al. (2020). Neutralising antibodies drive Spike mediated SARS-CoV-2 evasion (medRxiv).

Larsen, B.B., and Worobey, M. (2020). Identification of a novel SARS-CoV-2 Spike 69-70 deletion lineage circulating in the United States.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Lauring, A.S., and Hodcroft, E.B. (2021). Genetic Variants of SARS-CoV-2-What Do They Mean? JAMA.

Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods *12*, 733–735.

Maio, D., Walker, C., Borges, R., Weilguny, L., Slodkowicz, G., Goldman, N., and Nicola (2020). Masking strategies for SARS-CoV-2 alignments.

Maurano, M.T., Ramaswami, S., Zappile, P., Dimartino, D., Boytard, L., Ribeiro-Dos-Santos, A.M., Vulpescu, N.A., Westby, G., Shen, G., Feng, X., et al. (2020). Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region. Genome Res. *30*, 1781–1788.

Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188–1195.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534.

Moreno, G., Braun, K., Larsen, B.B., Alpert, T., Worobey, M., Grubaugh, N., Friedrich, T., O'Connor, D., Fauver, J., and Brito, A. (2021). Detection of non-B.1.1.7 spike ∆69/70 sequences (B.1.375) in the United States.

PHE (2020). Investigation of novel SARS-CoV-2 variant, Technical Briefing 3. Public Health England *3*.

Quick, J. (2020). nCoV-2019 sequencing protocol v3.

Quick, J., Grubaugh, N.D., Pullan, S.T., Claro, I.M., Smith, A.D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T.F., Beutler, N.A., et al. (2017). Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat. Protoc. *12*, 1261–1276.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D.L., and Volz, E. (2020). Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virological.

Sagulenko, P., Puller, V., and Neher, R.A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol *4*, vex042.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., et al. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell *182*, 1295–1310.e20.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., et al. (2020). Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa (medRxiv).

Vavrek, D., Speroni, L., Curnow, K.J., Oberholzer, M., Moeder, V., and Febbo, P.G. (2021). Genomic surveillance at scale is required to detect newly emerging strains at an early timepoint (medRxiv).

Vogels, C.B.F., Breban, M., Alpert, T., Petrone, M.E., Watkins, A.E., Hodcroft, E.B., Mason, C.E., Khullar, G., Metti, J., Dudley, J.T., et al. (2021). PCR assay to enhance global surveillance for SARS-CoV-2 variants of concern (medRxiv).

Volz, E., Mishra, S., Chand, M., Barrett, J.C., Johnson, R., Geidelberg, L., Hinsley, W.R., Laydon, D.J., Dabrera, G., O'Toole, Á., et al. (2021). Transmission of SARS-CoV-2 Lineage B.1.1.7 in England:

medRxiv preprint doi: https://doi.org/10.1101/2021.02.10.21251540; this version posted February 12, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Insights from linking epidemiological and genetic data (medRxiv).

Washington, N.L., Gangavarapu, K., Zeller, M., Bolze, A., Cirulli, E.T., Schiabor Barrett, K.M., Larsen, B.B., Anderson, C., White, S., Cassens, T., et al. (2021). Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. medRxiv 2021.02.06.21251159.

Wickham, H. (2016). GGPLOT2: Elegant Graphics for Data Analysis 2016 Springer-Verlag, New York.

Wise, J. (2021). Covid-19: The E484K mutation and the risks it poses. BMJ *372*, n359.