RESEARCH ARTICLE SUMMARY

CORONAVIRUS

Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events

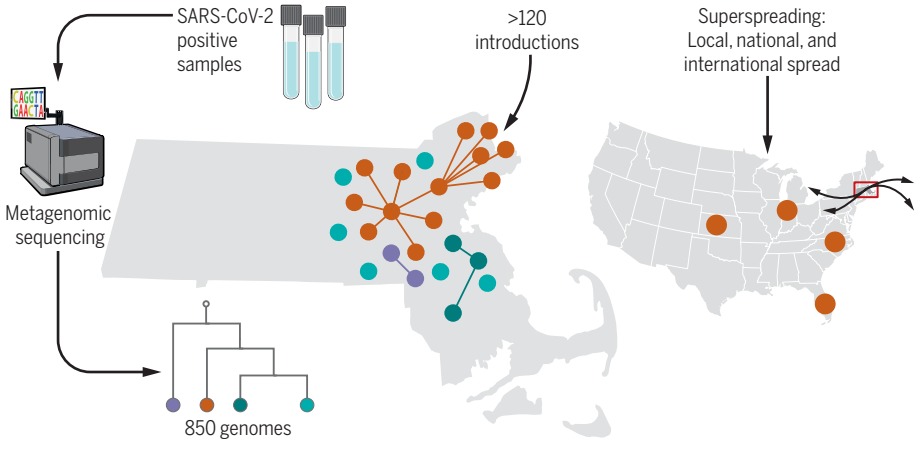
Jacob E. Lemieux*†, Katherine J. Siddle*, Bennett M. Shaw, Christine Loreth, Stephen F. Schaffner, Adrianne Gladden-Young, Gordon Adams, Timelia Fink, Christopher H. Tomkins-Tinch, Lydia A. Krasilnikova, Katherine C. DeRuff, Melissa Rudy, Matthew R. Bauer, Kim A. Lagerborg, Erica Normandin, Sinéad B. Chapman, Steven K. Reilly, Melis N. Anahtar, Aaron E. Lin, Amber Carter, Cameron Myhrvold, Molly E. Kemball, Sushma Chaluvadi, Caroline Cusick, Katelyn Flowers, Anna Neumann, Felecia Cerrato, Maha Farhat, Damien Slater, Jason B. Harris, John A. Branda, David Hooper, Jessie M. Gaeta, Travis P. Baggett, James O'Connell, Andreas Gnirke, Tami D. Lieberman, Anthony Philippakis, Meagan Burns, Catherine M. Brown, Jeremy Luban, Edward T. Ryan, Sarah E. Turbett, Regina C. LaRocque, William P. Hanage, Glen R. Gallagher‡, Lawrence C. Madoff‡, Sandra Smole‡, Virginia M. Pierce‡, Eric Rosenberg‡, Pardis C. Sabeti†‡, Daniel J. Park‡, Bronwyn L. MacInnis†‡

INTRODUCTION: We used genomic epidemiology to investigate the introduction and spread of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the Boston area across the first wave of the pandemic, from March through May 2020, including high-density sampling early in this period. Our analysis provides a window into the amplification of transmission in an urban setting, including the impact of superspreading events on local, national, and international spread.

RATIONALE: Superspreading is recognized as an important driver of SARS-CoV-2 transmission, but the determinants of superspreading—why apparently similar circumstances can lead to very different outcomes—are poorly understood. The broader impact of such events, both on local transmission and on the overall trajectory of the pandemic, can also be difficult to determine. Our dataset includes hundreds of cases that resulted from superspreading events with different epidemiological features, which allowed us to investigate the nature and effect of superspreading events in the first wave of the pandemic in the Boston area and to track their broader impact.

RESULTS: Our data suggest that there were more than 120 introductions of SARS-CoV-2 into the Boston area, but that only a few of these were responsible for most local transmission: 29% of the introductions accounted for 85% of the cases. At least some of this variation results from superspreading events amplifying some lineages and not others. Analysis of two superspreading events in our dataset illustrate how some introductions can be amplified by superspreading. One occurred in a skilled nursing facility, where multiple introductions of SARS-CoV-2 were detected in a short time period. Only one of these led to rapid and extensive spread within the facility, and significant mortality in this vulnerable population, but there was little onward transmission. A second superspreading event, at an international business conference, led to sustained community transmission, including outbreaks in homeless and other higher-risk communities, and was exported domestically and internationally, ultimately resulting in hundreds of thousands of cases. The two events also differed substantially in the genetic variation they generated, possibly suggesting varying transmission dynamics in superspreading events. Our results also show how genomic data can be used to support cluster investigations in real time—in this case, ruling out connections between contemporaneous cases at Massachusetts General Hospital, where nosocomial transmission was suspected.

CONCLUSION: Our results provide powerful evidence of the importance of superspreading events in shaping the course of this pandemic and illustrate how some introductions, when amplified under unfortunate circumstances, can have an outsized effect with devastating consequences that extend far beyond the initial events themselves. Our findings further highlight the close relationships between seemingly disconnected groups and populations during a pandemic: Viruses introduced at an international business conference seeded major outbreaks among individuals experiencing homelessness; spread throughout the Boston area, including to other higher-risk communities; and were exported extensively to other domestic and international sites. They also illustrate an important reality: Although superspreading among vulnerable populations has a larger immediate impact on mortality, the cost to society is greater for superspreading events that involve younger, healthier, and more mobile populations because of the increased risk of subsequent transmission. This is relevant to ongoing efforts to control the spread of SARS-CoV-2, particularly if vaccines prove to be more effective at preventing disease than blocking transmission. ■



Schematic outline of this genomic epidemiology study. Illustrated are the numerous introductions of SARS-CoV-2 into the Boston area; the minimal spread of most introductions; and the local, national, and international impact of the amplification of one introduction by a large superspreading event.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: lemieux@broadinstitute.org (J.E.L.); pardis@broadinstitute.org (P.C.S.); bronwyn@broadinstitute.org (B.L.M.)
‡These authors contributed equally to this work.

Cite this article as J. E. Lemieux *et al.*, *Science* 371, eabe3261 (2021). DOI: 10.1126/science.abe3261

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abe3261

RESEARCH ARTICLE

CORONAVIRUS

Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events

Jacob E. Lemieux[1,2]*†, Katherine J. Siddle[1,3]*, Bennett M. Shaw[1,2], Christine Loreth[1], Stephen F. Schaffner[1,3,4], Adrianne Gladden-Young[1], Gordon Adams[1], Timelia Fink[5], Christopher H. Tomkins-Tinch[1,3], Lydia A. Krasilnikova[1,3], Katherine C. DeRuff[1], Melissa Rudy[1], Matthew R. Bauer[1,6], Kim A. Lagerborg[1,6], Erica Normandin[1,7], Sinéad B. Chapman[1], Steven K. Reilly[1,3], Melis N. Anahtar[8], Aaron E. Lin[1,3], Amber Carter[1], Cameron Myhrvold[1,3], Molly E. Kemball[1,7], Sushma Chaluvadi[1], Caroline Cusick[1], Katelyn Flowers[1], Anna Neumann[1], Felecia Cerrato[1], Maha Farhat[9,10], Damien Slater[2], Jason B. Harris[2,11], John A. Branda[8], David Hooper[2], Jessie M. Gaeta[12,13], Travis P. Baggett[12,14,15], James O'Connell[12,14,15], Andreas Gnirke[1], Tami D. Lieberman[1,16], Anthony Philippakis[1], Meagan Burns[5], Catherine M. Brown[5], Jeremy Luban[1,17,18], Edward T. Ryan[2,4,15], Sarah E. Turbett[2,8,15], Regina C. LaRocque[2,15], William P. Hanage[19], Glen R. Gallagher[5]‡, Lawrence C. Madoff[5,20]‡, Sandra Smole[5]‡, Virginia M. Pierce[8,21,22]‡, Eric Rosenberg[2,8]‡, Pardis C. Sabeti[1,3,4,18,23]†‡, Daniel J. Park[1]‡, Bronwyn L. MacInnis[1,4,18]†‡

Analysis of 772 complete severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes from early in the Boston-area epidemic revealed numerous introductions of the virus, a small number of which led to most cases. The data revealed two superspreading events. One, in a skilled nursing facility, led to rapid transmission and significant mortality in this vulnerable population but little broader spread, whereas other introductions into the facility had little effect. The second, at an international business conference, produced sustained community transmission and was exported, resulting in extensive regional, national, and international spread. The two events also differed substantially in the genetic variation they generated, suggesting varying transmission dynamics in superspreading events. Our results show how genomic epidemiology can help to understand the link between individual clusters and wider community spread.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has now caused more than 50 million infections and more than 1 million reported deaths (*1*) in one of the worst public health crises of the past century. Cases are currently surging to unprecedented levels in the United States, reaching more than 180,000 cases reported daily during November 2020. Massive ongoing transmission globally underscores that most countries have not found effective ways to control spread of the virus; better understanding of transmission dynamics could contribute to more targeted and effective responses to the pandemic. Reports of COVID-19 transmission have featured clusters of cases linked to gatherings, including ones in workplaces (*2*) and churches (*3*) and especially in close living environments such as care homes (*4*) and homeless shelters (*5*). These clusters are thought to often involve superspreading (*6*, *7*), in which one individual infects many others (defined here as more than eight secondary cases) (materials and methods), yet the contribution of these events to regional and national transmission is not well understood. Instead, the evidence indicating that case clusters and superspreading events are major drivers of transmission has largely been based on time-series data showing an increase in cases after them (*8*), which has limited ability to determine the contribution of any event to overall transmission. Contact tracing from such events can be similarly uninformative because it is resource intensive, invasive, and often limited in scope. Likewise, without genetic data about the viruses involved, it is often not possible to distinguish superspreading events from other forms of locally intense transmission or from cases that occur in close proximity by chance. Yet, understanding the role of superspreading events in transmission is critical for prioritizing public health interventions. To further that understanding, we used genomic epidemiology to investigate the introduction and spread of SARS-CoV-2 in the Boston, Massachusetts area, which was severely affected in the first wave of the pandemic. These data allowed us to study early outbreak dynamics and to examine the role of importations and superspreading events in fueling epidemic spread.

Genomic epidemiology of Boston superspreading events

The first known case in the Boston area was confirmed on 1 February 2020 (*9*); case counts rapidly increased through March and peaked in the third week in April. We performed viral genome sequencing and phylogenetic analysis of SARS-CoV-2–positive nasopharyngeal (NP) samples collected between 4 March and 9 May 2020 by the Massachusetts Department of Public Health (MADPH) and Massachusetts General Hospital (MGH). Our dataset includes nearly all confirmed early cases of the epidemic (Fig. 1, A and B); samples from many of the highest-prevalence communities in the Boston area across the first wave (Fig. 1C), including Chelsea, Revere, and Everett (Fig. 1C and fig. S1); and samples from putative superspreading events that involved an international conference and congregate living environments, specifically among residents and staff at a skilled nursing facility (SNF) and in homeless shelters. As seen elsewhere, close-quarters living facilities such as these have been disproportionately affected by COVID-19 in Massachusetts, accounting for 22% of confirmed cases and 64% of reported deaths through 1 August 2020 (*10*).

We generated 778 complete SARS-CoV-2 assemblies (>98% complete) from 772 individuals, and an additional 72 partial genomes (>80%

[1]Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA. [2]Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [4]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA, USA. [5]Massachusetts Department of Public Health, Boston, MA, USA. [6]Harvard Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA. [7]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [8]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [9]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [10]Division of Pulmonary and Critical Care, Massachusetts General Hospital, Boston, MA, USA. [11]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [12]Institute for Research, Quality, and Policy in Homeless Health Care, Boston Health Care for the Homeless Program, Boston, MA, USA. [13]Section of General Internal Medicine, Boston University Medical Center, Boston, MA, USA. [14]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [15]Department of Medicine, Harvard Medical School, Boston, MA, USA. [16]Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [17]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA. [18]Massachusetts Consortium on Pathogen Readiness, Boston, MA 02115, USA. [19]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA. [20]University of Massachusetts Medical School, Infectious Diseases and Immunology, Worcester, MA 01655, USA. [21]Pediatric Infectious Disease Unit, Massachusetts General Hospital for Children, Boston, MA, USA. [22]Department of Pathology, Harvard Medical School, Boston, MA, USA. [23]Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, MD 20815, USA.

*These authors contributed equally to this work.

†Corresponding author. Email: lemieux@broadinstitute.org (J.E.L.); pardis@broadinstitute.org (P.C.S.); bronwyn@broadinstitute.org (B.L.M.) ‡These authors contributed equally to this work.

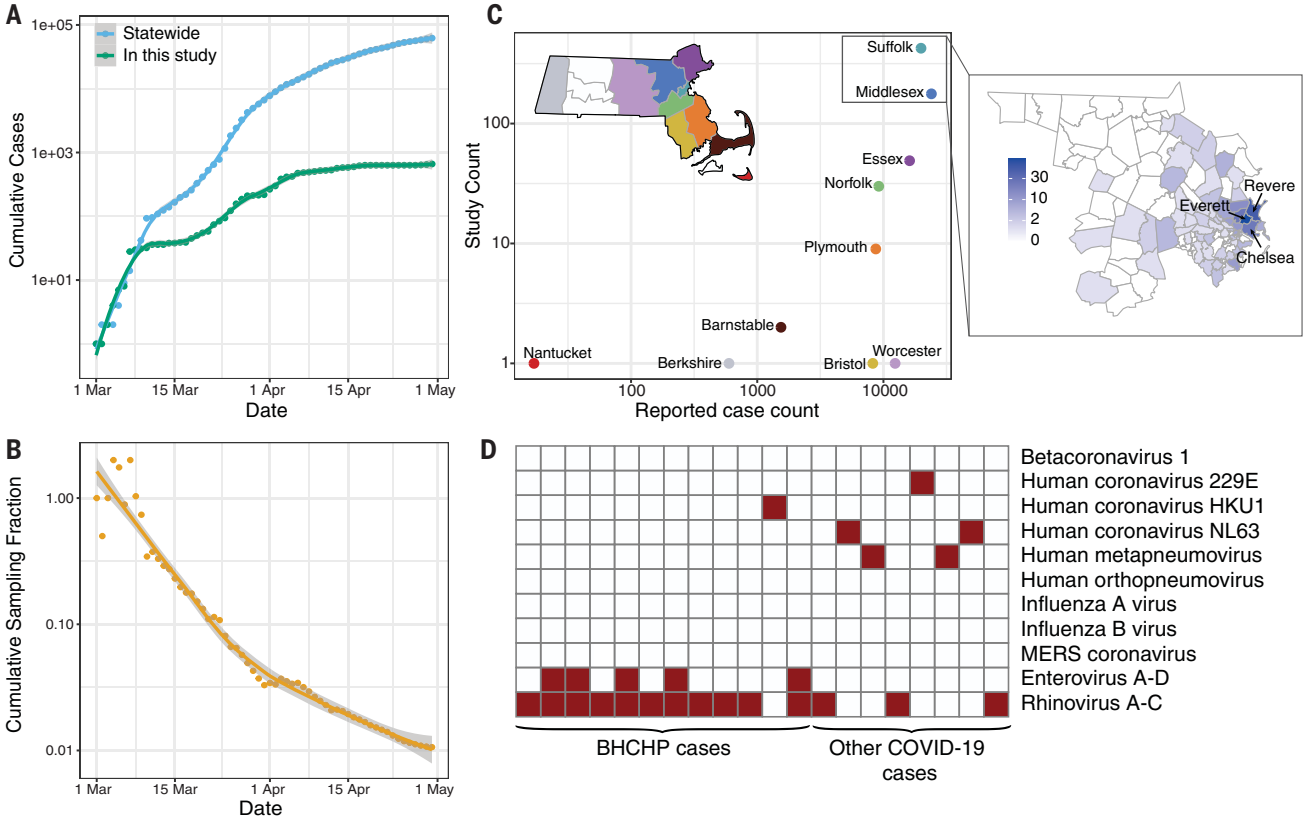


Fig. 1. Epidemiology of SARS-CoV-2 in Massachusetts and of sequenced viral genomes. (**A**) Cumulative confirmed and presumed cases reported statewide in Massachusetts (*10*) from 1 March through 1 May 2020 and the number of these cases that successfully yielded complete genomes with >98% coverage (green) in this study. (**B**) Cumulative proportion of all Massachusetts confirmed positive cases with complete genome sequences from distinct individuals that are part of this dataset over time. (**C**) Total number of cases compared with cases in this study by Massachusetts county. Points are colored by state as shown in the state map. Suffolk and Middlesex counties are shown in detail to the right, with counts from this study shown by ZIP code. (**D**) Detection of common respiratory viruses from metagenomic sequencing data. Samples with more than 10 reads that mapped to at least one of these viruses by using Kraken2 are shown in red. Enterovirus and Rhinovirus species have been grouped owing to the difficulty in discriminating at the sequence level.

complete), using Illumina-based unbiased metagenomic short-read sequencing, followed by reference-guided assembly using viral-ngs 2.0.21 software (*11*) with the Wuhan-Hu-1 sequence (NC_045512.2) as the reference (materials and methods). Genome recovery and coverage were strongly correlated with viral abundance (fig. S2) and clinical diagnostic test results (fig. S3). Genomes were separated from one another by a median of six single-nucleotide polymorphisms (SNPs) (interquartile range four to nine SNPs; range 0 to 85 SNPs) (fig. S4, A and B). As expected during rapid population expansion, most alleles were rare, as assessed from a strongly negative Tajima's D statistic throughout the genome (fig. S4C). In 20 samples (1.4% of sequenced cases), we identified the presence of at least one other common respiratory pathogen (Fig. 1D) through sequencing and confirmed it with a second assay (fig. S5). Co-infections were more commonly detected in residents and staff of homeless shelters (12 of 314) than in the other cases in the study (8 of 1117) (P = 0.0002, Fisher's exact test).

We constructed a phylogenetic tree from this SARS-CoV-2 dataset alone, and we constructed additional trees from these data combined with repeated subsampling (Fig. 2A) from the Global Initiative on Sharing All Influenza Data (GISAID) (materials and methods) (*12*). These trees form the basis of our analysis of the Boston-area epidemic. The presence of a temporal signal in our dataset (fig. S6) means that a molecular clock can be fitted to infer the timing of ancestral branching on the basis of the SARS-CoV-2 genomes.

We identified putative introductions into the Boston area by carrying out ancestral state reconstruction for these phylogenetic trees (materials and methods). In total, we identified more than 122 [95% (confidence interval (CI) 122 to 161, median 143] putative introductions into the Boston area through 9 May, stemming from sources on four continents (Table 1 and fig. S7, A and B). We characterize these introductions as putative because detailed ancestral reconstruction is limited by gaps in the global record of available genomes (*13*) and because the time scale of migration (hours to days) may exceed the rate of viral evolution (~1 new substitution every 13 days). Most of these inferred introductions occurred early in the pandemic, in March and early April, primarily from elsewhere in North America and from Europe (Table 1 and Fig. 2B). We observed close phylogenetic relatedness between genomes from the Boston area and genome sequences from elsewhere in the northeastern and eastern United States (fig. S8), which is consistent with frequent domestic travel that continued even after international routes were largely closed. The fraction of cases that were imported decreased over time (Fig. 2B), with the steepest decline during March (fig. S9), likely reflecting the expansion of existing local clades as the outbreak accelerated and travel restrictions were implemented. By April 2020, the vast majority of cases (median 90.7%, 89.2 to 91.9%, 95% CI) resulted from local populations, rather than from new importations (Table 1, Fig. 2B, and fig. S9).

The majority of cases in our dataset are associated with a minority of importation events; only 29% (26 to 32%, 95% CI) of importations

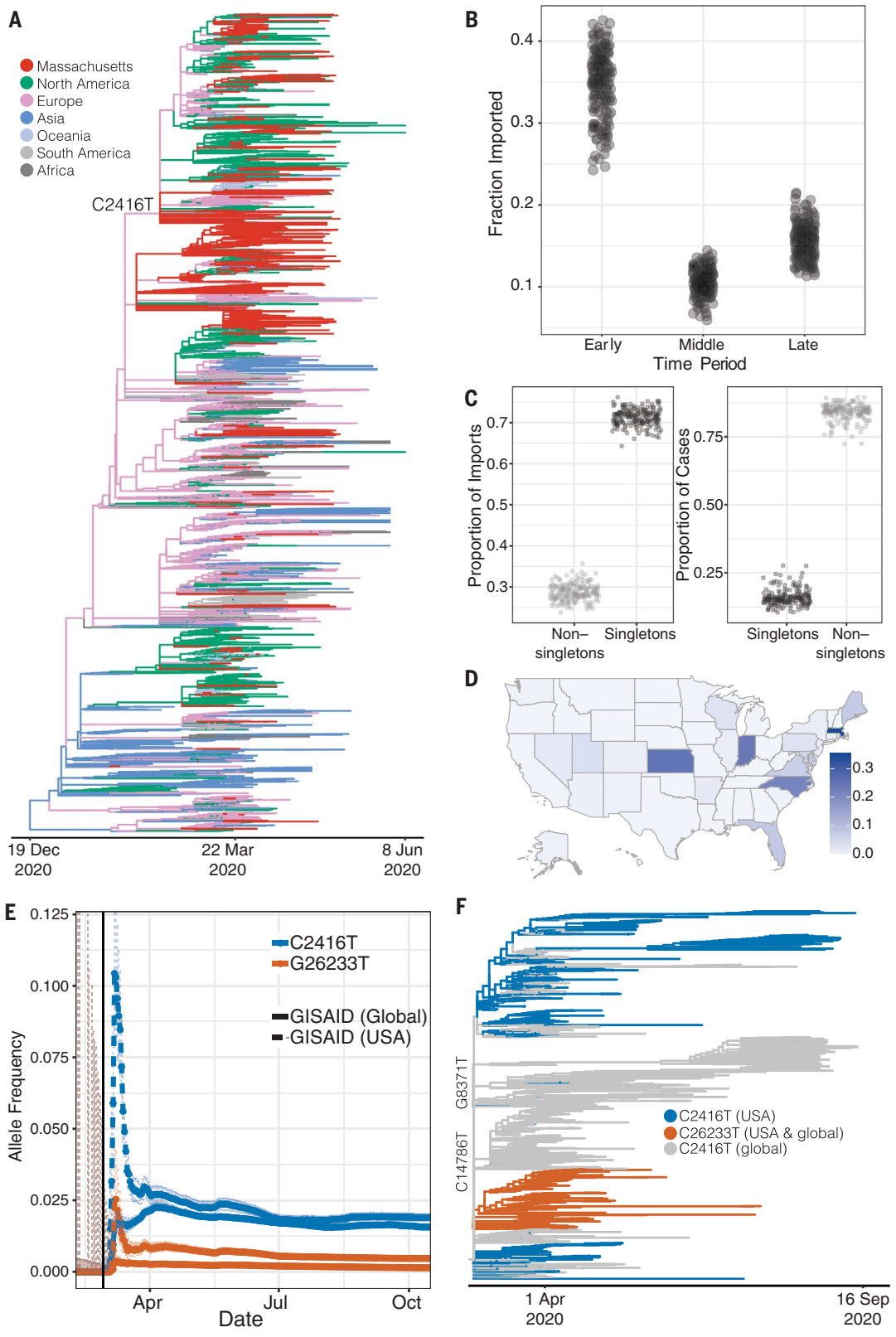


Fig. 2. Introductions of SARS-CoV-2 into Massachusetts. (**A**) Time tree of 772 Massachusetts genomes and a global set of 4011 high-quality genomes from GISAID. An interactive version of this tree and more information on specific subgroupings within the Massachusetts dataset is available at https://auspice.broadinstitute.org. (**B**) Proportion of genomes that were inferred as imported (ancestral state as not from Massachusetts) in the early (before 28 March 2020), middle (28 March to 14 April 2020), and late (after 15 April 2020) time periods of the Massachusetts epidemic. (**C**) The proportion of importation events and cases that were associated with singleton introductions (importation events associated with a single case in Massachusetts) into the Boston area over subsampled trees. (**D**) Allele frequency of the C2416T mutation by state. (**E**) Allele frequency of the C2416T and C26233T alleles in 159,043 GISAID samples reported through 17 October 2020. The vertical black line denotes the end of the business conference on 27 February. (**F**) Time tree of all sequences containing the C2416T variant collected before 30 September 2020.

involved more than one case, but those 29% accounted for 85% (78 to 88%) of the cases in our dataset (Fig. 2C and fig. S9C). As expected, early importation events resulted in large clades (fig. S9, B and C)—likely because of a combination of longer time to expand and unchecked spread before public health measures were implemented. Several clades established early in the Boston area showed continued community transmission throughout the study period (Table 2 and Fig. 3A), with the lineage containing C2416T, which is associated with a superspreading event early in the epidemic (described below), being the largest. The C2416T lineage was likely the first of these clades

imported into Boston [median estimated time to the most recent common ancestor (tMRCA), 14 February 2020; 95% highest posterior density (HPD) 4 to 20 February 2020) (Fig. 3B). The other four major lineages (G3892T, G105T, G28899T, and C20099T) appeared to enter the region between March and early April 2020. These major lineages, including the superspreading event–associated viruses, circulated widely in the Boston area (fig. S10). This included the communities of Chelsea, Revere, and Everett, which were among the most deeply affected in the state (fig. S11). Consistent with a larger global trend (*14*, *15*), we observed a rise in frequency of viruses harboring the D614G amino acid polymorphism in the Spike protein, conferred by a SNP at nucleotide 23,403 in the Wuhan reference strain, which rose to near-fixation in our dataset by the end of the study period (Fig. 3C) and is present in all of the dominant lineages.

On the basis of tMRCA estimates for the major Boston-area clades, we did not find evidence of cryptic transmission in the region before mid-February, and none of the importation events we inferred (Table 1) occurred before known cases. However, because testing for SARS-CoV-2 in Massachusetts was restricted to a narrow definition before established community spread (*16*), we cannot rule out the possibility that isolated importation events and small outbreaks may have escaped detection with the current resolution of sampling.

Spread of SARS-CoV-2 at an international business conference

Sustained local transmission of SARS-CoV-2 in the Boston area was first detected in early March, and with it, case clusters began to appear. The first large cluster was recognized in the context of an international business conference held in Boston from 26 to 27 February (*8*). Public health investigation with contact tracing identified approximately 100 cases associated with this conference (*17*), raising suspicion that a superspreading event had occurred there. We sequenced SARS-CoV-2 genomes from 28 of these cases. These genomes indeed showed the signature of superspreading: They form a tight phylogenetic cluster of highly similar viruses within a narrow time window.

All 28 conference-associated genomes were collected between 5 and 11 March and form a well-supported monophyletic cluster (posterior probability > 0.99) (Fig. 3A and fig. S12) marked by the presence of the SNP C2416T (Fig. 3A). The parent lineage of C2416T, defined by G25563T, was widely distributed in Europe in January and February 2020. The C2416T variant can serve as a marker for tracking the spread of SARS-CoV-2 from the conference, within Massachusetts and the United States; it is first reported in the United States in patients associated with the conference, and there is no evidence that it had entered the country independent of its appearance there. In our dataset, all C2416T-containing viruses collected before 10 March were sampled from individuals with conference exposure, and it was not seen in other publicly available genome data from cases anywhere in the United States before 7 March, when it appeared in cases that were also likely associated with the conference (*18*). Before that, it is seen in the global GISAID database in only two French patients, ages 87 and 88, on 29 February 2020 (Fig. 2E). The estimated tMRCA for C2416T-containing genomes is 14 February (95% HPD 4 to 20 February). Taken together, this strongly suggests low-level community transmission of C2416T in Europe in February 2020 before the allele came to Boston via a single introduction, which was then amplified by superspreading at the conference.

We also identified a second variant, G26233T, with a strong conference association. Evidence suggests that G26233T emerged during (or theoretically, immediately after) the conference because it was first seen in 7 of 28 individuals with known conference exposure, including in one sample at intermediate frequency (26%). It is not seen elsewhere in any public genome databases before cases associated with the conference (Figs. 2E and 3C). The presence of these two genetic signatures—C2416T in all conference-associated genomes in our dataset, and G26233T in a subset of them, with little or no evidence of transmission before the conference—provide markers to track the onward spread of SARS-CoV-2 from the event (Fig. 2F).

The conference-associated lineage was the most common one in our dataset, with C2416T representing 35% (261 of 744) and C2416T/G26233T representing 20% (151 of 744) of genomes (excluding those known to be directly associated with the conference). SARS-CoV-2 containing the C2416T allele spread extensively

Table 1. Estimate of SARS-CoV-2 introductions into Massachusetts. Results of ancestral trait inference using a binary model (MA versus non-MA) and regional model (regional geographic categories) are shown. 95% CIs are shown in parentheses and derived from subsampling the database of global strains (materials and methods).

Region	Before 28 March	28 March to 15 April	After 15 April
Binary model			
Non-MA	76 (61 to 86)	40 (33 to 46)	28 (23 to 33)
Regional model			
Africa	0 (0 to 1)	0 (0 to 1)	0 (0 to 1)
Asia	2 (1 to 4)	0 (0 to 1)	1 (0 to 2)
Europe	11 (7 to 16)	6 (3 to 9)	2 (0 to 3)
North America	56 (43 to 66)	29 (22 to 34)	22 (17 to 28)
Oceania	0 (0 to 0)	0 (0 to 0)	0 (0 to 1)
South America	1 (1 to 1)	0 (0 to 0)	0 (0 to 0)

Table 2. Major Boston-area lineages identified by lineage-defining mutation.

Lineage	Root	C20099T	G3892T	C2416T	G105T	G28899T
Number of genomes	772	21	77	288	98	34
Epidemiology		BHCHP	SNF	Conference, BHCHP	BHCHP	
Amino acid substitution		ORF1b: A2211V; NSP15: A160V	ORF1a: E1209D; NSP3: E391D			N: R56I, ORF14: E56*
Median tMRCA (95% HPD)	15 December 2019 (20 November 2019 to 4 January 2020)	4 April 2020 (30 March 2020 to 8 April 2020)	19 March 2020 (13 March 2020 to 23 March 2020)	14 February 2020 (4 February 2020 to 20 February 2020)	10 March 2020 (1 March 2020 to 16 March 2020)	15 March 2020 (4 March 2020 to 21 March 2020)

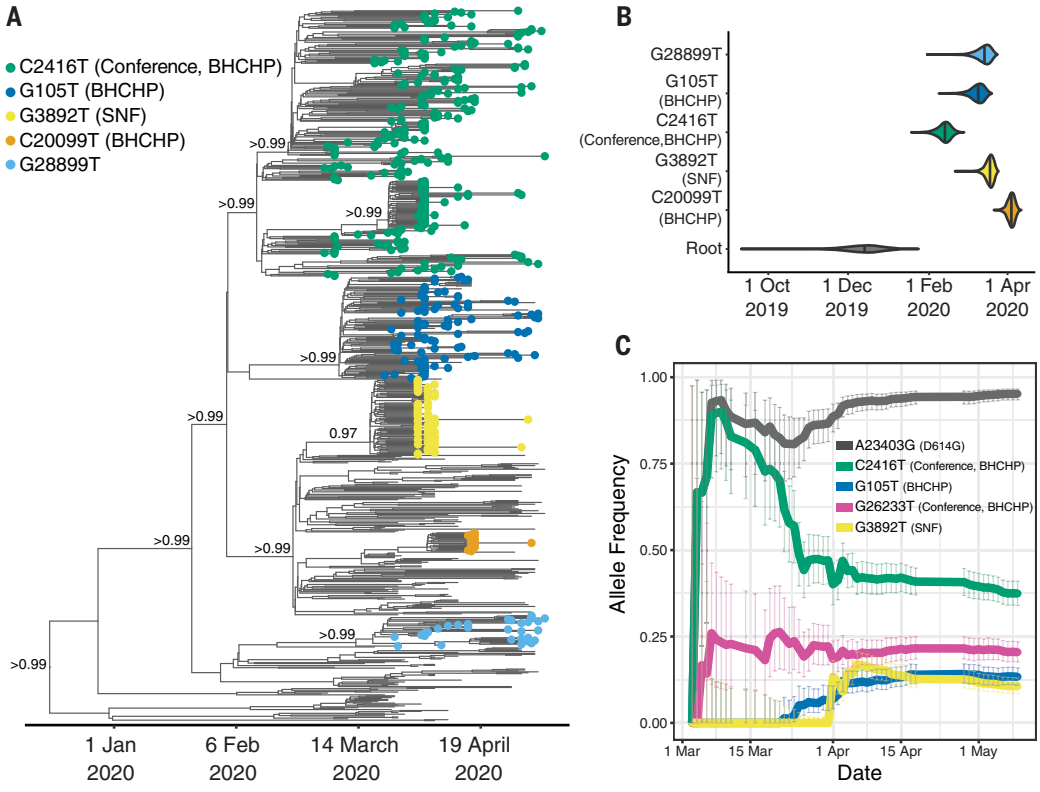


Fig. 3. SARS-CoV-2 spread in the Boston area. (**A**) Time-measured maximum clade credibility tree of 772 Massachusetts genomes with tips labeled by clade. Nodes with posterior support >0.8 are labeled. (**B**) Violin plots of tMRCA for the major Boston-area clades. (**C**) Estimated allele frequency in sequenced genomes over time for major Boston-area clades. Boston Healthcare for the Homeless Program, BHCHP; skilled nursing facility, SNF; international business conference, Conference.

in the Boston area (Fig. 3C and fig. S10A), accounting for between 30 and 46% of genomes from the four counties that make up the Boston area; by the end of the study period, these four counties had reported 51,718 cases. The allele was already at high frequency by the time it became clear that an epidemic was underway in the region (fig. S13B), establishing the conditions for extensive spread within Massachusetts and elsewhere.

C2416T began to appear in multiple other U.S. states in early March and increased rapidly in frequency (Fig. 2D and figs. S14 and S15). The effect of this spread was long-lasting. By 1 November 2020, viruses containing C2416T could be found in 29 states (fig. S15), and this lineage contributed 1.9% (675 of 35,566) of all U.S. SARS-CoV-2 genomes in GISAID. States with the largest numbers of cases included ones with known travel by or reported epidemiological links to conference participants returning from the meeting, including Florida, (125 of 1552 genomes contain C2416T), North Carolina (20 of 94 genomes) (*19*), and Indiana (10 of 42 genomes) (fig. S15A) (*20*).

Two additional lines of evidence suggest that the conference superspreading event in Boston contributed substantially to the spread of C2416T outside Massachusetts. First, the C2416T/G26233T sublineage, which arose in the context of the conference, was exported from Boston to at least 18 U.S. states as well as to other countries, including Australia, Sweden, and Slovakia (Fig. 2, D and F, and fig. S14A), with evidence of community spread in many places (fig. S15, C, D, and K). Second, there is evidence from other nonconference-associated C2416T sublineages that additional importations from Europe were not major contributors to C2416T prevalence in the United States. Two sublineages (C2416T/G8371T and C2416T/G20578T) appear frequently among European SARS-CoV-2 genomes in GISAID (295 genomes and 312 genomes, respectively) but are extremely rare among genomes from the United States (four and one genomes, respectively) (fig. S14, B and C). This evidence, along with the epidemiological data connecting multiple conference-linked cases to other U.S. states (*19*–*22*), suggests that most C2416T viruses in the United States likely derive from this initial introduction.

Genome data reveal that the impact of the conference was far larger than the approximately 100 cases directly associated with the event. Using state-reported case counts, we estimate that by the end of the study period, approximately 50,000 diagnosed cases (44,000 to 56,000) in the United States resulted from conference-associated viruses; of these, 46% (40.4 to 51.8%) were in Massachusetts. We estimate that through 1 November 2020, a total of 245,000 (205,000 to 300,000) cases marked by C2416T and 88,000 (56,000 to 139,000) cases marked by G26233T were linked to the conference in the United States. Although Massachusetts accounted for most early spread related to the conference, Florida accounted for the greatest proportion of cases overall [29.2% (22.8 to 36.0%)] (fig. S15G).

Although we have attempted to adjust for geography (by using state-level data) and time period as potential confounders, the accuracy of these estimates is limited by the available data: (i) GISAID is not a random sample of the U.S. epidemic, leading to unknown biases in the estimates; (ii) existing state-level data are too sparse for detailed spatiotemporal modeling; (iii) we have omitted states with 10 or fewer available genomes, leading to possible underestimation; (iv) diagnosed cases substantially underestimate true incidence (*23*); and (v) the estimates do not account for subsequent transmission of the virus (for example, 4 million new infections in the United States in November 2020). Although these estimates are provisional, they convey the

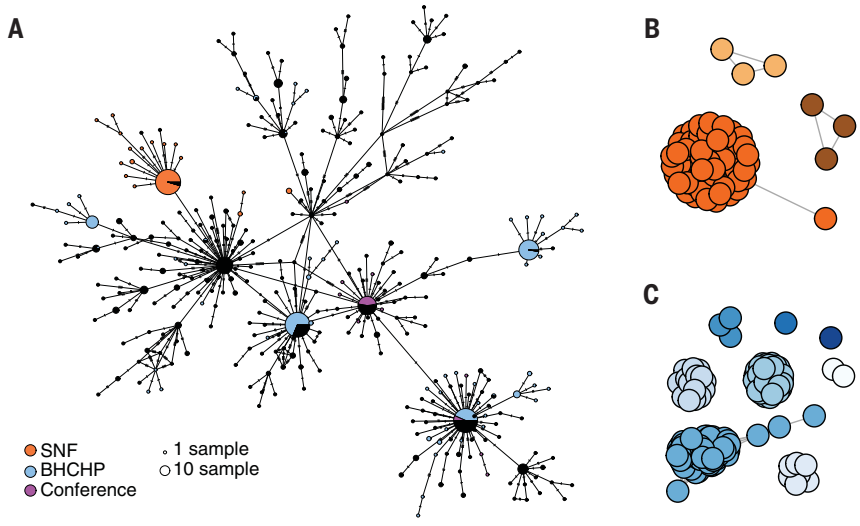


Fig. 4. SARS-CoV-2 superspreading events. (**A**) Minimal spanning network showing genetic similarity of SARS-CoV-2 genomes in the Massachusetts dataset, with genomes from major known superspreading events highlighted. (**B** and **C**) Gene graphs showing clusters of highly similar sequences among viral genomes from the (B) SNF and (C) BHCHP cohorts. Sequences are clustered when they are separated by less than four SNPs, and the lengths of lines between points reflect genetic distance.

likely scope of regional, national, and international spread resulting from a single superspreading event early in the pandemic.

Spread of SARS-CoV-2 in a SNF

We investigated a second large cluster of cases, this time at a SNF in the Boston area, that also proved to involve a superspreading event. The cluster was discovered accidentally: Screening of residents before a planned relocation in early April revealed widespread infection, and ultimately 85% (82 of 97) of the residents and 37% (36 of 97) of the staff (*24*) tested positive for SARS-CoV-2, even though none were known to be symptomatic when screening began. From these individuals, we assembled 83 SARS-CoV-2 genomes, 75 of which were found to compose a single cluster, part of the G3892T lineage described above (Fig. 3A). There was very little genetic variation within the cluster, and 59 of the genomes were identical (Fig. 4A), which is suggestive of a superspreading event. The estimated tMRCA for the cluster of 20 March (95% HPD: 13 to 24 March 2020) (Fig. 3B), along with the high proportion (30 of 45) of residents who tested negative on 1 April 2020 but were found to be positive 5 days later (*24*), suggests rapid spread within the facility in late March and early April 2020. Like other outbreaks reported from nursing facilities, the mortality rate was high. Although spread outside the facility appeared rare—only 1% (2 of 194) of samples in our dataset after 15 April 2020 harbored G3892T—24 residents who tested positive for SARS-CoV-2 died within 2 weeks of testing.

In addition to the major cluster, another one to two small clusters can be seen among the patients and staff in the SNF (95% HPD two to three total importations) (Fig. 4, A and B, and fig. S16). The different outcome of the introductions—one leading to massive spread within the facility (90% of sampled genomes) and the other(s) to little spread (10% of sampled genomes)—illustrates how superspreading can dramatically affect the transmission dynamics of SARS-CoV-2 and how under the right circumstances it can amplify the effect of any given introduction and associated lineage. These introductions occurred despite infection control policies—including a restriction on visitors (*25*), universal masking for all staff, masking for all residents when leaving their rooms, and vigilance with hand hygiene—in place for at least 2 weeks before the first detected infection (*24*).

Upon examination, we concluded that the genetic diversity in the main SNF cluster was very low even under the assumption of recent transmission from a single source. The 18 mutations seen in the cluster are significantly fewer than expected on the basis of the conference cluster (P = 0.019), which occurred over a similarly short time window, and much lower than the ~32 mutations expected under a simple model of SARS-CoV-2 substitution (P = 0.009) (materials and methods). This discrepancy might have resulted from low diversity in the SNF index patient, but it may also hint that heterogeneous mechanics of superspreading were at work in the two events. For example, if more virions than usual were transmitted from the SNF index patient to each secondary case—such as through unusually close or prolonged contact, or the initial case having a very high viral load at the time—then we would expect that the resulting infections would more often have the same consensus genome as that of the index case.

Cluster investigations in other close-contact settings

We studied several additional case clusters with the goal of providing viral genomic data to support public health investigations. These included potential transmission in homeless shelters and within a hospital. First, we analyzed the introduction and spread of SARS-CoV-2 among guests and staff at homeless shelters affiliated with the Boston Health Care for the Homeless Program (BHCHP). We produced 193 complete genomes from 314 samples collected in March and April 2020, including those collected during universal screening at Boston's largest homeless shelter (*5*). On the basis of the position of these 193 SARS-CoV-2 genomes from BHCHP in the overall Boston-area tree (Fig. 3A), we identified at least 14 introductions into the BHCHP community (95% HPD 14 to 18). Of these, four resulted in clusters consistent with superspreading, each containing 20 or more highly similar viral genomes (Fig. 4, A and C, and fig. S16B). Two of the clusters descended from the conference-associated C2416T lineage, including one that contained C2416T/G26233T. In total, 54% (105 of 193) of the genomes in this cohort contained C2416T, of which half (54 of 105) also contained G26233T, demonstrating that BHCHP guests and staff were affected by community transmission that resulted from amplification and spread of conference-associated SARS-CoV-2.

The other two case clusters occurred at Massachusetts General Hospital, where the Infection Control Unit sought genomic data to inform their investigations of possible nosocomial outbreaks. In the first cluster, two patients in the same hospital ward tested positive for SARS-CoV-2 during their hospital stay, after testing negative at the time of admission. In the second, unrelated cluster, four patients who received care in a specialty ward were diagnosed with SARS-CoV-2 infections over a period of several days. For each cluster, complete genomes (two of two from the first cluster and four of four from the second cluster) were genetically very distinct, a pattern that is inconsistent with having been infected from the same source during hospitalization (fig. S17). Although we cannot exclude the possibility of nosocomial transmission per se, because independent introductions from multiple asymptomatic staff could theoretically have occurred, this demonstrated that the individuals in each cluster were not part of the same transmission chain.

Conclusions

Genomic analysis of the first wave of the COVID-19 outbreak in the Boston area provides

powerful evidence of the importance of superspreading events in shaping the course of this pandemic. In this study, we show that importation events occurred very frequently—we identified more than 120 independent introductions during the 3-month study period—and that they varied enormously in terms of their subsequent impact on local transmission. Consistent with an overdispersed offspring distribution for SARS-CoV-2 (*26*), in our dataset, a small minority of importations accounted for the majority of observed cases. At least some of this variation in clade sizes results from superspreading events amplifying some lineages and not others. This can be seen in microcosm in one of the two superspreading events we studied in detail: SARS-CoV-2 was introduced at least twice into the skilled nursing facility; one introduction led to widespread transmission and numerous deaths, whereas the other one or two introductions led to a total of six cases.

The other superspreading event, which occurred at an international business conference early in the local epidemic, had a much greater impact on community transmission. Because SARS-CoV-2 viruses circulating at the conference happened to be marked by distinct genomic signatures, we were able to track its downstream effects far beyond the superspreading event itself, tracing the descendants of the virus as they made a large contribution to the local outbreak in the Boston area and as they spread throughout the United States and the world, likely causing hundreds of thousands of cases. The different genetic diversity seen in the two events raises the possibility that superspreading encompasses varied transmission dynamics.

Not all case clusters were the result of superspreading. Both hospital clusters consisted of unrelated cases that happened to occur in close proximity to one another. Cases associated with the homeless shelters likely resulted from a mix of superspreading events and more general transmission, although we lack the detailed epidemiological data to explore their history in depth. Where we were able to study superspreading events in detail, in the SNF and the conference, it was not because they were distinctive in size or character but because circumstances allowed close study. For both, we had dense sampling during a narrow time window of a clearly demarcated exposed population, aided by good data on prevailing genetic diversity to provide context.

Our findings highlight the close relationships between seemingly disconnected groups and populations: Viruses from international business travel seeded major outbreaks among individuals experiencing homelessness; spread throughout the Boston area, including to other higher-risk communities; and were exported to other domestic and international sites. It also illustrates the role of chance in the trajectory of an epidemic: A single introduction had an outsize effect on subsequent transmission because it was amplified by superspreading in a highly mobile population very early in the outbreak, before many public health precautions were put in place and when its effects would be further amplified by exponential growth and subsequent superspreading events (such as among the homeless). By contrast, other early introductions led to very little onward transmission, and the superspreading event in the SNF, although devastating to the residents, had little large-scale effect because it occurred later and in a more isolated population. Although superspreading events among medically vulnerable populations, such as nursing home residents, have a larger immediate impact on mortality, our findings raise the possibility that—paradoxically—the implications may be greater, when measured as a cost to society, for superspreading events that involve younger, healthier, and more mobile populations because of the increased risk of subsequent transmission. With the possibility of vaccines that protect against disease but not infection, this consideration may be increasingly important. This study provides clear evidence that superspreading events may profoundly alter the course of an epidemic and implies that prevention, detection, and mitigation of such events should be a priority for public health efforts.

Materials and methods

Full details of experimental and computational methods can be found in the supplementary materials, materials and methods.

Sample and data collection

This study was approved by the Partners Institutional Review Board under protocol 2019P003305 and MADPH IRB 00000701. We obtained samples and selected metadata from the MGH Microbiology Laboratory and MADPH under a waiver of consent for viral genome sequencing. All samples were nasopharyngeal (NP) swabs that tested positive for SARS-CoV-2 by means of quantitative reverse transcription polymerase chain reaction (RT-PCR). Epidemiological data on exposure and geography were obtained from medical record review (MGH) or collected by the MADPH laboratory in the process of clinical testing. Samples included individuals with known exposures to suspected superspreading events and individuals where no possible exposures were known. We compared known information about these cases to publicly available daily and weekly data on cases of SARS-CoV-2 in Massachusetts for the period 1 January to 1 August (www.mass.gov/info-details/covid-19-response-reporting).

Viral sequencing and analysis

Total RNA was extracted from inactivated NP swabs and presence of virus was confirmed using an RT-qPCR assay detecting the N1 gene of the virus. Metagenomic sequencing libraries were prepared as previously described (*27*). Briefly, following DNase treatment to remove residual DNA and depletion of human ribosomal RNA (rRNA), cDNA was synthesized using random hexamer priming. Illumina sequencing libraries were prepared from cDNA and sequenced with 100-nucleotide paired-end reads.

We conducted all analyses using viral-ngs 2.0.21 on the Terra platform (app.terra.bio). All of the workflows named below are publicly available via the Dockstore Tool Registry Service (https://dockstore.org/organizations/BroadInstitute/collections/pgs). Code is also archived at doi:10.5281/zenodo.4306358 and doi:10.5281/zenodo.4306362. Briefly, samples were demultiplexed (demux_only workflow), filtered for known sequencing contaminants and SARS-CoV-2 genomes were assembled using a reference-based assembly approach (assemble_refbased) with the reference genome NC_045512.2. Following a stringent quality control and filtering, we identified a final set of 772 high-quality assemblies from unique individuals that was used for all subsequent analyses and deposited in GenBank and GISAID. We used R (*28*), Bioconductor (*29*), ggplot2, tidyverse (*30*), and ggtree (*31*) to clean and plot data and trees, and choroplethr to draw maps.

To detect the presence of 20 common respiratory viruses in sequenced samples, we used Kraken2 (*32*) implemented in the classify_single and merge_metagenomics workflows. A virus was determined to be present if more than 10 reads mapped to that species. Wherever possible, these co-infections were confirmed using the BioFire FilmAssay Respiratory Panel.

We constructed phylogenetic maximum likelihood (ML) and time trees with associated visualizations using the Augur pipeline (augur_with_assemblies) and SARS-CoV-2-specific procedures taken from github.com/nextstrain/ncov for our 772 genomes and a representative background set of 4,011 subsampled from the GISAID database on 15 June 2020. We separately constructed ML trees from trimmed alignments to estimate root-to-tip distances and obtain branch support for ML phylogenies. To estimate coalescence dates of major lineages we constructed Bayesian time-trees using BEAST 2.6.2 with a general time reversible substitution model with four rate categories drawn from a gamma distribution (GTR4G), a strict clock, coalescent exponential tree prior, a uniform [-inf, inf] prior for the clock rate, a 1/x [-inf, inf] prior for the coalescent exponential population size; and a laplace [-inf, inf] prior for the growth rate.

Ancestral state reconstruction

We used three orthogonal approaches to reconstruct the ancestral location of unsampled nodes:

(i) a ML approach using the augur pipeline, (ii) a maximum parsimony approach using the Narushima and Hanazawa method as implemented in the MPR function of the ape package in R, and (iii) a Bayesian approach using BEAST1.10.4. In each case, we use a binary classification of "MA" vs "non-MA" to identify nodes that represent a likely importation event into Massachusetts. Full details of each approach are provided in the supplementary materials, materials and methods.

Analysis of superspreading events

To estimate the number of cases linked to the conference we estimated the proportion of genomes with C2416T and C2416T/G26233T per state by multiplying the observed proportion in genomes reported in GISAID through 2 November 2020 by case counts reported in the New York Times COVID data repository (https://github.com/nytimes/covid-19-data). We summed across states using a Monte Carlo simulation (n = 10,000).

To show clustering within the SNF and BHCHP cases, we constructed a minimal spanning haplotype network from the trimmed ML alignment of 772 genomes using PopART v1.7 (*33*) with masking of regions where any sequence had ambiguous bases. Gene graphs were constructed using pairwise distance matrices computed on aligned SARS-CoV-2 genomes and clustered using the R package adegenet (*34*). Importations into the SNF and BHCHP populations were calculated using a Bayesian approach similar to that described above (see Supplementary Materials and Methods for more details).

We define a superspreading event as the transmission from a single source to a large number of secondary infections, where the number is large enough that it would occur <1% of the time in a simple Poisson model of transmission (*35*). For this study, using an R_{eff} value of 3.0, we set the threshold at a minimum of nine transmissions. We compared the number of mutations among conference-associated and SNF-associated genomes with the expected number based on a generation time of 5.0 days (*36*) and a mean substitution rate of 1.04×10^{-3}/bp/year (fig. S6C) and calculated a P value based on the fraction of draws yielding fewer mutations than observed.

REFERENCES AND NOTES

1. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, (JHU), COVID-19 Dashboard; https://coronavirus.jhu.edu/map.html.
2. M. A. Waltenburg *et al*., Coronavirus disease among workers in food processing, food manufacturing, and agriculture workplaces. *Emerg. Infect. Dis.* **27**, (2020). doi: 10.3201/eid2701.203821; pmid: 33075274
3. W. E. Wei *et al*., Presymptomatic Transmission of SARS-CoV-2—Singapore, January 23–March 16, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 411–415 (2020). doi: 10.15585/mmwr.mm6914e1; pmid: 32271722
4. T. M. McMichael *et al*., Epidemiology of COVID-19 in a long-term care facility in King County, Washington. *N. Engl. J. Med.* **382**, 2005–2011 (2020). doi: 10.1056/NEJMoa2005412; pmid: 32220208
5. T. P. Baggett, H. Keyes, N. Sporn, J. M. Gaeta, Prevalence of SARS-CoV-2 infection in residents of a large homeless shelter in Boston. *JAMA* **323**, 2191–2192 (2020). doi: 10.1001/jama.2020.6887; pmid: 32338732
6. L. Hamner *et al*., High SARS-CoV-2 attack rate following exposure at a choir practice—Skagit County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 606–610 (2020). doi: 10.15585/mmwr.mm6919e6; pmid: 32407303
7. A. James *et al*., High COVID-19 attack rate among attendees at events at a church—Arkansas, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 632–635 (2020). doi: 10.15585/mmwr.mm6920e2; pmid: 32437338
8. A. Schuchat, CDC COVID-19 Response Team, Public health response to the initiation and spread of pandemic COVID-19 in the United States, February 24–April 21, 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, 551–556 (2020). doi: 10.15585/mmwr.mm6918e2; pmid: 32379733
9. MA Department of Public Health, Man returning from Wuhan, China is first case of 2019 Novel Coronavirus confirmed in Massachusetts (2020); www.mass.gov/news/man-returning-from-wuhan-china-is-first-case-of-2019-novel-coronavirus-confirmed-in.
10. COVID-19 Response Reporting (Massachusetts Department of Public Health, 2020); www.mass.gov/info-details/covid-19-response-reporting.
11. D. J. Park *et al*., Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161**, 1516–1526 (2015). doi: 10.1016/j.cell.2015.06.007; pmid: 26091036
12. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017). doi: 10.1002/gch2.1018; pmid: 31565258
13. M. Worobey *et al*., The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020). doi: 10.1126/science.abc8169; pmid: 32912998
14. L. Yurkovetskiy *et al*., SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. bioRxiv 187757 [Preprint] 4 July 2020. .doi: 10.1101/2020.07.04.187757
15. B. Korber *et al*., Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e19 (2020). doi: 10.1016/j.cell.2020.06.043; pmid: 32697968
16. MA Department of Public Health, Update and interim guidance on outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China (2020); www.mass.gov/clinical-advisory/update-and-interim-guidance-on-outbreak-of-2019-novel-coronavirus-2019-ncov-in.
17. MA Department of Public Health, Coronavirus Disease 2019 (COVID-19) Cases in MA, March 15 2020 (2020).
18. The U.S. Centers for Disease Control and Prevention (CDC) sequenced 19 Massachusetts genomes before 8 March 2020; 17 of 19 cases (89%) contained C2416T. The CDC Massachusetts genomes are not annotated with exposure information, but given official MADPH data reporting that 23 of 28 cases as of 8 March 2020 were linked to the conference (*20*), and the five nonconference associated cases include the travel-associated cases from this time period (MA-1, DPH_00002, and DPH_00003) and one from the Berkshire county cluster (all of which lack C2416), it can be inferred that a minimum of 16 or 17 C2416T-containing samples sequenced by the CDC were conference-associated.
19. North Carolina Department of Health and Human Services, Five more people in north carolina test positive for COVID-19 (2020); www.ncdhhs.gov/news/press-releases/five-more-people-north-carolina-test-positive-covid-19.
20. Indiana State Department of Health, State health department announces 2nd COVID-19 case (2020); https://calendar.in.gov/site/isdh/event/isdh-news-release-state-health-department-announces-2nd-covid-19-case.
21. Tennesse Department of Health, TDH releases further information regarding COVID-19 case (2020); www.tn.gov/health/news/2020/3/5/tdh-releases-further-information-regarding-covid-19-case.html.
22. Indiana State Department of Health, State health department confirms 1st case of COVID-19 in Hoosier with recent travel (2020).
23. H. Reese *et al*., Estimated incidence of COVID-19 illness and hospitalization—United States, February–September, 2020. *Clin. Infect. Dis.* ciaa1780 (2020). doi: 10.1093/cid/ciaa1780; pmid: 33237993
24. S. A. Goldberg *et al*., Presymptomatic transmission of severe acute respiratory syndrome coronavirus 2 among residents and staff at a skilled nursing facility: Results of real-time polymerase chain reaction and serologic testing. *Clin. Infect. Dis.* ciaa991 (2020). doi: 10.1093/cid/ciaa991; pmid: 32667967
25. M. Bharel, Order of the Commisioner of Public Health, (2020); www.mass.gov/doc/march-15-2020-assisted-living-visitor-restrictions-order/download.
26. A. Endo, S. Abbott, A. J. Kucharski, S. Funk, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020). doi: 10.12688/wellcomeopenres.15842.3; pmid: 32685698
27. C. B. Matranga *et al*., Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014). doi: 10.1186/s13059-014-0519-7; pmid: 25403361
28. R. Ihaka, R. Gentleman, R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5**, 299–314 (1996).
29. R. C. Gentleman *et al*., Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004). doi: 10.1186/gb-2004-5-10-r80; pmid: 15461798
30. H. Wickham *et al*., Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019). doi: 10.21105/joss.01686
31. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T. Lam, ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017). doi: 10.1111/2041-210X.12628
32. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019). doi: 10.1186/s13059-019-1891-0; pmid: 31779668
33. J. W. Leigh, D. Bryant, POPART: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015). doi: 10.1111/2041-210X.12410
34. T. Jombart, I. Ahmed, adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011). doi: 10.1093/bioinformatics/btr521; pmid: 21926124
35. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreading and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005). doi: 10.1038/nature04153; pmid: 16292310
36. L. Ferretti *et al*., Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **368**, eabb6936 (2020). doi: 10.1126/science.abb6936; pmid: 32234805

ACKNOWLEDGMENTS

We gratefully acknowledge the microbiology laboratory staff and infection control personnel at MGH and MADPH and all members of the COVID-19 emergency response efforts at MGH, BHCHP, and MADPH. We also thank H. C. Metsky, A. Regev, A. I. Feller, and members of the CDC SPHERES consortium for valuable feedback and helpful discussions. This study was approved by the Partners Institutional Review Board under protocol 2019P003305 and MADPH IRB 00000701. We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID genetic sequence data on which this research is based (additional acknowledgment of 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, is provided in table S4). **Funding:** This work was sponsored by the National Institute of Allergy and Infectious Diseases (U19AI110818 to P.C.S; R37AI147868 and R01AI148784 to J.L.), the National Human Genome Research Institute (K99HG010669 to S.K.R.), the National Institute of General Medical Sciences of the National Institutes of Health (U54GM088558 to W.P.H.), the U.S. Centers for Disease Control and Prevention (U01CK000490; MGH), the Bill and Melinda Gates Foundation (Broad Institute), and the U.S. Food and Drug Administration (HHSF223201810172C), with in-kind support from Illumina as well as support from the Doris Duke Charitable Foundation (J.E.L.), the Howard Hughes Medical Institute and Merck KGaA Future Insight Prize (P.C.S.), the Herchel Smith Fellowship (K.A.L.), and the Evergrande COVID-19 Response Fund Award from the Massachusetts Consortium on Pathogen Readiness (J.L.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes

of Health. **Author contributions:** K.J.S., B.M.S., A.G.-Y., G.A., T.F., K.C.D., M.R., M.R.B., K.A.L., E.N., S.K.R., and A.G. guided and/or performed laboratory experiments and prepared samples for sequencing. J.E.L., K.J.S., C.L., S.F.S., C.H.T.-T., L.A.K., S.C., W.P.H., D.J.P, and B.L.M. performed data management, processing, and/or analysis. S.C., A.C., M.E.K., C.C., K.F., A.N., and F.C. provided project management. J.E.L., D.H., J.M.G., T.P.B., J.O., E.T.R., S.E.T., R.C.L., G.R.G., L.C.M., S.S., V.M.P., and E.R. oversaw research activities at clinical or public health sites and provided study guidance. M.N.A., A.E.L., C.M., M.F., D.S., J.B.H., J.A.B., A.G., T.D.L., A.P., M.B., C.M.B., J.L., and W.P.H. provided critical insights and/or study guidance. J.E.L, K.J.S., C.L., S.F.S., P.C.S., D.J.P., and B.L.M. oversaw study design, implementation, analysis, and drafted and revised the manuscript. All authors contributed to interpreting results and reviewing the manuscript. **Competing interests:** J.E.L. has received consulting fees from Sherlock Biosciences. J.A.B. has been a consultant for T2 Biosystems, DiaSorin, and Roche Diagnostics. A.P. is a venture partner at Google Ventures. P.C.S. is a cofounder and shareholder of Sherlock Biosciences and a Board member and shareholder of Danaher Corporation. **Data and materials availability:** Sequences and genome assembly data are publicly available in the Broad Institute's Terra platform (https://terra.bio) in a featured workspace for COVID-19 (https://app.terra.bio/#workspaces/pathogen-genomic-surveillance/COVID-19_Broad_Viral_NGS). Researchers can use this workspace to reproduce analyses described here or perform similar analyses on their own viral sequence data. Assembled genomes and raw metagenomic reads from this dataset have been deposited at NCBI's GenBank and SRA databases under BioProject PRJNA622837 in accordance with NIAID's Data Sharing policy (www.niaid.nih.gov/research/data-sharing-and-release-guidelines) and will soon be available to visualize on nextstrain.org/ncov. Experimental protocols are publicly available on Benchling and can be accessed here: https://benchling.com/sabetilab/f_/gaLGu5X9-sabeti_group_sars-cov-2_metagenomic_sequencing_protocols.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/371/6529/eabe3261/suppl/DC1
Materials and Methods
Figs. S1 to S17
References (*37*–*44*)
Tables S1 to S4
MDAR Reproducibility Checklist

19 August 2020; accepted 7 December 2020
Published online 10 December 2020
10.1126/science.abe3261