**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.

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.); brownwy@broadinstitute.org (B.L.M.)
‡These authors contributed equally to this work.

This is an open-access article distributed under the terms of the Creative Commons Attribution license (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

Jacob E. Lemieux1,2,†, Katherine J. Siddle3,‡, Bennett M. Shaw1,†, Christine Loreth3, Stephen F. Schaffner1,3,‡, Adrienne Gladden-Young4, Gordon Adams5, Timelia Fink5, Christopher H. Tomkins-Tinch5, Lydia A. Krasilnikova5,‡, Katherine C. DeRuff5, Melissa Rudy5, Matthew R. Bauer1,†, Kim A. Lagerborg1,§, Erica Normand1,§, Sinéad B. Chapman1, Steven K. Reilly1,§, Melis N. Anahtar8, Aaron E. Lin1,3, Amber Carter1, Cameron Myhrvold1,3,27, Molly E. Kemball1,7, Jeremy Luban1,17,18, Edward T. Ryan2,4,15, Sarah E. Turbett2,8,15, Regina C. LaRocque2,15, Matthew R. Bauer1,6, Kim A. Lagerborg1,6, Erica Normand1,§, Sinéad B. Chapman1, Steven K. Reilly1,§, Melis N. Anahtar8, Aaron E. Lin1,3, Amber Carter1, Cameron Myhrvold1,3,27, Molly E. Kemball1,7, Jeremy Luban1,17,18, Edward T. Ryan2,4,15, Sarah E. Turbett2,8,15, Regina C. LaRocque2,15, William P. Hanage19, Glen R. Gallagher1,‡, Lawrence C. Madoff5,20,§, Sandra Smole5,‡, Virginia M. Pierce6,21,22,‡, Eric Rosenberg2,23,§, Pardis C. Sabeti1,3,4,18,23,§, Daniel J. Park1,‡, Bronwyn L. MacInnis1,4,18,23,§.

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 18,000 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 772 complete SARS-CoV-2 assemblies (>98% complete) from 772 individuals, and an additional 72 partial genomes (>80%
Fig. 1. Epidemiology of SARS-CoV-2 in Massachusetts and of sequenced viral genomes. (A) Cumulative confirmed and presumed cases reported statewide in Massachusetts (JD) 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.

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). 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 (I3) 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
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

![Image](https://www.science.org)
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 1,552 genomes contain C2416T), North Carolina (20 of 94 genomes) (I9), 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. S14, 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 in November 2020). 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...
SARS-CoV-2 genomes in the Massachusetts dataset, with genomes from major known superspreading events ultimately 85% (82 of 97) of the residents and early April revealed widespread infection, and of residents before a planned relocation in 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 (26), 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 (3). 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 resulted 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 communities 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 MAPH IRB 00000701. We obtained samples and selected metadata from the MGH Microbiology Laboratory and MAPH 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 MAPH 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 choropleth to clean and 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 FilmArray 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/nov 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 (GTR+G), 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:
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 (39) 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$^{-4}$/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. Walthenburg et al., Coronavirus disease among workers in food processing, food manufacturing, and agriculture workplaces. Emerg. Infect. Dis. 27, doi:10.3201/ eid2701.200101
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.mm6916e2; pmid:3227722
4. T. M. McMahan et al., Epidemiology of COVID-19 in a long-term care facility in King County, Washington. N. Engl. J. Med.
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., I.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 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. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.
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, Adrienne Gladden-Young, Gordon Adams-Tinch, Melissa Rudy, Matthew R. Bauer, Kim A. Lagerberg, Erica Normandin, Sinead B. Chapman, Steven K. Reilly, Anahtar Aaron E. Lin, Amber Carter, Cameron Myhrvold, Molly E. Kemball, Sushma Chaluvadi, Caroline Cusick, Katelyn Flowers, Anna Neumann, Felecia Cerrato-Math, Farhat Damien, Slater, Jason B. Harris, John A. Branda, David Hooper, Jessie M. Gaeta, Travis P. Baggett, James O’Connell, Andreas Ginirke, 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.

*Science*, 371 (6529), eabe3261.

**Phylogenetics of superspreading**
One important characteristic of coronavirus epidemiology is the occurrence of superspreading events. These are marked by a disproportionate number of cases originating from often-times asymptomatic individuals. Using a rich sequence dataset from the early stages of the Boston outbreak, Lemieux *et al.* identified superspreading events in specific settings and analyzed them phylogenetically (see the Perspective by Alizon). Using ancestral trait inference, the authors identified several importation events, further investigated the context and contribution of particular superspreading events to the establishment of local and wider SARS-CoV-2 transmission, and used viral phylogenies to describe sustained transmission.

*Science*, this issue p. eabe3261; see also p. 574