

Cite as: M. Worobey *et al.*, *Science* 10.1126/science.abc8169 (2020).

# The emergence of SARS-CoV-2 in Europe and North America

Michael Worobey<sup>1\*</sup>, Jonathan Pekar<sup>2,3</sup>, Brendan B. Larsen<sup>1</sup>, Martha I. Nelson<sup>4</sup>, Verity Hill<sup>5</sup>, Jeffrey B. Joy<sup>6,7,8</sup>, Andrew Rambaut<sup>5</sup>, Marc A. Suchard<sup>9,10,11\*</sup>, Joel O. Wertheim<sup>12\*</sup>, Philippe Lemey<sup>13\*</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. <sup>2</sup>Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093, USA. <sup>3</sup>Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093, USA. <sup>4</sup>Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA. <sup>5</sup>Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK.

<sup>6</sup>Department of Medicine, University of British Columbia, Vancouver, BC, Canada. <sup>7</sup>BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada. <sup>8</sup>Bioinformatics Programme, University of British Columbia, Vancouver, BC, Canada. <sup>9</sup>Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. <sup>10</sup>Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA 90095, USA.

<sup>11</sup>Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. <sup>12</sup>Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA. <sup>13</sup>KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Clinical and Evolutionary Virology, Leuven, Belgium.

\*Corresponding author. Email: worobey@arizona.edu (M.W.); msuchard@ucla.edu (M.A.S.); jwertheim@health.ucsd.edu (J.O.W.); philippe.lemey@kuleuven.be (P.L.)

**Accurate understanding of the global spread of emerging viruses is critically important for public health responses and for anticipating and preventing future outbreaks. Here, we elucidate when, where and how the earliest sustained SARS-CoV-2 transmission networks became established in Europe and North America. Our results suggest that rapid early interventions successfully prevented early introductions of the virus into Germany and the US from taking hold. Other, later introductions of the virus from China to both Italy and to Washington State founded the earliest sustained European and North America transmission networks. Our analyses demonstrate the effectiveness of public health measures in preventing onward transmission and show that intensive testing and contact tracing could have prevented SARS-CoV-2 from becoming established.**

In late 2019 the emergence of SARS-CoV-2, which causes COVID-19, ignited a pandemic that has been associated with over 500,000 deaths globally as of July, 2020. As the original outbreak in Hubei province, China, spilled into other countries, containment strategies focused on travel restrictions, isolation, and contact tracing. Given the virus's exponential growth rate, delaying the onset of community transmission by even a few weeks likely bought government officials valuable time to establish diagnostic testing capacity and implement social distancing plans.

Viral genetic sequence data can provide critical information about whether viruses separated by time and space are likely to be epidemiologically linked. Genomic data have suggested differences in the timing, spatial origins and transmission dynamics of early SARS-CoV-2 outbreaks in multiple North American locations, including Washington State (1, 2), the East Coast of the US (3, 4), California (5) and British Columbia (5, 6). The first confirmed US case was associated with a virus ("WA1") isolated in Washington State from a traveler who returned from Wuhan, China on January 15th, 2020 (7). No onward transmission was detected after extensive follow-up in what appeared to be successful containment of the country's first known incursion of the virus (8). However, subsequent identification of viruses that were genetically

similar to WA1, first in Washington, then in Connecticut (3), California (5), British Columbia (6) and elsewhere, raised the possibility that WA1 had actually established chains of cryptic transmission that started on January 15th and went undetected for several weeks (1, 2). If true, this introduction would predate early SARS-CoV-2 community transmission chains documented elsewhere on the continent (3–5) and establish the Seattle area as the epicenter of the North American epidemic. Hence it is necessary to resolve this question to determine where the virus first initiated substantial community outbreaks, and whether the earliest coast-to-coast spread of the virus within the US (3) was from west-to-east or east-to-west.

In Europe, the first diagnosed case was an employee of an auto supplier who visited the company's headquarters in Bavaria, Germany, from Shanghai, China, on January 20th, 2020 (9). She had been infected with SARS-CoV-2 in Shanghai (after her parents had visited from Wuhan) (10) and transmitted the virus to a German man who tested positive on January 27th (11) and whose viral genome ("BavPat1") was sampled on January 28th (10). All told, the outbreak infected 16 employees but was apparently contained through rapid testing and isolation (9). Italy's first major outbreak in Lombardy, which was apparent by ~February 20th, 2020, was

associated with viruses closely related to BavPat1, but in a separate lineage designated “B.1,” differing from BavPat (a lineage “B” virus) by just one nucleotide in the nearly 30,000 nucleotide genome. A narrative took hold that the virus from Germany had not been contained but had been transmitting undetected for weeks and had been carried to Italy by an infected German (9, 12). In addition to igniting a devastating outbreak in Italy, this B.1 lineage subsequently spread widely across Europe and beyond, initiating outbreaks in many countries including the intense one in New York City (13, 14). Greater clarity about the effectiveness of Germany’s early contact tracing efforts has implications for the feasibility of controlling the virus through nonpharmaceutical interventions.

There are a number of limitations in phylogenetic and spatial inferences drawn from SARS-CoV-2 genomic data. The SARS-CoV-2 virus has a relatively long (~29 kB) positive-sense ssRNA genome that evolves at a rate less than  $1 \times 10^{-3}$  substitutions/site/yr, amounting to ~2 substitutions per genome per month. The rate is slower than most RNA viruses owing to the proofreading activity encoded by the non-structural gene *nsp14* (15). Consequently, the entire global population of SARS-CoV-2 viruses through March, 2020 differed by only 0–12 nucleotide substitutions from the inferred ancestor of the whole pandemic. Transmission clusters tend to be defined by 1–3 nucleotide differences across the entire viral genome. Phylogeographic inferences are further confounded by the relatively low availability of genomic sequence data from locations that experienced early outbreaks, including Italy, Iran and the original epicenter in Hubei. The combination of the relatively slow rate of SARS-CoV-2 evolution, its rapid dissemination within and between locations, and unrepresentative sampling risks serious misinterpretation.

Here, we investigate fundamental questions about when, where and how the SARS-CoV-2 virus established itself globally. We integrate multiple sources of information into phylogenetic inferences, including airline passenger flow data between potential sources and destinations of viral dispersals early in the pandemic, as well as disease incidence data in Hubei province and other locales that likely impacted the probability of infected travelers moving the virus around the globe. By combining a genomic epidemiology approach, which aims to account for the effects of undersampling viral genetic diversity in the epicenter of the pandemic, with consideration of expected evolutionary patterns for a novel pathogen with low diversity, we resolve key questions about how and when the SARS-CoV-2 pandemic unfolded in Europe and North America.

### **Emergence of SARS-CoV-2 in the US**

A key turning point in the US outbreak occurred when researchers sequenced the first viral genome recovered from a

putative case of community transmission in the US (“WA2,” sampled in the Seattle area on February 24th, 2020), reporting on February 29th that it was similar to WA1, the viral variant from the first-diagnosed COVID-19 patient (1). This led to the suggestion that WA1 might have established cryptic transmission in Washington State in mid-January (1). (The researchers did however acknowledge the possibility of an independent introduction of WA2 separate from WA1). This finding fundamentally altered the picture of the SARS-CoV-2 situation in the US, playing a decisive role in Washington State’s early adoption of intensive social distancing efforts. This, in turn, appeared to explain Washington State’s relative success in controlling the outbreak, compared with states that delayed, such as New York.

The availability of hundreds of SARS-CoV-2 genomes sampled in Washington State by mid-March revealed that WA2 belonged to a large, monophyletic clade of “A.1” lineage viruses that accounted for about 85% of cases in Washington State at that point, designated the “Washington State outbreak clade” (2) (hereafter “WA outbreak clade”). These data provided an opportunity to investigate whether the WA outbreak clade was initiated in mid-January by WA1 by simulating the epidemic under the constraint that it had been established by WA1 and then comparing observed evolutionary patterns with those expected under that scenario. A range of phylogenetic patterns could have been observed in this large sample (e.g., Fig. 1, A to C), yet were not (Fig. 1D).

To investigate whether the observed pattern of evolution reported in (1, 2) was consistent with the WA outbreak clade having descended from WA1, we simulated outbreaks using FAVITES (16) (fig. S1 and table S1). These simulated outbreaks had a median doubling time of 4.7 days (95% range across simulations: 4.2–5.1)—including so-called “super-spreading” events (fig. S2)—and a fixed evolutionary rate of  $0.8 \times 10^{-3}$  substitutions/site/year. A duration of two months (61 days) was chosen to reflect the time-period between WA1 and the implementation of disease mitigation efforts that would affect the median doubling time.

We examined the phylogenetic structure of maximum likelihood trees inferred from sub-sampled simulated viral sequences to determine how frequently they matched the observed relationship between WA1 and the WA outbreak clade. Specifically, a simulation tree matching the observed tree must produce a single branch emanating from WA1 that experiences at least two mutations (C17747T and A17858G in the observed tree) prior to establishment of a single outbreak clade (Fig. 2A). Alternative patterns include: (i) a virus identical to WA1 (Fig. 2B); (ii) a virus that differs from WA1 by a single mutation (Fig. 2C); (iii) a viral lineage forming a basal polytomy with WA1 and the outbreak clade (Fig. 2D); and (iv) a viral lineage that is sibling to the outbreak clade but experienced fewer than two mutations before divergence (Fig. 2E).

The frequency of alternative phylogenetic patterns in the simulated epidemics represents the probability that the true topology (Fig. 2A) could not have occurred if the WA outbreak clade had been initiated by WA1.

In 70.1% of simulations, we observed at least one virus genetically identical to WA1, with a median of 12 identical viruses in each simulation (95% range: 0–85 identical viruses) (Fig. 2). Not observing a virus identical to WA1 in the real Washington data does not significantly differ from expectation ( $p = 0.299$ ). However, viruses with one mutation from WA1 were observed in 95.5% of simulations, indicating a low probability of failing to detect even a single sequence from Washington within one mutation of WA1 ( $p = 0.045$ ). Lineages forming a basal polytomy with WA1 and the epidemic clade were observed in 99.7% of populations ( $p = 0.003$ ) and 100% of simulations had at least one sibling lineage diverging prior to experiencing two mutations and the formation of the outbreak clade ( $p < 0.001$ ). Therefore, even if C17747T and A17858G were linked—a possibility since they are both non-synonymous mutations located in the nsp13 helicase gene—we would still expect to see descendants of their predecessors in Washington prior to March 15th. In summary, when we simulated the Washington outbreak beginning with WA1 on January 15th, 2020 and sampled 294 genomes in the first two months of this outbreak, we failed to observe a single simulated epidemic that had the characteristics of the real phylogeny (Fig. 2). These findings were robust to simulations that used a slower epidemic doubling time of 5.6 days (95% range 5.2–5.9) or an accelerated substitution rate of  $1.6 \times 10^{-3}$  substitutions/site/year (16) (Supplementary text).

Although WA outbreak-related genomes lacking one or the other of the clade-defining substitutions C17747T and A17858G (Fig. 2, C and E) were absent in this initial large sample from Washington State, such genomes have been reported to be very common in nearby British Columbia (BC), Canada (Supplementary text). Indeed, genomes with the ancestral C17747 state constituted 16 of the first 27 WA outbreak-related genomes sequenced in BC and have been sampled occasionally at much lower frequency in several US states (3). Such a high frequency of these viruses in BC but not in Washington State raises the possibility that BC rather than Washington State was the site of introduction of the founding virus of this important lineage. Another possibility is that these BC genomes are descendants of a separate A.1 lineage introduction from China. The first scenario seems unlikely because of epidemiological evidence that the outbreak was larger in February and March in Washington State than in BC; the second scenario is unlikely because it would necessitate both introduced lineages to independently acquire the C17747T mutation.

We therefore considered a third hypothesis: that these 16 BC viral genomes contain a sequencing error at position 17747

and in reality bear the derived C17747T mutation. We reasoned that if this were the case, some of these genomes might share additional derived mutations with C17747/A17858G genomes sampled in the same location (i.e., they might be identical or highly similar except for a spurious C17747 base) (Supplementary text). As shown in Fig. 3, this is indeed the case: each of the 6 C17747 genomes from BC that contained one or more derived mutations at positions other than 17747 and 17858 shared one to four of these mutations with others sampled locally. Such a pattern is virtually impossible to explain through homoplasy events. Observing even one such homoplasy in a genome with more than 29,000 bases is rare; the probability of observing more than one is infinitesimally small. Similarly, the hypothesis that the C17747 state in these genomes is due to multiple, independent T17747C reversions is untenable. Occasional C17747 genomes from California, Oregon, Wyoming, Minnesota, Washington State and elsewhere also share derived mutations with viruses sampled in the same location (Fig. 3, table S2, Supplementary Text). Most of these genomes were generated using the amplicon-based “ARTIC Protocol,” and we speculate that mistaken incorporation of a primer sequence containing C17747 (“nCoV2019\_58\_RIGHT”) may be the cause.

When we investigated an exhaustive collection of genomes sampled in Washington State, including viruses sampled after March 15th related to the WA outbreak clade (Supplementary text), we detected a single virus, “WA-S566” sampled on March 29th, 2020, that lacked the derived C17747T and A17858G mutations found in the rest of the WA outbreak clade. The phylogenetic position of this virus matches the pattern in Fig. 2D, though it differs from WA1 at 7 additional sites. Hence, the observed pattern in this larger, and later, sample of approximately 1000 viral genomes reflects the scenario depicted in Fig. 1A rather than Fig. 1D. We therefore revisited our WA simulations, sampling 1000 genomes instead of the original 294, looking for instances in which more than two lineages diverged prior to the formation of the outbreak clade. In 88.8% of the simulations, we observed two or fewer basally divergent lineages and, therefore, cannot reject a scenario in which WA1 gave rise to only two lineages that diverge as a basal polytomy ( $p = 0.112$ ). However, in 99.0% of simulations, we do observe three or more divergent lineages before two mutations (i.e., lineages that experienced zero or one mutation from WA1 before diverging; fig. S3). Therefore, it is unlikely that if WA1 were the ancestral virus it would have given rise to only the S566 lineage and the WA outbreak clade ( $p = 0.010$ ). We must therefore take seriously the possibility of multiple introductions into the US of genetically similar viruses to explain the presence of S566 and the WA outbreak clade.

We thus turned to a distinct phylogeographic approach that explicitly considers the relatively late sampling time of

WA-S566, along with other temporal, epidemiological and geographic data. This method accounts for geographical gaps in sampling and integrates relevant covariates for global spatial spread in a Bayesian framework (16). We investigated how tree topologies were affected by the inclusion of unsampled viruses assigned to 12 of the most severely undersampled locations both in China and globally, based on COVID-19 incidence data (16). Realistic sampling time distributions also were inferred from COVID-19 incidence data. To better inform placement of unsampled viruses on the phylogeography, we adopted a generalized linear model (GLM) formulation of the phylogenetic diffusion process (17). This approach estimates a significant contribution for air passenger flow and asymmetric flow in and out of Hubei (both with Bayes factors > 8,042 and positive log effect sizes, Supplementary text).

The resulting phylogeny (Fig. 4) provides one reconstruction of the possible evolutionary relationships of WA outbreak viruses and their closest relatives that realistically accounts for major gaps in sequence data. For low-diversity data, a single phylogeny has a resolution that is, to a large extent, not supported by the full posterior tree distribution that contains several plausible phylogeographic scenarios that need to be considered, all of which are compatible with the genetic data (e.g., the mutation trees in ref. (2) and those available at [nextstrain.org](http://nextstrain.org)). The posterior maximum clade credibility (MCC) tree (Fig. 4) hypothesizes that the WA outbreak clade (plus S566 and a sibling virus sampled in New York, “NY”) resulted from an introduction from Zhejiang, China, supported by the clustering of sampled and unsampled taxa from this location. And while an introduction from a Chinese location other than Hubei yields considerable posterior support (bar chart inset in Fig. 4), Hubei is preferred over Zhejiang by the entire posterior sample as the most likely source for this introduction. Interestingly, although the genome from NY (near S566 in Fig. 4) is identical to WA1, its much more recent sampling time separates it from WA1 (and similarly early Chinese sampling) in the time-calibrated phylogeographic reconstruction. The more recent collection date for both this NY sample and S566, and the modest support (posterior probability [pp] = 0.67) that they share the US location state with the WA outbreak viruses, results in a reconstruction with a single introduction for these viruses. Using Markov jump estimates that account for phylogenetic uncertainty (18), we inferred February 1st, 2020 (95% HPD Jan. 14th–February 15th) as the time for this introduction, consistent with the observation that viruses from the WA outbreak clade were likely present during the voyage of the Grand Princess cruise ship to Mexico starting on February 11th (5).

Through a comparison with a time-inhomogeneous model, we show that our estimates are relatively robust to the

assumption of constant covariate effect sizes through time (fig. S4). Although the time-inhomogeneous model was fitted to a data set without unsampled viruses, it also provides strong support for an independent introduction from Hubei (fig. S5). Without unsampled taxa, we estimate a somewhat earlier date for the introduction of the ancestor of the WA outbreak clade plus S566 (January 26th, 2020 [95% HPD: January 15th–February 7th]), likely because the time-homogeneous analysis allows unsampled taxa from Hubei or other Chinese locations (as in the MCC tree in Fig. 5) to branch off closely to the WA outbreak clade. In the light of the travel restrictions, specifically from Hubei, the earlier mean date obtained without unsampled taxa may be the more realistic estimate.

We note that the MCC tree suggests that a Malaysian virus also descends from this introduction (i.e., that it resulted from a subsequent US to Malaysia jump). It is however much more plausible that this Malaysian virus was introduced directly from China to Malaysia, but both the sequence and covariate data in the phylogeographic model lack the information to strongly support this scenario. In light of the simulation results, we think there is a distinct possibility that S566 and the related NY virus may have descended from a separate introduction from Asia, with the site of arrival in the US unresolved based on the presence of both a West Coast and East Coast virus in the clade. Accordingly, an analysis that does not assign a known location to S566 and the related NY virus supports independent introductions from Hubei for these viruses and for the WA outbreak clade (fig. S6), with February 7 (95% HPD: January 23rd - February 18th) as the date for the latter.

Consistent with estimates of the introduction date of this viral lineage into Washington State, the Seattle Flu Study tested 6,908 archived samples from January and February, of which only 10, from the end of February, were positive (19). Our estimates of the introduction date of the WA outbreak clade into Washington State around the end of January or beginning of February, 2020 are approximately two weeks later than if the outbreak had originated with WA1’s arrival on January 15th (2), implying: (a) archived “self-swab” samples retrospectively detected the virus within a few weeks of its arrival (19), (b) this Washington State outbreak may have been smaller than estimates based on the assumption of a January 15th arrival of WA1, and (c) the individual who introduced the founding virus likely arrived in the US close to the initiation of the “Suspension of Entry” of non-US residents from China on February 2nd, 2020 (20) perhaps during the period when an estimated 40,000 US residents were repatriated from China, with screening described as cursory or lax (21). These passengers were directed to a short list of airports including Los Angeles, San Francisco, New York, Chicago, Newark, Detroit and Seattle (21). The timing of COVID-19

cases in Solano County and Santa Clara County in California later in February (5) (Supplemental text) suggest self-limited outbreaks may have originated from returning US residents during this period. So, although our reconstructions incorporating unsampled lineages do not account for travel restrictions, the remaining influx likely provided opportunity for a second introduction (distinct from WA1), or even multiple such introductions, into Washington State. Recent inferences that there have been more than 1,000 independent introductions of SARS-CoV-2 into the UK (22) lend support to this idea.

### **Early establishment of SARS-CoV-2 in Europe**

We used a similar approach to investigate whether the Northern Italy SARS-CoV-2 outbreak was introduced from the German outbreak or independently from China by simulating the Northern Italy outbreak under the hypothetical constraint that it was initiated by a virus imported from the German outbreak (fig. S7) and by conducting phylogeographic analyses (Fig. 5). Our simulation framework suggested that the outbreak in Bavaria, Germany was unlikely to be responsible for initiating the Italian outbreak (see fig. S7 and supplementary results for detailed phylogenetic scenarios). We simulated the origins of the Italian outbreak under the assumption that it was associated with viruses genetically related to the German virus BavPat1, again using realistic epidemiological parameters. Simulations with a median doubling time of 3.4 days (95% range: 2.9–4.4 days) resulted in a median epidemic size (including outbreaks that died out) of 725 infections (95% range: 140–2,847) after 36 days. In the observed phylogeny, the Italian outbreak is the sole descendant lineage from BavPat1. Within the Italian outbreak, there are zero viruses identical to BavPat1 and four of the 27 related viruses included in this analysis are separated from BavPat1 by a single mutation. In simulation, the distributions of identical and one-mutation divergent viruses are not significantly different from expectation ( $p = 0.156$  and  $p = 0.157$ , respectively). However, the lack of at least one descendant lineage that forms a polytomy with BavPat1 and the Italian outbreak significantly differs from expectation ( $p = 0.004$ ). Therefore, it is highly unlikely that BavPat1 or a virus identical to it initiated the Italian outbreak (fig. S7). As with the WA outbreak, these findings were robust to different infection rates and faster evolutionary rates (see Supplementary Text). Importantly, therefore, both a WA1-origin of the WA outbreak and a German origin of the Italian outbreak are rejected even by misspecified models of the epidemiological and evolutionary process.

An alternative scenario in which the outbreaks in both Germany and Italy were independently introduced from China is further supported by our phylogeographic inference (Fig. 5). The resulting reconstruction attributes higher

support to independent viral introductions from China into Germany and into Italy ( $pp = 0.84$ ), compared with a direct connection between Germany and Italy ( $pp = 0.16$ , Fig. 5). Similar support is obtained for this scenario by a time-inhomogeneous inference without unsampled taxa (fig. S8). These findings emphasize that epidemiological linkages inferred from genetically similar SARS-CoV-2 associated with outbreaks in different locations can be highly tenuous, given low levels of sampled viral genetic diversity and insufficient background data from key locations.

Our approach infers that the European B.1 clade (emanating from the green node labeled 0.86 in Fig. 5), the same one that dominates in New York City (14) and Arizona (23), had an origin in Italy, as might be expected from the epidemiological evidence. Both travel history and unsampled diversity contribute to this inference. While only two samples in our data set are from Italy, five additional genomes were obtained from people who arrived from Italy (Fig. 5). The unsampled taxa from Italy further contributed to a reconstruction with a higher support for Italy at the origin of the entire clade (Fig. 5 vs. fig. S8; also see fig. S9). The introduction from Hubei to Italy was dated to January 28th, 2020 (95% HPD: January 20th–February 6th). This Italian/European cluster, in turn, was the source of multiple introductions to New York City (NYC) (14). Using the same approach, we date the introduction leading to the largest NYC transmission cluster to February 12th, 2020 (95% HPD February 3th–February 22th). This is consistent with the finding that the earliest seropositive samples in NYC were from the week of February 17th through February 23rd (24).

Hence, even though a second introduction into Washington State (independent of WA1) implies a more recent date of origin of that transmission cluster than under the WA1-origin scenario (~February 1st versus January 15th, if it had originated with WA1), the WA outbreak clade still predates the earliest genomically-identified transmission clusters elsewhere in the US: the large one in NYC (4) plus two smaller, apparently self-limited clusters from California (in Solano County and Santa Clara County) that appear to have been introduced from China (5). Of these, the transmission cluster from Santa Clara County appears older, dating to before February 22nd, 2020 (95% HPD February 5th–February 29th) (Supplementary text).

### **Discussion**

Despite the early successes in containment, SARS-CoV-2 eventually took hold in both Europe and North America during the first two months of 2020: first in Italy around the end of January, then in Washington State around the beginning of February, and followed by New York City later that month. Our analyses therefore delineate when widespread community transmission was first established on both continents

(Fig. 6) and clarify the period before SARS-CoV-2 establishment when contact tracing and isolation might have been most effective.

Our findings highlight the potential value of establishing intensive, community-level respiratory virus surveillance architectures, such as the Seattle Flu Study, during a pre-pandemic period. The value of detecting cases early, before they have bloomed into an outbreak, cannot be overstated in a pandemic situation (25). Given that every delay in case detection reduces the feasibility of containment, it is also worth assessing the impact of lengthy delays in FDA approval of testing the Seattle Flu Study's stored samples for SARS-CoV-2.

By delaying COVID-19 outbreaks by even a few weeks in the US and Europe, the public health response to the WA1 case in Washington State, and a particularly impressive response in Germany to an early outbreak, bought crucial time for their own cities, as well as other countries and cities, to prepare for the virus when it finally did arrive. Surveillance efforts and genomic analyses subsequently helped close the gap between the onset of sustained community transmission and mitigation measures in Washington State, compared to other locales like New York City. However, our evidence suggests that the period between the founding of the outbreak and the initiation of mitigation measures in Washington State was not as long as supposed under the WA1-origin hypothesis and that the outbreak may therefore have been somewhat smaller than some estimates based on that hypothesis.

Because the SARS-CoV-2 evolutionary rate is slower than its transmission rate, many identical genomes are rapidly spreading. This genetic similarity places limitations on some inferences such as calculating the ratio of imported cases to local transmissions in a given area. Yet we have shown that, precisely because of this slow rate, when as little as one mutation separates viruses, this difference can provide enough information for hypothesis testing when appropriate methods are employed. Bearing this in mind will put us in a better position to understand SARS-CoV-2 in the coming years.

## REFERENCES AND NOTES

1. T. B. (@trvrb), "The team at the @seattleflustudy have sequenced the genome the #COVID19 community case reported yesterday from Snohomish County, WA, and have posted the sequence publicly to <http://gisaid.org>. There are some enormous implications here." Twitter (2020); <https://twitter.com/trvrb/status/1233970271318503426>.
2. T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M.-L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, The Seattle Flu Study Investigators, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautam, G. Melly, B. Hiatt, P. Dykerna, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, Cryptic transmission of SARS-CoV-2 in Washington state. *Science* 10.1126/science.abc0523 (2020). doi:[10.1126/science.abc0523](https://doi.org/10.1126/science.abc0523)
3. J. R. Fauver, M. E. Petrone, E. B. Hodcroft, K. Shiota, H. Y. Ehrlich, A. G. Watts, C. B. F. Vogels, A. F. Brito, T. Alpert, A. Muyombwe, J. Razeq, R. Downing, N. R. Cheemarla, A. L. Wyllie, C. C. Kalinich, I. M. Ott, J. Quick, N. J. Loman, K. M. Neugebauer, A. L. Greninger, K. R. Jerome, P. Roychoudhury, H. Xie, L. Shrestha, M.-L. Huang, V. E. Pitzer, A. Iwasaki, S. B. Omer, K. Khan, I. I. Bogoch, R. A. Martinello, E. F. Foxman, M. L. Landry, R. A. Neher, A. I. Ko, N. D. Grubaugh, Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* 181, 990–996.e5 (2020). doi:[10.1016/j.cell.2020.04.021](https://doi.org/10.1016/j.cell.2020.04.021) Medline
4. A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Albuquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, H. van Bakel, Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* 369, 297–301 (2020). doi:[10.1126/science.abc1917](https://doi.org/10.1126/science.abc1917) Medline
5. X. Deng, W. Gu, S. Federman, L. du Plessis, O. G. Pybus, N. R. Faria, C. Wang, G. Yu, B. Bushnell, C.-Y. Pan, H. Guevara, A. Sotomayor-Gonzalez, K. Zorn, A. Gopez, V. Servellita, E. Hsu, S. Miller, T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, H. Y. Chu, J. Shendure, K. R. Jerome, C. Anderson, K. Gangavarapu, M. Zeller, E. Spencer, K. G. Andersen, D. MacCannell, C. R. Paden, Y. Li, J. Zhang, S. Tong, G. Armstrong, S. Morrow, M. Willis, B. T. Matyas, S. Mase, O. Kasirye, M. Park, G. Masinde, C. Chan, A. T. Yu, S. J. Chai, E. Villarino, B. Bonin, D. A. Wadford, C. Y. Chiu, Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* 369, 582–587 (2020). doi:[10.1126/science.abb9263](https://doi.org/10.1126/science.abb9263) Medline
6. T. B. (@trvrb), "This separate introduction may have been to British Columbia or may have been elsewhere. Better resolving this introduction geographically would benefit from additional sequencing of samples collected closer in time to the introduction event. 14/18." Twitter (2020); <https://twitter.com/trvrb/status/1265063937663328256>.
7. M. L. Holshue, C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, A. Cohn, L. Fox, A. Patel, S. I. Gerber, L. Kim, S. Tong, X. Lu, S. Lindstrom, M. A. Pallansch, W. C. Weldon, H. M. Biggs, T. M. Uyeki, S. K. Pillai; Washington State 2019-nCoV Case Investigation Team, First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* 382, 929–936 (2020). doi:[10.1056/NEJMoa2001191](https://doi.org/10.1056/NEJMoa2001191) Medline
8. A. Harmon, "Inside the Race to Contain America's First Coronavirus Case." *The New York Times* (2020); [www.nytimes.com/2020/02/05/us/corona-virus-washington-state.html](http://www.nytimes.com/2020/02/05/us/corona-virus-washington-state.html).
9. D. A. Bolduc, "Webasto disputes link to Italy coronavirus outbreak." *Automotive News* (2020); [www.autonews.com/suppliers/webasto-disputes-link-italy-coronavirus-outbreak](http://www.autonews.com/suppliers/webasto-disputes-link-italy-coronavirus-outbreak).
10. M. M. Böhmer, U. Buchholz, V. M. Corman, M. Hoch, K. Katz, D. V. Marosevic, S. Böhm, T. Woudenberg, N. Ackermann, R. Konrad, U. Eberle, B. Treis, A. Dangel, K. Bengs, V. Fingerle, A. Berger, S. Hörmansdorfer, S. Ippisch, B. Wicklein, A. Grahl, K. Pörtnar, N. Müller, N. Zeitlmann, T. S. Boender, W. Cai, A. Reich, M. An der Heiden, U. Rexroth, O. Hamouda, J. Schneider, T. Veith, B. Mühlemann, R. Wölfel, M. Antwerpen, M. Walter, U. Protzer, B. Liebl, W. Haas, A. Sing, C. Drosten, A. Zapf, Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet Infect. Dis.* 20, 920–928 (2020). doi:[10.1016/S1473-3099\(20\)30314-5](https://doi.org/10.1016/S1473-3099(20)30314-5) Medline
11. C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, T. Zimmer, V. Thiel, C. Janke, W. Guggemos, M. Seilmaier, C. Drosten, P. Vollmar, K. Zwirglmaier, S. Zange, R. Wölfel, M. Hoelscher, Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *N. Engl. J. Med.* 382, 970–971 (2020). doi:[10.1056/NEJMc2001468](https://doi.org/10.1056/NEJMc2001468) Medline
12. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9241–9243 (2020). doi:[10.1073/pnas.2004999117](https://doi.org/10.1073/pnas.2004999117) Medline
13. A. Rambaut, E. C. Holmes, V. Hill, Á. O'Toole, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *bioRxiv* (2020); <https://doi.org/10.1101/2020.04.17.046086>.

14. M. T. Maurano, S. Ramaswami, G. Westby, P. Zappile, D. Dimartino, G. Shen, X. Feng, A. M. Ribeiro-dos-Santos, N. A. Vulpescu, M. Black, M. Hogan, C. Marier, P. Meyn, Y. Zhang, J. Cadley, R. Ordonez, R. Luther, E. Huang, E. Guzman, A. Serrano, B. Belovarac, T. Gindin, A. Lytle, J. Pinnell, T. Vougiouklakis, L. Boytard, J. Chen, L. H. Lin, A. Rapkiewicz, V. Raabe, M. I. Samanovic-Golden, G. Jour, I. Osman, M. Aguero-Rosenfeld, M. J. Mulligan, P. Cotzia, M. Snuderl, A. Heguy, Sequencing identifies multiple, early introductions of SARS-CoV2 to New York City Region. *medRxiv* (2020); doi:[10.1101/2020.04.15.20064931](https://doi.org/10.1101/2020.04.15.20064931)
15. E. Minskaia, T. Hertzig, A. E. Gorbalenya, V. Campanacci, C. Cambillau, B. Canard, J. Ziebuhr, Discovery of an RNA virus 3'-5' exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5108–5113 (2006). doi:[10.1073/pnas.0508200103](https://doi.org/10.1073/pnas.0508200103) Medline
16. Materials and methods are available as supplementary materials at the *Science* website.
17. P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, M. A. Suchard, Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLOS Pathog.* **10**, e1003932 (2014). doi:[10.1371/journal.ppat.1003932](https://doi.org/10.1371/journal.ppat.1003932) Medline
18. V. N. Minin, M. A. Suchard, Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. B* **363**, 3985–3995 (2008). doi:[10.1098/rstb.2008.0176](https://doi.org/10.1098/rstb.2008.0176) Medline
19. H. Y. Chu, J. A. Englund, L. M. Starita, M. Famulare, E. Brandstetter, D. A. Nickerson, M. J. Rieder, A. Adler, K. Lacombe, A. E. Kim, C. Graham, J. Logue, C. R. Wolf, J. Heimonen, D. J. McCulloch, P. D. Han, T. R. Sibley, J. Lee, M. Ilcisin, K. Fay, R. Burstein, B. Martin, C. M. Lockwood, M. Thompson, B. Lutz, M. Jackson, J. P. Hughes, M. Boeckh, J. Shendure, T. Bedford; Seattle Flu Study Investigators, Early Detection of Covid-19 through a Citywide Pandemic Surveillance Platform. *N. Engl. J. Med.* **383**, 185–187 (2020). doi:[10.1056/NEJMcp2008646](https://doi.org/10.1056/NEJMcp2008646) Medline
20. The White House, "Proclamation on Suspension of Entry as Immigrants and Nonimmigrants of Persons who Pose a Risk of Transmitting 2019 Novel Coronavirus"; [www.whitehouse.gov/presidential-actions/proclamation-suspension-entry-immigrants-nonimmigrants-persons-pose-risk-transmitting-2019-novel-coronavirus/](http://www.whitehouse.gov/presidential-actions/proclamation-suspension-entry-immigrants-nonimmigrants-persons-pose-risk-transmitting-2019-novel-coronavirus/).
21. S. Eder, H. Fountain, M. H. Keller, M. Xiao, A. Stevenson, "430,000 People Have Traveled From China to U.S. Since Coronavirus Surfaced." *The New York Times* (2020); [www.nytimes.com/2020/04/04/us/coronavirus-china-travel-restrictions.html](http://www.nytimes.com/2020/04/04/us/coronavirus-china-travel-restrictions.html).
22. "Preliminary analysis of SARS-CoV-2 importation and establishment of UK transmission lineages." *Virological* (2020); <https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507>.
23. J. T. Ladner, B. B. Larsen, J. R. Bowers, C. M. Hepp, E. Bolyen, M. Folkerts, K. Sheridan, A. Pfeiffer, H. Yaglom, D. Lemmer, J. W. Sahl, E. A. Kaelin, R. Maqsood, N. A. Bokulich, G. Quirk, T. D. Watt, K. Komatsu, V. Waddell, E. S. Lim, J. G. Caporaso, D. M. Engelthaler, M. Worobey, P. Keim, Defining the Pandemic at the State Level: Sequence-Based Epidemiology of the SARS-CoV-2 virus by the Arizona COVID-19 Genomics Union (ACGU). *medRxiv* (2020); doi:[10.1101/2020.05.08.20095935](https://doi.org/10.1101/2020.05.08.20095935).
24. D. Stadlbauer, J. Tan, K. Jiang, M. Hernandez, S. Fabre, F. Amanat, C. Teo, G. Asthagiri Arunkumar, M. McMahon, J. Jhang, M. Nowak, V. Simon, E. Sordillo, H. van Bakel, F. Krammer, Seroconversion of a city: Longitudinal monitoring of SARS-CoV-2 seroprevalence in New York City. *medRxiv* (2020); doi:[10.1101/2020.06.28.20142190](https://doi.org/10.1101/2020.06.28.20142190).
25. M. Worobey, Epidemiology: Molecular mapping of Zika spread. *Nature* **546**, 355–357 (2017). doi:[10.1038/nature22495](https://doi.org/10.1038/nature22495) Medline
26. Zenodo, <http://doi.org/10.5281/zenodo.3979896>.
27. Z. Dezső, A.-L. Barabási, Halting viruses in scale-free networks. *Phys. Rev. E* **65**, 055103 (2002). doi:[10.1103/PhysRevE.65.055103](https://doi.org/10.1103/PhysRevE.65.055103) Medline
28. J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, W. J. Edmunds, Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Med.* **5**, e74 (2008). doi:[10.1371/journal.pmed.0050074](https://doi.org/10.1371/journal.pmed.0050074) Medline
29. 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](https://doi.org/10.1038/nature04153) Medline
30. S. J. Spielman, C. O. Wilke, Pvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLOS ONE* **10**, e0139047 (2015). doi:[10.1371/journal.pone.0139047](https://doi.org/10.1371/journal.pone.0139047) Medline
31. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020). doi:[10.1093/molbev/msaa015](https://doi.org/10.1093/molbev/msaa015) Medline
32. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016). doi:[10.1093/molbev/msw046](https://doi.org/10.1093/molbev/msw046) Medline
33. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). doi:[10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010) Medline
34. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016). doi:[10.1093/ve/vew007](https://doi.org/10.1093/ve/vew007) Medline
35. N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, S. Mirarab, FAVITES: Simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics* **35**, 1852–1861 (2019). doi:[10.1093/bioinformatics/bty921](https://doi.org/10.1093/bioinformatics/bty921) Medline
36. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018). doi:[10.1093/ve/vey016](https://doi.org/10.1093/ve/vey016) Medline
37. D. L. Ayres, M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, M. A. Suchard, BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019). doi:[10.1093/sysbio/syz020](https://doi.org/10.1093/sysbio/syz020) Medline
38. P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao, Z.-L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020). doi:[10.1038/s41586-020-2012-7](https://doi.org/10.1038/s41586-020-2012-7) Medline
39. F. Bielejec, P. Lemey, G. Baele, A. Rambaut, M. A. Suchard, Inferring heterogeneous evolutionary processes through time: From sequence substitution to phylogeography. *Syst. Biol.* **63**, 493–504 (2014). doi:[10.1093/sysbio/syu015](https://doi.org/10.1093/sysbio/syu015) Medline
40. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018). doi:[10.1093/sysbio/syy032](https://doi.org/10.1093/sysbio/syy032) Medline

## ACKNOWLEDGMENTS

We thank the patients and healthcare workers who made the collection of this global viral data set possible and all those who made viral genomic data available for analysis. We thank Niema Moshiri for his guidance on FAVITES, Trevor Bedford for insights into how viral genomic inferences influenced public health responses in Washington State, and Louis du Plessis for insights into the timing of the origin of the WA outbreak clade based on Grand Princess voyage dates. **Funding:** MW was supported by the David and Lucile Packard Foundation as well as the University of Arizona College of Science. This work was supported by the Multinational Influenza Seasonal Mortality Study (MISMS), an on-going international collaborative effort to understand influenza epidemiology and evolution, led by the Fogarty International Center, NIH. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 725422-ReservoirDOCS) and from the European Union's Horizon 2020 project MOOD (grant agreement no. 874850). The Arctic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z. JOW acknowledges funding from the National Institutes of Health (K01AI110181, AI135992, and AI136056). PL acknowledges support by the Research Foundation–Flanders ("Fonds voor Wetenschappelijk Onderzoek–Vlaanderen," G066215N, G0D5117N and G0B9317N). MAS acknowledges support from

National Institutes of Health U19 AI135995. JBJ is thankful for support from the Canadian Institutes of Health Research Coronavirus Rapid Response Programme 440371 and Genome Canada Bioinformatics and Computational Biology Programme grant 287PHY. JP acknowledges funding from the National Institutes of Health (T15LM011271). VH acknowledges funding from the Biotechnology and Biological Sciences Research Council (BBSRC) [grant number BB/M010996/1]. The content is solely the responsibility of the authors and does not necessarily represent official views of the National Institutes of Health. We gratefully acknowledge support from NVIDIA Corporation with the donation of parallel computing resources used for this research. **Author contributions:** Conceptualization: MW. Methodology: MW, JP, MAS, PL, JOW. Software: JP, MAS, PL, JOW. Validation: JP, MAS, PL. Formal analysis: MW, JP, PL, MAS. Investigation: MW, JP, BBL, JBJ, AR, MIN, VH. Resources: MW, PL, MAS. Data Curation: BBL, JBJ, VH. Writing - original draft preparation: MW, MIN. Writing - review and editing: MW, BBL, MAS, JOW, JBJ, AR. Visualization: BBL, JOW, AR. Supervision: MW, JOW. Project administration: MW. Funding acquisition: MW, MAS, JOW. **Competing interests:** JOW has received funding from Gilead Sciences, LLC (completed) and the CDC (ongoing) via grants and contracts to his institution unrelated to this research. MAS receives funding from Janssen Research & Development, IQVIA and Private Health Management via contracts unrelated to this research. **Data and materials availability:** All data used in this analysis are free to access: BEAST .xml file example, FAVITES simulated phylogenies, the GISAID accession numbers for all sequences used in the analysis, and alignments are hosted at GitHub ([https://github.com/Worobeylab/SC2\\_outbreak](https://github.com/Worobeylab/SC2_outbreak)) and Zenodo (26). 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.

## SUPPLEMENTARY MATERIALS

[science.scienmag.org/cgi/content/full/science.abc8169/DC1](http://science.scienmag.org/cgi/content/full/science.abc8169/DC1)

Materials and Methods

Supplementary Text

Figs. S1 to S9

Tables S1 to S3

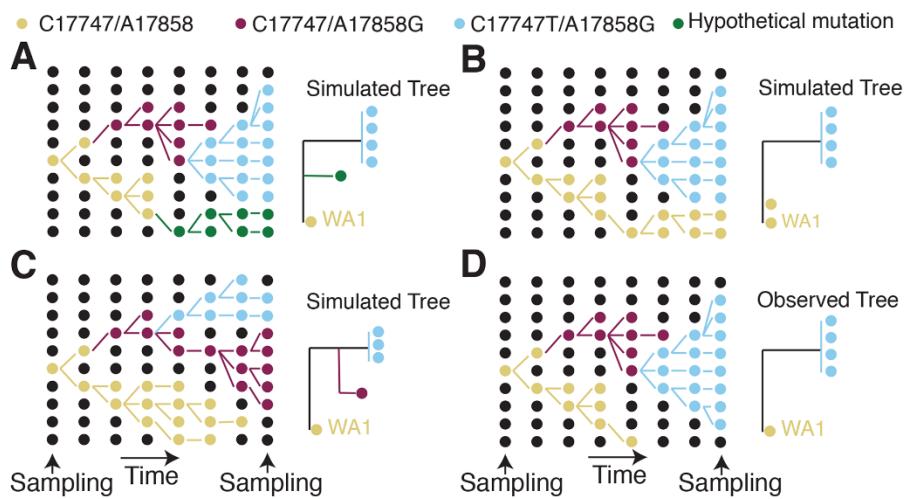
References (27–40)

MDAR Reproducibility Checklist

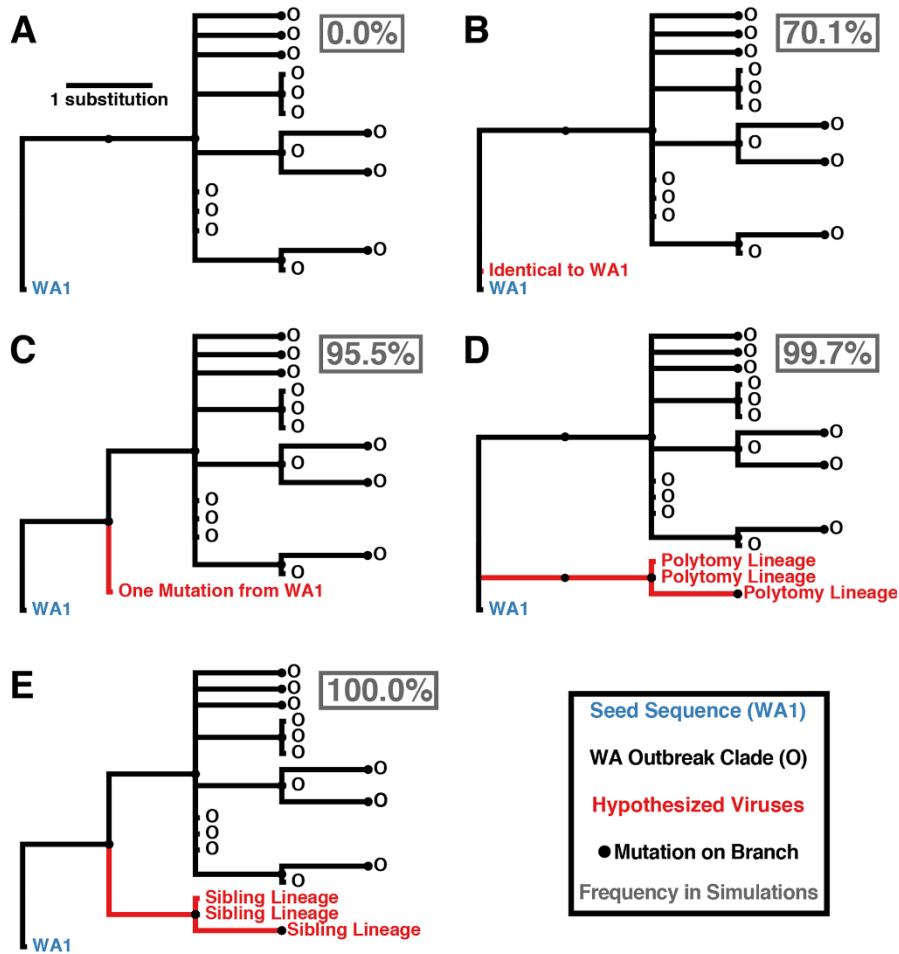
18 May 2020; accepted 3 September 2020

Published online 10 September 2020

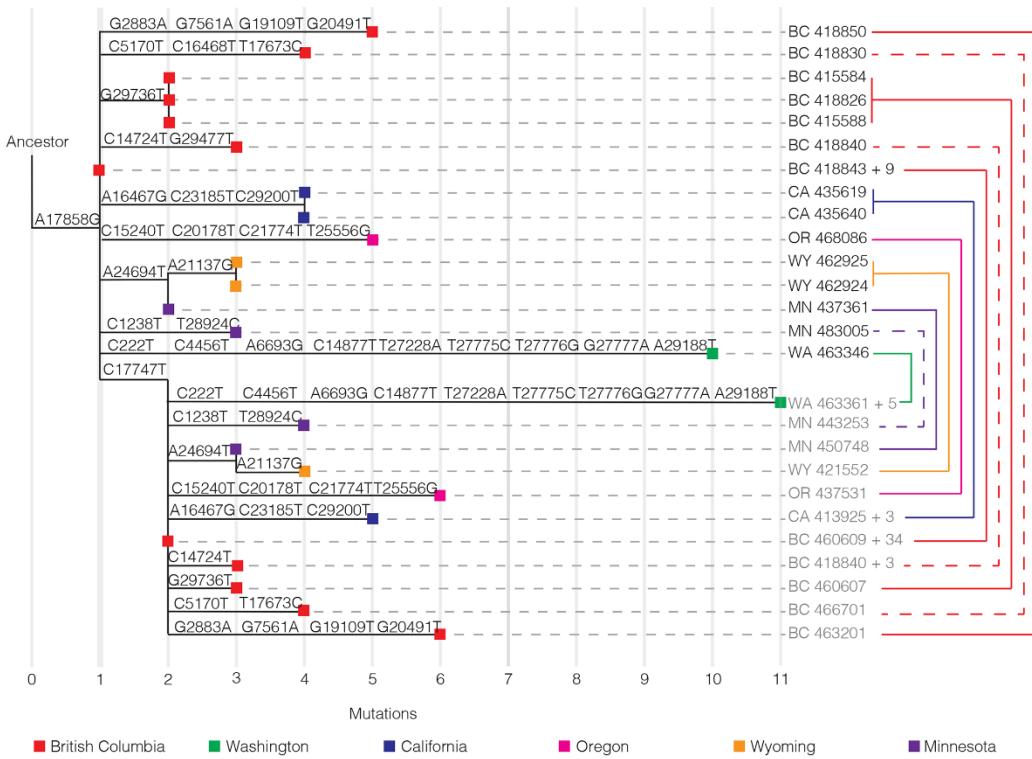
10.1126/science.abc8169



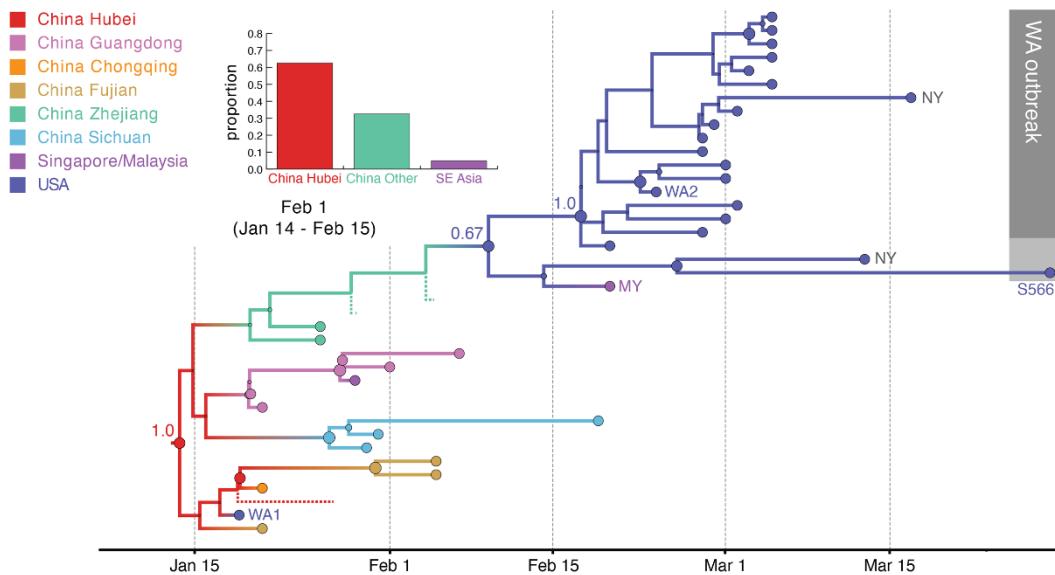
**Fig. 1. Schematic showing a hypothetical path along which the key mutations in the WA outbreak could have taken in a susceptible population, alongside the inferred phylogeny.** (A) Scenario where a hypothetical mutation occurs from WA1-like genomes. (B) A hypothetical phylogeny where A17747 and C17858 from the original WA1 virus are maintained in the population and sampled at the end. (C) Hypothetical scenario where a virus one mutation (A17858G) different from WA1 is maintained in the population. (D) The observed tree from the WA outbreak.



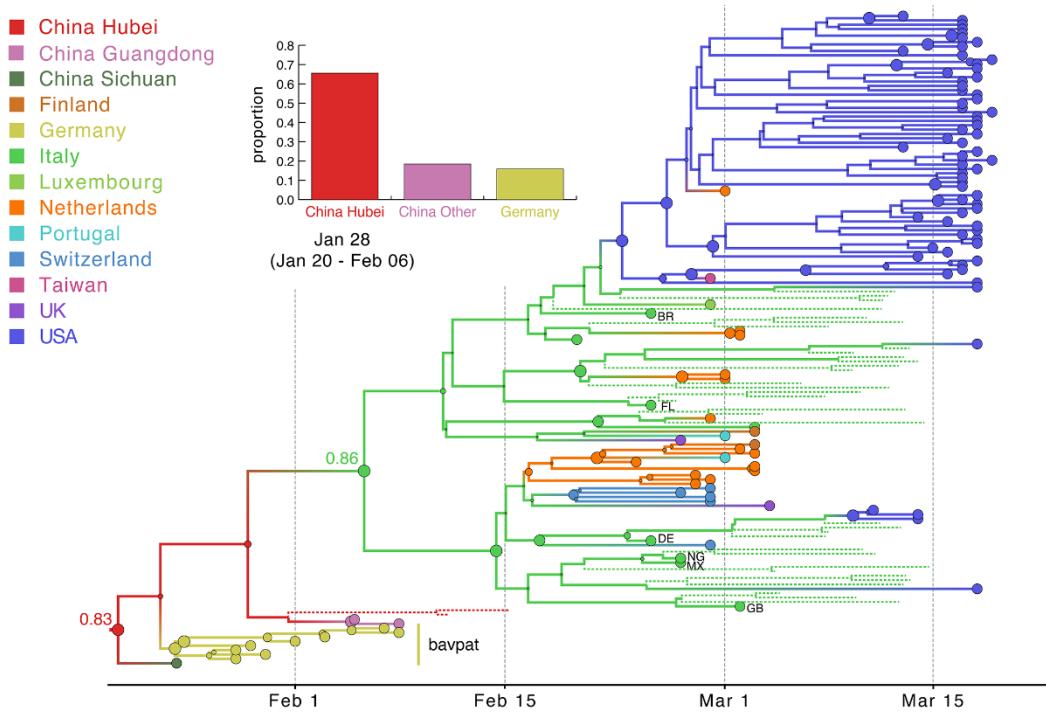
**Fig. 2. Potential phylogenetic relationships between WA1 and the WA outbreak clade and their occurrence probabilities.** (A) Observed pattern where the WA1 genome is the direct ancestor of the outbreak clade, separated by at least two mutations. (B) Identical sequence to WA1. (C) Sequence that is one mutation divergent from WA1. (D) Lineage forming a basal polytomy with WA1 and the outbreak clade. (E) Sibling lineage to the outbreak clade experiencing fewer than two mutations from WA1 before divergence. The frequency of each relationship across 1000 simulations is reported in the gray box.



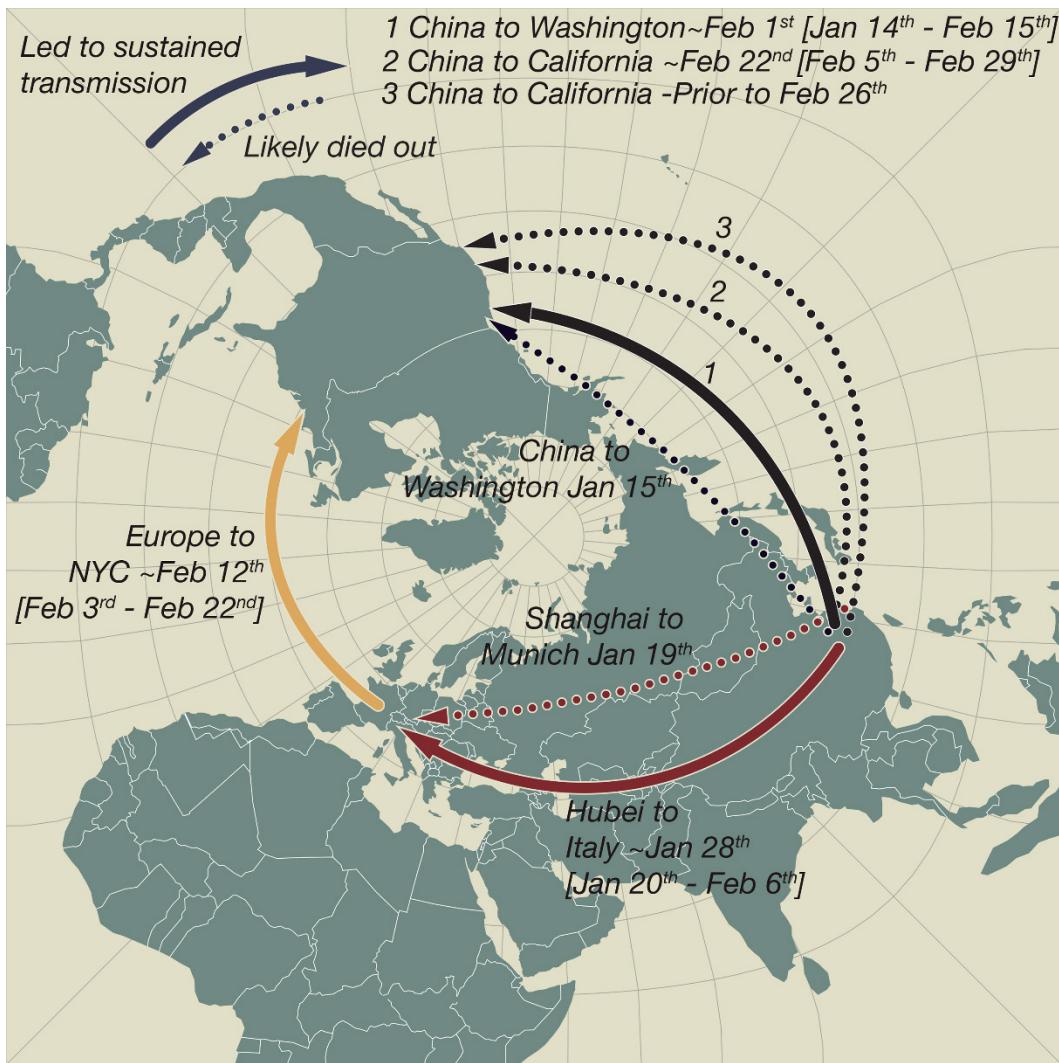
**Fig. 3. Phylogeny of representative sequences showing connections between sequences that share derived mutations despite differences at the key site 17747.** Derived mutations away from ancestral states are shown above each branch with position number (relative to the reference sequence hCoV-19/Wuhan/Hu-1/2019|EPI\_ISL\_402125). Branches are connected to taxon names with a horizontal dotted line. The taxon names include a two-letter state or province code, as well as the GISAID accession number. In cases where more than one sequence is represented, the total number of additional, identical sequences is shown following “+.” Sequences that share derived mutations are connected with colored lines on the right, with the colors of the line indicating the location the connected sequences were sampled. Some lines on the right are dashed for clarity. Sequences that contain the derived nucleotide at 17747 have names shaded in gray.



**Fig. 4. Hypothesis of SARS-CoV-2 entry into Washington State.** A subtree of the maximum clade credibility (MCC) tree depicting the evolutionary relationships inferred between (i) the first identified SARS-CoV-2 case in the US (WA1); (ii) the clade associated with the Washington State outbreak (including WA2) and related viruses (WA-S566, and a virus from NY); and (iii) closely related viruses that were identified in multiple locations in Asia. Genome sequences sampled at the tips of the phylogeny are represented by circles shaded according to location of sampling. Internal node circles, representing posterior clade support values, and branches are shaded similarly by location. Dotted lines represent branches associated with unsampled taxa assigned to Hubei and Zhejiang, China. Posterior location state probabilities are shown for three well-supported key nodes (with the color of the circle indicating inferred location state). The bar chart summarizes the probability by location for a second introduction giving rise to the WA outbreak clade. The mean date and 95% HPD intervals represent the estimate for the time of the introduction from Hubei.



**Fig. 5. MCC tree of SARS-CoV-2 entry into Europe.** A subtree inferred for viruses from (i) the first outbreak in Europe (Germany, BatPat) and identical viruses from China, (ii) outbreaks in Italy and New York, and (iii) other locations in Europe. Dotted lines represent branches associated with unsampled taxa assigned to Italy and Hubei, China (CN). Country codes are shown at tips for genomes sampled from travelers returning from Italy. The bar chart summarizes the probability distribution for the location state ancestral to the Italian clade. Other features as described in Fig. 4.



**Fig. 6. SARS-CoV-2 introductions to Europe and the US.** Pierce projection mapping early and apparently “dead-end” introductions of SARS-CoV-2 to Europe and the US (dashed arrows). Successful dispersals between late January and mid-February are shown with solid arrows: from Hubei Province, China to Northern Italy, from China to Washington State, and later from Europe (as the Italian outbreak spread more widely) to New York City and from China to California.

## The emergence of SARS-CoV-2 in Europe and North America

Michael Worobey, Jonathan Pekar, Brendan B. Larsen, Martha I. Nelson, Verity Hill, Jeffrey B. Joy, Andrew Rambaut, Marc A. Suchard, Joel O. Wertheim and Philippe Lemey

published online September 10, 2020 originally published online September 10, 2020

ARTICLE TOOLS

<http://science.scienmag.org/content/early/2020/09/11/science.abc8169>

SUPPLEMENTARY MATERIALS

<http://science.scienmag.org/content/suppl/2020/09/10/science.abc8169.DC1>

REFERENCES

This article cites 34 articles, 9 of which you can access for free

<http://science.scienmag.org/content/early/2020/09/11/science.abc8169#BIBL>

PERMISSIONS

<http://www.scienmag.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of Service](#)

---

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2020 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution License 4.0 (CC BY).



[science.scienmag.org/cgi/content/full/science.abc8169/DC1](https://science.scienmag.org/cgi/content/full/science.abc8169/DC1)

## Supplementary Materials for

### The emergence of SARS-CoV-2 in Europe and North America

Michael Worobey\*, Jonathan Pekar, Brendan B. Larsen, Martha I. Nelson, Verity Hill,  
Jeffrey B. Joy, Andrew Rambaut, Marc A. Suchard\*, Joel O. Wertheim\*,  
Philippe Lemey\*

\*Corresponding author. Email: [worobey@arizona.edu](mailto:worobey@arizona.edu) (M.W.); [msuchard@ucla.edu](mailto:msuchard@ucla.edu) (M.A.S.);  
[jwertheim@health.ucsd.edu](mailto:jwertheim@health.ucsd.edu) (J.O.W.); [philippe.lemey@kuleuven.be](mailto:philippe.lemey@kuleuven.be) (P.L.)

Published 10 September 2020 on *Science* First Release  
DOI: [10.1126/science.abc8169](https://doi.org/10.1126/science.abc8169)

#### This PDF file includes:

Materials and Methods  
Supplementary Text  
Figs. S1 to S9  
Tables S1 to S3  
References

**Other Supplementary Materials for this manuscript includes the following:**  
(available at [science.scienmag.org/cgi/content/full/science.abc8169/DC1](https://science.scienmag.org/cgi/content/full/science.abc8169/DC1))

MDAR Reproducibility Checklist (.pdf)

## Materials and Methods

### Methods

*Epidemic Simulation using FAVITES.* To compare observed evolutionary patterns and those expected if the WA outbreak clade had evolved from WA1, we downloaded from the Global Initiative on Sharing All Influenza Data (GISaid) the 294 SARS-CoV-2 genomes sampled in Washington state up until March 15<sup>th</sup>, a period prior to rolling out of statewide mitigation efforts that likely dramatically altered the dynamics of the SARS-CoV-2 outbreak there. We then performed a series of epidemic simulations using FAVITES v1.1.35 (Fig. S1). First, we generated static contact networks in FAVITES under a preferential-attachment model using the Barabási-Albert algorithm (Fig. S1). We used this network because its scale-free properties are well-suited to infectious diseases (27), and since we are primarily focused on phylogenetic tree structure, a dynamic contact network is unnecessary. The reported number of reported contacts per day varies, with Mossong et al. reporting a mean of 13.4 contacts per day across 8 European countries, but Italy having 19.8 contacts per day (28). Therefore, we selected an intermediate value of 16 contacts per day (i.e., mean degree) within a 20,000 node contact network.

Concurrently, we performed a forward simulation SIR (Susceptible-Infected-Recovered) model to generate a transmission network over this contact network (Fig. S1). Starting with a single seed among our 20,000 susceptible individuals, we applied an infectiousness parameter of 2.7 inverse person-years, or the probability that an infected neighbor would transmit the virus, and a recovery time of 20 days. The SIR model and contact network are linked to document the propagation of infection through the network, creating the transmission network (Fig. S1). A single viral lineage from each infected individual was sampled at a random point during their infection time to represent viral genotype sampling. The viral phylogeny in units of time (years) was sampled under a coalescent model using the Virus Tree Simulator package embedded in FAVITES. The final output utilized from FAVITES is the viral time-based phylogeny (Fig. S1). All inputs for FAVITES can be found in Table S1 and the JSON input files are available at [https://github.com/Worobeylab/SC2\\_outbreak](https://github.com/Worobeylab/SC2_outbreak).

We constrained our simulated epidemics to have a median doubling time of 4.7 days (95% range: 4.2–5.1), corresponding to prior phylodynamic analysis (2). All resulting time trees were subsampled to match the dates corresponding to the 294 sampling dates in the WA outbreak (Fig. S1). If 294 precisely matched dates were not found in the phylogeny, we subsampled tips that were closest to the desired dates by expanding the range to a 3-day window. If there were not enough viruses to properly capture the sampling dates for the WA outbreak clade using this method, the simulation was excluded. We produced 1000 simulations for each set of data.

Across 1000 simulations, our epidemics encompassed high-levels of variance in transmission, consistent with the presence of so-called “superspreading events” (29) (Fig. S2). Because the contact network was generated under a preferential attachment model, the most connected individuals in the contact networks had a median number of 507 contacts (95% range across simulations: 403–812) (Fig. S2A). This network feature was retained in simulated transmission networks, where the maximum number of transmission events from a single individual ranged between 38 and 122, with a median of 66 (Fig. S2B). Further, we observed multiple super-spreading events in each simulation (Fig. S2C). The right-most distribution of these events exceeded those expected under a negative-binomial distribution (Fig. S2D).

*Sequence Simulation and Phylogenetic Inference.* We simulated genetic sequences over the sub-sampled epidemic trees using Pyvolve v1.0.1 (30). There were 262 variant sites found in an alignment of length 29,850 with 295 taxa comprising WA1 and the Washington outbreak clade. The simulations were restricted to two times the amount of variant sites ( $2 \times 262 = 524$ ) to avoid an upper limit and better capture the amount of variants in the original data. We used the 262 nucleotides from WA1 that correspond to the variant sites, with each nucleotide occurring twice, as the root of the sub-sampled tree. We applied a

genome-wide evolutionary rate of  $0.8 \times 10^{-3}$  substitutions/site/year to the variant sites, which translates to 0.0456 substitutions/site/year across the 542 sites:  $(0.8 \times 10^{-3} \times 29,850) / 524$  (Fig. S1). To parameterize the sequence evolution, we first inferred a maximum likelihood tree for the 295 taxa WA outbreak clade alignment using a GTR + Inv model in IQ-TREE2 v2.0.4 (31). We then applied the GTR substitution model to evolve the 524 variant sites to generate simulated alignments. Each simulation used two repetitions of the WA1 seed sequence variant sites as the starting point of the evolution (Fig. S1). We then added the WA1 seed sequence along with the invariant genomic positions. A maximum likelihood phylogeny was then inferred in IQ-TREE2 using a GTR + Inv substitution model (Fig. S1).

We determined the number of identical and 1-mutation divergent strains by examining the simulated sequence alignment. We used ETE3 v3.1.1 (32) to determine how the number of lineages that diverged basally or before experiencing two mutations from WA1.

*WA Sensitivity Analysis—Slower rate of infection.* The same methods for epidemic and sequence simulation and tree inference were done on the WA data to test a more slowly spreading virus. We used the aforementioned parameters, except the contact network was 5,000 nodes and the infectiousness rate was 2.0-2.2 (Table S1). We constrained our output to be between 500 and 2,000 infected people and produced 1,000 replicates. The same methods as earlier were used for the subsequent tree pruning and scaling, sequence evolution, and tree inference.

*WA Sensitivity Analysis—Faster rate of evolution.* We took the subsampled mutation-scaled trees from the original WA epidemic and sequence simulation and multiplied each branch length by 2 to simulate an accelerated evolutionary rate of  $1.6 \times 10^{-3}$  substitutions/site/year. The same methods as earlier were used for the subsequent sequence evolution and tree inference.

*WA Sensitivity Analysis—Sampling 1000 genomes.* Starting with the time trees from the original WA epidemic and sequence simulation, we subsampled tips that were closest to the desired dates by expanding the range to a 3-day window, and then randomly subsampled an additional 706 tips at dates postdating than February 20th, bringing the total to 1000 leaves on the time tree. The same methods as earlier were used for the subsequent sequence evolution and tree inference. The null hypothesis for this analysis was having 2 divergent lineages prior to the formation of the WA outbreak clade.

*Italy Outbreak Epidemic Simulation.* The same methods for epidemic and sequence simulation and tree inference were used to model the Italy outbreak. We used a contact network of 10,000 nodes and the infectiousness rate was 2.6-2.8 for the epidemic simulation within FAVITES: 500 applied an infectiousness rate of 2.6, and the other 500 applied an infectiousness rate of 2.8. Prior epidemiological analysis of the initial German outbreak identified 16 cases of COVID-19 16 days after an introduction, leading to a doubling time of 3.2 days (10). We constrained our simulated epidemics to have a similar median doubling time of 3.4 days (95% range: 2.9-4.4 days). The subsampling method remained the same. There were 45 variant sites found in an alignment of length 29,775 with 28 taxa comprising Italian outbreak. The simulations were restricted to 90 nucleotides (two times the amount of variant sites), using 45 nucleotides from BavPat1 that correspond to the variant sites, with each nucleotide occurring twice, as the root of the sub-sampled tree. The evolutionary rate of  $0.8 \times 10^{-3}$  substitutions/site/year was translated to 0.265 substitutions/site/year across the 90 sites:  $(0.8 \times 10^{-3} \times 29,775) / 90$ . The remainder of the pipeline was the same, except using the Germany/Italy data and seeding the evolution with the BavPat1 sequence. The files used for FAVITES, the viral time trees, and the subsampled mutation trees can be found in [https://github.com/Worobeylab/SC2\\_outbreak](https://github.com/Worobeylab/SC2_outbreak).

*Italy Outbreak Sensitivity Analysis—Slower rate of infection.* The same methods for epidemic and sequence simulation and tree inference were done on the Italy outbreak data to test a more slowly spreading virus. We used the aforementioned parameters, except the contact network was 2,500 nodes

and the infectiousness rate was 2.0-2.2 (Table S1). We constrained our output to be less than 1,600 infected people and produced 1,000 replicates. The same methods as earlier were used for the subsequent tree pruning and scaling, sequence evolution, and tree inference.

*Italy Outbreak Sensitivity Analysis—Faster rate of evolution.* We took the subsampled mutation-scaled trees from the original Italy outbreak and sequence simulation and multiplied each branch length by 2 to simulate an accelerated evolutionary rate of  $1.6 \times 10^{-3}$  substitutions/site/year. The same methods as earlier were used for the subsequent sequence evolution and tree inference.

*SARS-CoV-2 genome data set for phylogeographic analyses.* To investigate the evolutionary origins of early outbreaks of the SARS-CoV-2 virus in the Europe and US, we compiled a curated SARS-CoV-2 genome sequence data based on available genomes from the GISAID on March 10<sup>th</sup>, 2020. In addition, we added genomes that have since become available from Wuhan (Hubei, China), genomes from New York City and genomes identical or closely related to Bavpat1 (EPI\_ISL\_406862) and WA1 (EPI\_ISL\_404895). This extends the sampling time range to March 29<sup>th</sup>. Acknowledgements of all laboratories that contributed SARS-CoV-2 sequence data used in our study are found in [https://github.com/Worobeylab/SC2\\_outbreak](https://github.com/Worobeylab/SC2_outbreak). The genomes were aligned using MAFFT v.7 and partially trimmed at the 5' and 3' ends (from position 1 to 39 and from position 29830 to 29903 according to the NC\_045512.2 reference genome) (33). Viruses with evidence of sequencing error, based on having stretches with many apparent single nucleotide polymorphisms, were identified based on visual inspection and removed. Only single genomes were retained from patients for which multiple genomes were available. For additional quality control, root-to-tip divergence was visualized as a function of sampling time using TempEst (34), based on a maximum likelihood (ML) tree inferred with IQTree (31). One potential outlier was identified and removed (hCoV-19/South Korea/SNU01/2020, EPI\_ISL\_411929). Genomes sampled from cruise ship travelers were not considered. The final data set consisted of 555 genomes that were sampled from 28 countries. Chinese samples were available from 13 provinces, one municipality (Beijing), and one special administrative area (Hong Kong), which were considered as separate locations in our phylogeographic reconstructions. The exact date of virus collection was available for all sequences except for one genome from Anhui, China, for which only the month of viral collection was available. In this case, the lack of tip date precision was accommodated by sampling uniformly across a 30-day window (35). The different genomes from the Bavpat cluster that resulted from a single introduction were constrained to monophyletic in order to avoid phylogenetic intermingling with identical genomes from China.

*Bayesian phylogeographic inference.* The evolutionary history of SARS-CoV-2 was inferred using a Bayesian approach, implemented through the Markov chain Monte Carlo (MCMC) framework available in BEAST 1.10.4 (36), using the BEAGLE library v3 (37) to increase computational performance. We specify an HKY85 nucleotide substitution model using a continuous-time Markov chain (CTMC) process, with a gamma-distributed rate variation among sites, a strict molecular clock model and an exponential growth coalescent tree prior. The relatively low sequence variability over the short time scale of viral dissemination introduces considerable uncertainty into phylogenetic reconstructions of SARS-CoV-2. Uncertainty in inferences of viral spatial movements is further compounded by the low availability of genetic sequence data from a number of critical countries and regions. To address such limitations, Bayesian reconstructions of the evolutionary history and spatial origins of SARS-CoV-2 were augmented by incorporating various sources of information. For genomes from patients with a recent travel history, we used the travel locations in our ancestral location reconstructions, except for WA1 because we did not want to bias our analysis against a direct connection between WA1 and the WA outbreak in our joint Bayesian inference of sequences and locations. In this case, we specified an ambiguous location state that includes both USA and Hubei as possible location states. For locations that are poorly represented (e.g. Italy), genomes sampled from travelers upon returning from these locations assist in uncovering the diversity in these locations.

*Incorporating unsampled viruses into Bayesian phylogeographic inference.* For poorly sampled locations, we included in the Bayesian inference unsampled genomes that were associated with an epidemiologically-informed sampling time and location but not observed sequence data. Phylogeographic reconstructions are highly sensitive to sample bias, and 12 locations were identified as critically undersampled by contrasting the number of available genomes against the cumulative number of recorded COVID-19 cases in each location on March 10<sup>th</sup>, 2020 (available from Our World in Data, <https://ourworldindata.org/coronavirus-source-data>). A standard approach to reduce geographical bias would be to down-sample data sets by removing genomes from more densely sampled locations. However, sequence data are sparse during the pre-pandemic stages of SARS-CoV-2 emergence and global spread, particularly from locations that experienced early outbreaks. In order to retain all available sequences, we opted instead to add unsampled taxa from underrepresented locations until achieving an arbitrary minimal ratio of 0.005 of taxa (sampled and unsampled) to cumulative number of cases in each location. To meet this threshold, 289 unsampled taxa were included from 12 locations that experienced early SARS-CoV-2 outbreaks but were poorly represented in the data available in GISAID. The majority of unsampled taxa were assigned to Hubei ( $n = 138$ ), followed by Italy ( $n = 47$ ), Iran ( $n = 40$ ), and South Korea ( $n = 30$ ).

Epidemiological information also was utilized to specify realistic sampling times for the 289 unsampled taxa. For each of the 12 locations associated with unsampled taxa, the time of sampling was described by a probability distribution, the shape of which was based on prevalent infections over time, estimated using the methods and data sources recently described by Fauver et al. (3). Probabilistic prior distributions were normally distributed (and appropriately truncated) in east Asian locations that effectively controlled early SARS-CoV-2 outbreaks through intensive social distancing and contact tracing (Table S2). Probabilistic prior distributions were exponentially distributed in most European countries where SARS-CoV-2 cases were still increasing exponentially in early March (Table S3). Inaccuracies in COVID-19 case counts arising from underreporting and low testing are likely to affect the absolute number of COVID-19 cases in a given location, but less likely to impact the general shape and timing of epidemic curves, as used here. The probabilistic distributions for the sampling times of unsampled taxa were integrated over all possible times using MCMC in the full Bayesian analysis. To aid mixing in the presence of unsampled taxa, we used an informative prior on the growth rate of the exponential growth coalescent prior based on the analysis without unsampled taxa. To illustrate the behavior of unsampled taxa, we perform an analysis that includes only a single unsampled taxon from four different locations and summarize their plausible phylogeographic clustering (Supplementary text, Fig. S9).

*Bayesian phylogeographic inference with covariates.* The final discrete spatial diffusion model included 44 locations, including 12 locations associated with unsampled taxa. To avoid an exceedingly high number of location transition parameters in a high dimensional CTMC, and to further inform the phylogenetic position of unsampled taxa, we utilized a generalized linear model (GLM) formulation of the discrete trait CTMC that parametrizes transition rates as a function of potential covariates (17). We considered three covariates related to human mobility that potentially predict long-distance routes of early SARS-CoV-2 spread: (i) air travel [as previously used in (17)], (ii) geographic distance (within continents), and (iii) an estimable asymmetry coefficient for Hubei, since in the early stage SARS-CoV-2 spread is dominated by an asymmetric flow out of Hubei, where the pandemic originated (38). The effect size of each covariate was estimated, along with an inclusion probability. In our phylogeographic inference, we also included a Markov jump counting procedure in order to estimate the times of specific transitions between locations (17, 18).

In order to evaluate the sensitivity of our phylogeographic estimates to the assumption of time-homogeneous covariate contributions in our GLM diffusion model, we construct a time-inhomogeneous diffusion model by borrowing concepts of epoch modelling from Bielejec et al. (2014) (39). We specify

an epoch model with two-week time intervals over the unknown phylogeny allowing independent effect sizes for the two significant covariates in each of these time intervals. We specify a first order random walk prior over the log effect sizes with a mean of 0 and an estimable precision. Due to the computational burden associated with this model, we restrict this analysis to the data set without unsampled diversity.

The posterior distribution of the full probabilistic models were estimated using MCMC sampling, running chains sufficiently long (e.g. > 100M iterations for analyses including unsampled taxa) to ensure adequate effective sample sizes for continuous parameters in combined runs (ESS > 100), as diagnosed using Tracer (40). Posterior tree distributions were summarized using maximum clade credibility (MCC) trees that were visualized in FigTree v1.4.4 (<https://github.com/rambaut/figtree/releases>).

*Analysis of discordance between derived mutations and position 17747.* In order to test the hypothesis that the BC data were spurious using the full collection of WA outbreak sequences generated, we examined derived mutations shared by sequences with either C or T at position 17747. We reasoned that, if the C at position 17747 in those 16 genomes were erroneous and those genomes in reality bore the C17747T mutation present in the overwhelming majority of WA outbreak (approaching 2000), the C17747 genomes might possess derived mutations that had occurred after the C17747T and A17585G substitutions. We downloaded from GISAID all genomes worldwide with C8782T, C18060T and T28144C (GISAID database accessed 7/7/2020) and then compared each of the 6 BC genomes with C17747 plus one or more derived mutations to all others in the data set. We then did the same thing for each of the other genomes with C17747T and A17585G, including one or two from California, Oregon, Minnesota (including the suspicious one mentioned above), Wyoming, Connecticut, Illinois, Florida, and Washington State.

## Supplementary Text

### Detection of WA outbreak sequences

To determine the time frame that sequences identical to the WA outbreak clade putative common ancestor were sampled, we downloaded an aligned, complete genome dataset provided by GISAID (accessed May 11th, 2020) and inferred a neighbor joining (NJ) phylogeny with the HKY substitution model in Geneious Prime 2019.1.3. Sequences that appeared to be identical in the phylogeny to the WA outbreak putative common ancestor were extracted in Geneious and manually examined to make sure they did not contain an ‘N’ at either site 17747 or 17858. The earliest and latest genomes containing the key mutations C17747T and A17858G were sampled at the earliest on Feb 20<sup>th</sup> (EPI\_ISL\_413456) and at the latest on April 27th (EPI\_ISL\_444051). This implies sequences identical to the WA outbreak clade putative common ancestor were circulating and detected for a minimum of 67 days. In contrast, the WA1 sequence was not detected again in another patient in Washington, despite intensive sampling. We do note however, that a single sequence identical to WA1 was sampled on March 13<sup>th</sup>, 2020 in NYC (EPI\_ISL\_427531).

Separately, we downloaded all available WA outbreak genomes sampled in Washington State (930) to see if any sampled after the pre-mitigation window (up to March 15th) fit any of the patterns represented in Fig. 2, yielding the basal polytomy genome WA-S566 (EPI\_ISL\_434326).

### Primary epidemic simulation in Washington State

The primary analysis consisted of 1,000 simulated epidemics seeded using the WA1 genome sequence on January 15<sup>th</sup> 2020. These epidemics had a median doubling time of 4.7 days (95% range of median doubling time across simulations: 4.2–5.1) and an evolutionary rate of  $0.8 \times 10^{-3}$  substitutions/site/year. These simulated epidemics produced a median of 4,269 total infections after 61 days (95% range: 1,993–

11,053 infections). The median genetic distance from WA1 to the sub-sampled viruses was three mutations, which is consistent with the observed phylogeny. The results from these simulations can be found in Main Text Fig. 2.

#### Simulating a slower epidemic doubling time in Washington State.

We simulated 1000 epidemics seeded using the WA1 genome sequence on January 15<sup>th</sup> 2020 with a slower median doubling time of 5.6 days (95% range: 5.2–5.9) and the original evolutionary rate of  $0.8 \times 10^{-3}$  substitutions/site/year. These simulated epidemics produced a median of 938 cases after 61 days (95% range: 681–1,779 cases). Of the 1000 simulations, 500 applied an infectiousness rate of 2.0, and the other 500 applied an infectiousness rate of 2.2. As in the primary simulations, we observed a median distance from WA1 to the subsampled viruses of 3 mutations. In 79.4% of simulations, we observed at least one identical virus to WA1, with a median of 10 identical viruses in each simulation (95% range: 0–62 identical viruses). Not observing a viral sequence identical to WA1 in the real Washington data does not significantly differ from expectation ( $p=0.206$ ). However, viruses with one mutation from WA1 were observed in 98.1% of simulations, indicating a low probability of not detecting a single sequence from Washington within one mutation of WA1 in the real data ( $p=0.019$ ). Lineages forming a basal polytomy with WA1 and the epidemic clade were observed in 99.6% of populations ( $p=0.004$ ) and 100% of simulations had at least one sibling lineage diverging prior to the formation of the outbreak clade ( $p<0.001$ ). Therefore, our results are robust to an assumption of slower epidemic growth and smaller size.

#### Simulating a faster rate of evolution in Washington State.

To explore the influence of a faster evolutionary rate on our simulations, we used the 1000 time trees from the main WA outbreak simulations and evolved the sequences with a substitution rate of  $1.6 \times 10^{-3}$  substitutions/site/year. These 1000 simulated epidemics produced a median doubling time of 4.7 days (95% range: 4.2–5.1) and a median of 4269 cases after 61 days (95% range: 1993–11053 cases). In a departure from the true phylogeny and our previous simulations, the median distance from WA1 to the subsampled viruses evolved under a faster evolutionary rate was 6 mutations. In 28.9% of simulations, we observed at least one identical virus to WA1, with a median of 0 identical viruses in each simulation (95% range: 0–16 identical viruses). Not observing a virus identical to WA1 in the real Washington data does not significantly differ from expectation ( $p=0.711$ ). In this case, viruses with one mutation from WA1 also did not significantly differ from expectation either ( $p=0.336$ ). However, lineages forming a basal polytomy with WA1 and the epidemic clade were observed in 98.6% of populations ( $p=0.013$ ) and 100% of simulations had at least one sibling lineage diverging prior to the formation of the outbreak clade ( $p<0.001$ ). In sum, the faster evolution scenario produced trees that experienced more evolutionary divergence than observed in real-data, we still find strong evidence against WA1 serving as the founder of WA outbreak clade, as all of the simulated epidemics experienced a divergent lineage prior to the formation of the outbreak clade.

#### Simulating a slower epidemic doubling time in the Italy Outbreak.

We simulated 1000 epidemics seeded using the BavPat1 genome sequence on January 28<sup>th</sup> 2020 with a slower median doubling time of 4.2 days (95% range: 3.6–5.0) and the original evolutionary rate of  $0.8 \times 10^{-3}$  substitutions/site/year. These simulated epidemics produced a median of 180 total infections after 36 days (95% range: 71–498 cases). Of the 1000 simulations, 500 applied an infectiousness rate of 2.0, and the other 500 applied an infectiousness rate of 2.2. The distributions of identical and one-mutation divergent viruses are not significantly different from expectation ( $p=0.186$  and  $p=0.273$ , respectively). However, the lack of at least one descendent lineage that forms a polytomy with BavPat1 and the Italian outbreak significantly differs from expectation ( $p=0.006$ ). Therefore, the results for the Italian outbreak are robust to an assumption of slower epidemic growth and smaller size.

#### Simulating a faster rate of evolution in the Italy Outbreak.

To explore the influence of a faster evolutionary rate on our simulations, we used the 1000 time trees from the main Italy outbreak simulations and evolved the sequences with a substitution rate of  $1.6 \times 10^{-3}$  substitutions/site/year. Of those simulations, 500 had an infectiousness rate of 2.6, and the other 500 had an infectiousness rate of 2.8. The 1000 simulated epidemics produced a median doubling time of 3.4 days (95% range: 2.9-4.4) and a median of 725 total infections after 36 days (95% range: 140-2847 cases). The distributions of identical and one-mutation divergent viruses are not significantly different from expectation ( $p=0.576$  and  $p=0.551$ , respectively). However, the lack of at least one descendant lineage that forms a polytomy with BavPat1 and the Italian outbreak significantly differs from expectation ( $p=0.01$ ). In sum, although the faster evolution scenario produced trees that experienced more evolutionary divergence than observed in real-data, we still find evidence against BavPat1 serving as the founder of the Italy outbreak clade, as 99% of the epidemics experienced basal divergence prior to the formation of the outbreak clade.

#### Time-homogeneous and time-inhomogeneous GLM-diffusion estimates

The time-homogeneous GLM diffusion model yields strong Bayes factor support ( $BF > 8,042$ ) for both air travel and the asymmetry coefficient, but no support for geographic distance ( $BF = 0.24$ ). The two significant covariates were associated with relatively large positive log effect sizes (0.97 [0.70,1.26] and 2.01 [1.42,2.60] for air travel and the asymmetry coefficient respectively).

The estimates for the time-inhomogeneous model we applied to a data set without unsampled taxa are summarized in Fig. S4. This model allows for time-variable effect sizes for the two significant covariates in the time-homogeneous model. The effect size estimates in the upper panel show that the contribution of air travel remains roughly constant through time whereas the flux out of Hubei varies strongly over time. In the lower panel, we summarize the posterior distribution of transitions out-of-Hubei over time based on the time-homogeneous analysis. This demonstrates that although the process is parameterized using a homogeneous model, the realizations of the process are strongly time-variable and in line with the time-variable flux out of Hubei, providing reassurance that our estimates are sufficiently robust with respect to the assumption of time-homogeneity.

#### Assessing the behavior of unsampled taxa.

We assess the placement of four unsampled taxa from Hubei, Italy, Iran and South Korea respectively in our phylogeographic approach (Fig. S9). We visualize all the nodes that are 'visited' by a single unsampled taxon from these particular locations; by visited, we mean that the unsampled taxon forms a sister lineage of that node. We note that many internal nodes in the tree we visualize may not appear to be visited by these unsampled taxa simply because the frequency of the node is low in the posterior distribution. However, the summary depicts in which parts of the phylogeny the unsampled taxa can be placed and how this differs by location.

The unsampled taxon from Hubei visits different parts of the phylogeny that are primarily associated with this location. As it represents the source of the outbreak, all other lineages are embedded in the Hubei diversity. The unsampled taxon from Italy is primarily placed in the European clade that gave rise to the cluster of sequences from New York City. This is where most other Italian isolates or isolates from travellers returning from Italy are clustering. The unsampled taxon from Italy also visits a cluster at the bottom of the circular tree where two other Italian viruses are clustering, and it is occasionally suggested as a single introduction from Hubei. The unsampled taxon from Iran exclusively visits a small cluster of viruses from travellers returning from Iran (primarily Australian but also a Canadian traveller). In this case the spatial structure represents a strong constraint on its placement. Finally, the unsampled taxon from South Korea has the most volatile placement as it seems to be able to visit almost every part of the phylogeny. This reflects the fact that i) the few sampled South Korean viruses are also dispersed over the tree as a number of independent lineages and ii) South Korea is well connected to different global locations by air travel. In contrast to the Italian and Iranian taxa, which reinforce specific spatial patterns, the unsampled taxa from South Korea contribute very little to the analysis.

### Estimating the date of early reported transmission clusters in Northern California

In order to determine the earliest case of sustained transmission in the United States based on genome data, we looked at other early transmission clusters reported by Deng et al (5), where they report 3 distinct early transmission chains in Northern California. One of these clusters is descended from the WA outbreak clade, and another transmission chain is represented by only 3 sequences, all collected from Solano county on 2/27. The 3rd transmission cluster contains 10 sequences from Santa Clara county collected from 2/29-3/8. To estimate the date of the tMRCA of this transmission chain, we downloaded the 10 publically available genomes and analyzed them in BEASTv1.10.4 (36) with the HKY substitution model, a strict clock rate prior of 8e-4 and an exponential growth tree prior. Following 10 million generations, all ESS values were >200. The estimated mean date for the TMRCA of these 10 sequences was 2/22 (95% HPD 2/5-2/29).

### Discordance between derived mutations and position 17747

An initial release of genomes from BC from early March showed 59% (16 out of 27) of sequences within or closely related to the WA outbreak clade only contain one of the two derived mutations which define the WA outbreak clade sequences compared to WA1 (17747/18858). These 16 early sequences from BC with C17747 suggested either (1) that the virus had actually first arrived in BC, then moved down into WA, since it is implausible that genomes having the transitional pattern would be sampled at high frequency in BC but would not be sampled in WA if the virus had first emerged in WA then spread to BC, or (2) that the apparently transitional genome in BC descended from an introduction from China/Asia to BC, separate from the one that led the WA outbreak genomes with both 17747 and 17858 derived mutations. This would also mean that in BC you had a mix of two separately-introduced lineages, since we also see C17747T/A17858G viruses there which must have spread north from WA.

Based on the totality of evidence, which we review below, we conclude that it is highly likely that all of the BC genomes in the Washington outbreak clade with the C17747 ancestral mutation are spurious and we speculate it represents a failure to remove PCR primer sequence prior to read assembly. First, more than 100 additional BC genomes from the WA outbreak clade were posted on GISAID beginning 6/6/2020, also sampled in March and April like the original 16 genomes that contained the spurious C17747 above. None of the Washington outbreak clade genomes from BC uploaded after 6/6/2020 have C17747; they all have C17747T, like the Washington State ones. In other words, the WA outbreak genomes sampled in BC from the same two-month period, but sequenced at different time points show strikingly different patterns: 59% of the ones generated earlier show the C17747 pattern, while 0% of the ones sequenced later do.

Second, we have heard from the author of another WA outbreak genome with C17747 (EPI\_ISL\_437361, from Minnesota) who responded that he has low confidence in the C17747 called at that position using the ARTIC Protocol as there was an almost even mix of C17747 and C17747T in sequencing reads for that sample (Jacob Garfin, personal communication). This suggests that there may be other genomes within the WA outbreak clade that show similar patterns.

Third, we used an approach of comparing derived mutations between genomes from similar geographic areas with either C or T at 17747. Globally, there are 1902 sequences (as of 7/6/2020) that contain the prototypical WA1-mutations C8782T, C18060T and T28144C. Of these, 28 genomes are C17747/A18858G rather than the derived T at position 17747 that all other WA outbreak sequences have. We examined all additional derived mutations in the 28 genomes. 18 of these genomes contain at least one derived mutation. We calculated the number of derived mutations that are shared by at least 2 sequences, the geographic location, and the relative abundance of each mutation out of the database frequency. Each genome shared derived mutations with genomes sampled in BC but having C17747T,

and in most cases were an exact match except for 17747. Some shared as many as 4 derived mutations, and these mutations were very rare, typically seen only in other BC viruses. In each case we found the same association between C17747 and C17747T genomes, with derived mutations being shared by sequences sampled in the same or nearby locations. The Washington one is the record-holder, sharing 9 such mutations with its C17747T counterpart. We summarize our results in Figure 3, which shows the pattern of shared derived mutations at all other sites besides 17747. Although this pattern does suggest the original 16 BC genomes with C17747 are spurious, there are a total of 10 ‘early’ BC genomes with C17747 and 35 ‘later’ genomes released following 6/6/2020 that contain no other derived mutations which can be compared. Regardless, we conclude based on the different strands of evidence that the BC genomes with C17747 are spurious. However we do note, if real, these genomes would have an impact on our inferences about the origin of the WA outbreak clade. We also note that there is a possibility these genomes with C17747 with other shared derived mutations are the result of reversions, but this seems unlikely given the huge difference in C17747 frequency between the ‘early’ and ‘late’ BC genomes from March which show completely different site frequencies.

**Table S1.** Parameters for epidemic and sequence simulations.

Parameter	Epidemic Simulation					
	WA (main text)	WA - slow spreading	WA - fast evolving	Germany/Italy (main text)	Germany/Italy - slow spreading	Germany/Italy - fast evolving
Days simulated	61	61	61	36	36	36
Number nodes in contact network	20000	5000	20000	5000	2500	5000
Average number of edges per node (degree)	16	16	16	16	16	16
Infectiousness	2.7	2.0-2.2	2.7	2.60-2.85	2.0-2.2	2.60-2.85
Scaled substitution rate (variant site subs/site/year)	0.0456	0.0456	0.0911	0.2647	0.2647	0.5293

<sup>1</sup>Additional parameters can be found in the example JSON files at [https://github.com/Worobeylab/SC2\\_outbreak](https://github.com/Worobeylab/SC2_outbreak).

Table S2. Derived mutations shared between C17747- and C17747T-containing viruses

Reference Position	2883	7561	17747	19109	20491
ancestral nucleotide	G	G	C	G	G
hCoV-19/Canada/BC_8606204/2020 EPI_ISL_418850 2020-03-13 NorthAmerica	A	A	C	T	T
hCoV-19/Canada/BC_24232310/2020 EPI_ISL_463201 2020-03-00 NorthAmerica	A	A	T	T	T
hCoV-19/Canada/BC_21282517/2020 EPI_ISL_462804 2020-03-00 NorthAmerica	G	A	T	G	T
hCoV-19/Canada/BC_22056219/2020 EPI_ISL_462823 2020-03-00 NorthAmerica	G	A	T	G	T
Reference Position	5170	16468	17673	17747	
ancestral nucleotide	C	C	T	C	
hCoV-19/Canada/BC_4799711/2020 EPI_ISL_418830 2020-03-09 NorthAmerica	T	T	C	C	
hCoV-19/Canada/BC_00714140/2020 EPI_ISL_466701 2020-04-00 NorthAmerica	T	C	C	T	
hCoV-19/USA/WA-S539/2020 EPI_ISL_434299 2020-03-23 NorthAmerica	T	T	C	T	
hCoV-19/USA/UN-UW-1331/2020 EPI_ISL_422968 2020-03-00 NorthAmerica	T	T	C	T	
hCoV-19/USA/WA-UW-4014/2020 EPI_ISL_437805 2020-03-28 NorthAmerica	T	T	C	T	
hCoV-19/USA/CA-SR0247/2020 EPI_ISL_437563 2020-03-27 NorthAmerica	T	C	C	T	
Reference Position	16457	17747	22480		
ancestral nucleotide	C	C	C		
hCoV-19/USA/FL_5091/2020 EPI_ISL_419560 2020-02-28 NorthAmerica	T	C	T		
hCoV-19/USA/WI-UW-59/2020 EPI_ISL_421317 2020-03-24 NorthAmerica	T	T	C		

hCoV-19/USA/WI-WSLH-200078/2020 EPI_ISL_471243 2020-04-03 NorthAmerica	T	T	C
hCoV-19/Australia/VIC222/2020 EPI_ISL_419917 2020-03-20 Oceania	C	C	T
hCoV-19/Australia/VIC130/2020 EPI_ISL_419823 2020-03-21 Oceania	C	C	T

Reference Position	222	4456	6693	14877	17747	27228	27775	27776	27777	29188
<b>ancestral nucleotide</b>	C	C	A	C	C	T	T	T	G	A
hCoV-19/USA/WA-S856/2020 EPI_ISL_463346 2020-04-24 NorthAmerica	T	T	G	T	C	A	C	G	A	T
hCoV-19/USA/WA-S871/2020 EPI_ISL_463361 2020-04-24 NorthAmerica	T	T	G	T	T	A	C	G	A	T
hCoV-19/USA/WA-S894/2020 EPI_ISL_463384 2020-04-28 NorthAmerica	T	T	G	T	T	A	C	G	A	T
hCoV-19/USA/WA-S853/2020 EPI_ISL_463343 2020-04-24 NorthAmerica	T	T	G	T	T	A	C	G	A	T
hCoV-19/USA/WA-UW-1538/2020 EPI_ISL_424169 2020-03-19 NorthAmerica	T	C	A	C	T	T	T	T	G	A
hCoV-19/USA/WA-S852/2020 EPI_ISL_463342 2020-04-24 NorthAmerica	T	T	G	T	T	A	C	G	A	T
hCoV-19/USA/WA-UW-4051/2020 EPI_ISL_437813 2020-04-01 NorthAmerica	T	C	G	C	T	T	T	T	G	A
hCoV-19/USA/WA-S857/2020 EPI_ISL_463347 2020-04-24 NorthAmerica	T	T	G	T	T	A	C	G	A	T
hCoV-19/USA/WA-S854/2020 EPI_ISL_463344 2020-04-24 NorthAmerica	T	T	G	T	T	A	C	G	A	T
hCoV-19/USA/UNKNOWN-UW-5622/2020 EPI_ISL_461438 2020-04-08 NorthAmerica	C	T	G	T	T	A	T	T	G	T
hCoV-19/USA/WA-S872/2020 EPI_ISL_463362 2020-04-24 NorthAmerica	C	T	G	T	T	A	T	T	G	T
hCoV-19/USA/UN-UW-4645/2020 EPI_ISL_429628 2020-04-06 NorthAmerica	C	T	G	T	T	A	T	T	G	T
46 sequences from Washington	C	C	G	T	T	T	T	G	A	T

Reference Position	16467	17747	23185	29200
<b>ancestral nucleotide</b>	A	C	C	C
hCoV-19/USA/CA-SCCPHD-UC161/2020 EPI_ISL_435640 2020-03-12 NorthAmerica	G	C	T	T

<b>hCoV-19/USA/CA-SCCPHD-UC140/2020 EPI_ISL_435619 2020-03-10 NorthAmerica</b>	<b>G</b>	<b>C</b>	<b>T</b>	<b>T</b>
hCoV-19/USA/CA-CDPH-UC7/2020 EPI_ISL_413925 2020-03-05 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/UT-03470/2020 EPI_ISL_450539 2020-03-15 NorthAmerica	<b>G</b>	<b>C</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/WI-UW-399/2020 EPI_ISL_480381 2020-03-16 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/MN_0100/2020 EPI_ISL_426426 2020-03-05 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/MN1-MDH1/2020 EPI_ISL_414588 2020-03-05 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/FL_9656/2020 EPI_ISL_424856 2020-03-06 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/CA-CDPH-UC9/2020 EPI_ISL_413928 2020-03-05 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/CA-SCCPHD-UC152/2020 EPI_ISL_435631 2020-03-11 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>T</b>
hCoV-19/USA/CA-CDPH-UC8/2020 EPI_ISL_413926 2020-03-05 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/Mexico/CDMX-InDRE_09/2020 EPI_ISL_455432 2020-03-11 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>T</b>
hCoV-19/USA/CA-CDPH-UC11/2020 EPI_ISL_413931 2020-03-05 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>
hCoV-19/USA/CA-SR0311/2020 EPI_ISL_445165 2020-03-17 NorthAmerica	<b>G</b>	<b>T</b>	<b>T</b>	<b>C</b>

Reference Position	14724	17747	29477
<b>ancestral nucleotide</b>	<b>C</b>	<b>C</b>	<b>G</b>
<b>hCoV-19/Canada/BC_6502001/2020 EPI_ISL_418840 2020-03-12 NorthAmerica</b>	<b>T</b>	<b>C</b>	<b>T</b>
hCoV-19/Canada/BC_22032711/2020 EPI_ISL_462821 2020-03-00 NorthAmerica	<b>T</b>	<b>T</b>	<b>G</b>
hCoV-19/USA/HI_7881/2020 EPI_ISL_426421 2020-03-07 NorthAmerica	<b>T</b>	<b>T</b>	<b>G</b>
hCoV-19/Canada/BC_3842755/2020 EPI_ISL_418820 2020-03-09 NorthAmerica	<b>T</b>	<b>T</b>	<b>G</b>
hCoV-19/Canada/BC_27222254/2020 EPI_ISL_463262 2020-03-00 NorthAmerica	<b>T</b>	<b>T</b>	<b>G</b>

hCoV-  
 19/Canada/BC\_26590284/2020|EPI\_ISL\_463240|2020-  
 03-00|NorthAmerica      T      T      G

Reference Position	17747	29736
ancestral nucleotide	C	G
hCoV- 19/Canada/BC_83109/2020 EPI_ISL_415588 2020-03- 05 NorthAmerica	C	T
hCoV- 19/Canada/BC_4122951/2020 EPI_ISL_418826 2020-03- 10 NorthAmerica	C	T
hCoV- 19/Canada/BC_41851/2020 EPI_ISL_415584 2020-03- 02 NorthAmerica	C	T
hCoV- 19/Canada/BC_08719828/2020 EPI_ISL_460607 2020- 03-00 NorthAmerica	T	T
hCoV- 19/Canada/BC_18111443/2020 EPI_ISL_460616 2020- 03-00 NorthAmerica	T	T
hCoV- 19/Canada/BC_20432317/2020 EPI_ISL_462774 2020- 03-00 NorthAmerica	T	T
hCoV- 19/Canada/BC_83163/2020 EPI_ISL_415589 2020-03- 05 NorthAmerica	T	T

Reference Position	15240	17747	20178	21774	25556
ancestral nucleotide	C	C	C	C	T
hCoV-19/USA/OR-OHSU- 0049/2020 EPI_ISL_468086 2020-04-16 NorthAmerica	T	C	T	T	G
hCoV-19/USA/OR-OHSU- 0051/2020 EPI_ISL_437531 2020-04-17 NorthAmerica	T	T	T	T	G
hCoV-19/USA/OR-UW- 1849/2020 EPI_ISL_424286 2020-03-20 NorthAmerica	T	T	C	T	T
hCoV-19/USA/CT-Yale- 041/2020 EPI_ISL_428372 2020-03-16 NorthAmerica	T	T	C	C	T
hCoV-19/USA/WA-S90/2020 EPI_ISL_417143 2020-02- 29 NorthAmerica	T	T	C	T	T
hCoV-19/USA/AZ- TG268903/2020 EPI_ISL_426513 2020-03- 11 NorthAmerica	T	T	C	T	T

hCoV-19/USA/WA-S89/2020|EPI\_ISL\_417142|2020-02-29|NorthAmerica T T C T T

**Reference Position** 10771 17747

	T	C
ancestral nucleotide		
hCoV-19/USA/CT-Yale-007/2020 EPI_ISL_416421 2020-03-13 NorthAmerica	C	C
27 other sequences (14 from Canada, 7 from E. or Midwest US states)	C	T

**Reference Position** 17747 21137 24694

	C	A	A
ancestral nucleotide			
hCoV-19/USA/WY-WYPHL-00031/2020 EPI_ISL_462924 2020-03-23 NorthAmerica	C	G	T
hCoV-19/USA/WY-WYPHL-00032/2020 EPI_ISL_462925 2020-03-23 NorthAmerica	C	G	T
hCoV-19/USA/MN-MDH-284/2020 EPI_ISL_437361 2020-03-29 NorthAmerica	C	A	T
hCoV-19/USA/WY-WYPHL018/2020 EPI_ISL_421552 2020-03-19 NorthAmerica	T	G	T
hCoV-19/USA/MN-MDH-322/2020 EPI_ISL_450780 2020-03-30 NorthAmerica			
hCoV-19/USA/IL-NM09/2020 EPI_ISL_444522 2020-03-17 NorthAmerica	C	A	T
hCoV-19/USA/IL-NM022/2020 EPI_ISL_444532 2020-03-15 NorthAmerica	C	A	T
hCoV-19/Australia/VIC636/2020 EPI_ISL_426877 2020-03-28 Oceania	C	A	T
215 additional sequences	T	A	T

**Reference Position** 1238 17747 28924

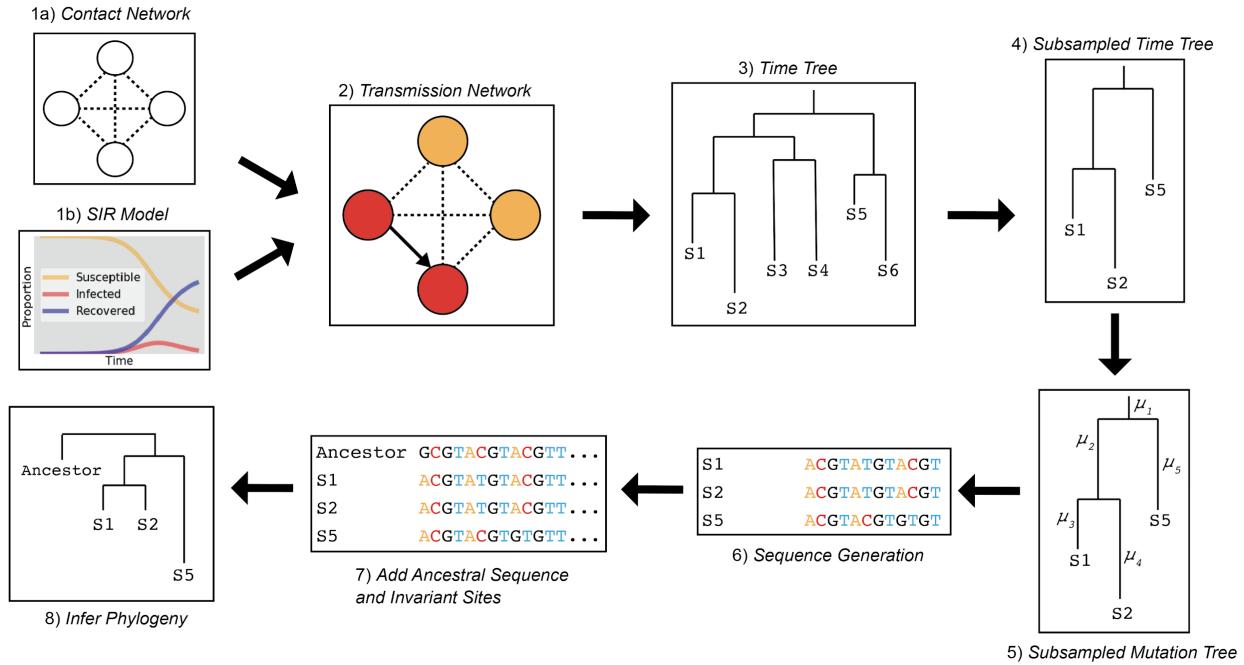
	C	C	T
ancestral nucleotide			
hCoV-19/USA/MN-MDH-1284/2020 EPI_ISL_483005 2020-03-20	T	C	C
hCoV-19/USA/MN-UMGC-2/2020 EPI_ISL_443253 2020-04-04	T	T	C

hCoV-19/USA/MN-MDH- 163/2020 EPI_ISL_460058 2020-03-23	T	T	C
hCoV-19/USA/MN-MDH- 148/2020 EPI_ISL_460053 2020-03-21	T	T	C
10 additional sequences from WA	T	T	C
	T		

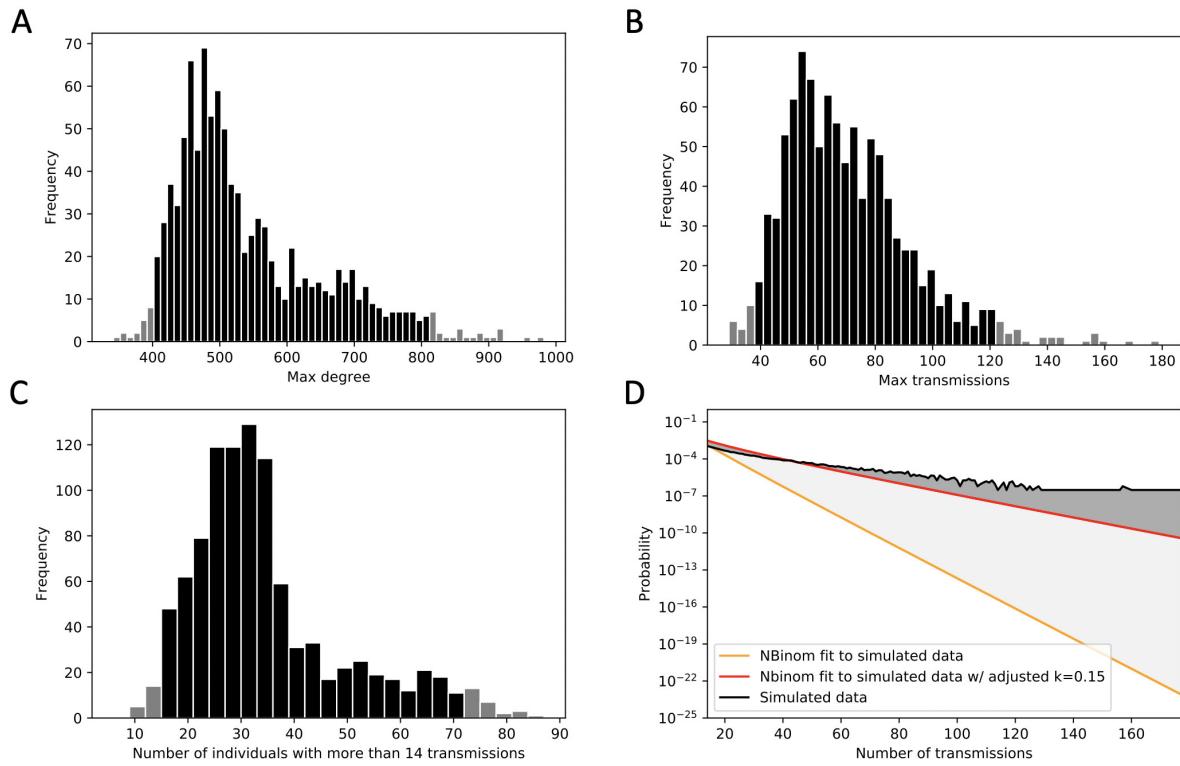
**Table S3.** Prior distributions for the ages of the unsampled taxa by location.

Unsampled taxon location	Unsampled taxon age prior
CN Hubei	normal(0.139, 0.021)*
CN Henan	normal(0.151, 0.025)
CN Zhejiang	normal(0.161, 0.018)
CN Hunan	normal(0.156, 0.022)
CN Jiangxi	normal(0.151, 0.018)
Italy	normal(0.0167, 0.038)
Iran	normal(0.066, 0.015)
South Korea	normal(0.086, 0.017)
France	exp(0.011), offset = 0.065
Spain	exp(0.009), offset = 0.065
Germany	exp(0.009), offset = 0.065
Sweden	exp(0.012), offset = 0.065

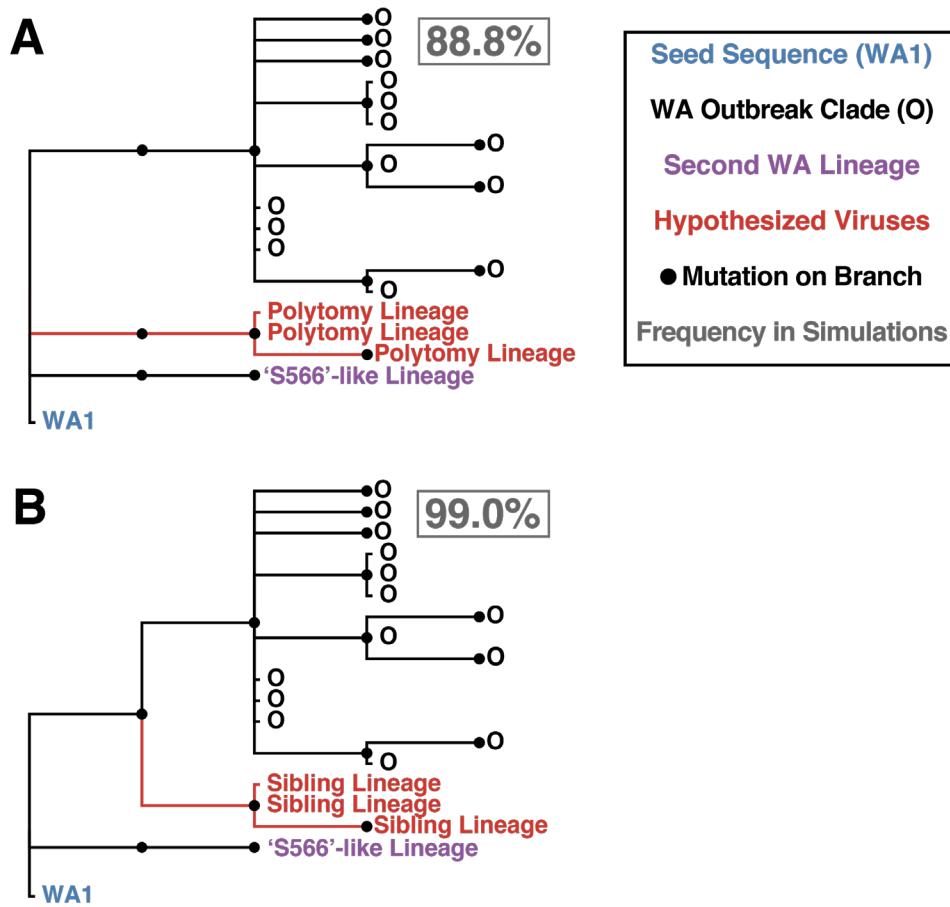
\*We list the means and standard deviations for the normal prior distributions and the means for the exponential prior distributions. Normal priors are truncated at 0 because tip and node heights are strictly positive parameters. We use offsets for the exponential priors because these distributions were determined based on case count data available up to March 5th while more recent available sequences were added that effectively set time point 0 on March 29th.



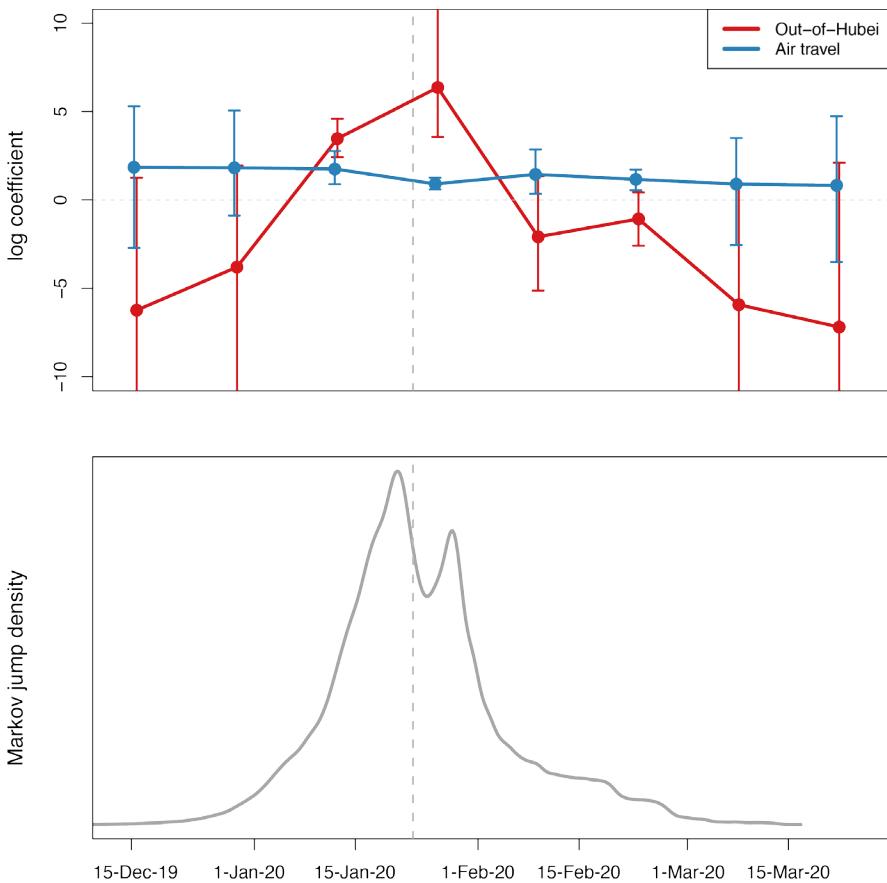
**Fig. S1 Epidemic simulation workflow.** (1) FAVITES generates a preferential attachment contact network (1a), runs an SIR model (1b) to simulate spread through the contact network and (2) produces a transmission network. (3) FAVITES outputs a viral time tree based on the infected individuals in the transmission network from which (4) 294 individuals are subsampled to match the dates of the original epidemic (e.g., WA outbreak clade). (5) Evolutionary rates are applied to the time tree based on the number of variant sites from the original alignment, converting the branches from years to substitutions/site ( $\mu$ ). (6) Genetic sequences at variant sites are evolved over the subsampled tree using Pyvolve, starting with the simulation seed sequence (i.e., WA1), using a general time-reversible (GTR) parameters inferred from the original alignment. (7) The ancestral sequence and invariant sites are added to the sequence data so (8) a maximum likelihood phylogeny can be inferred in IQ-TREE2.



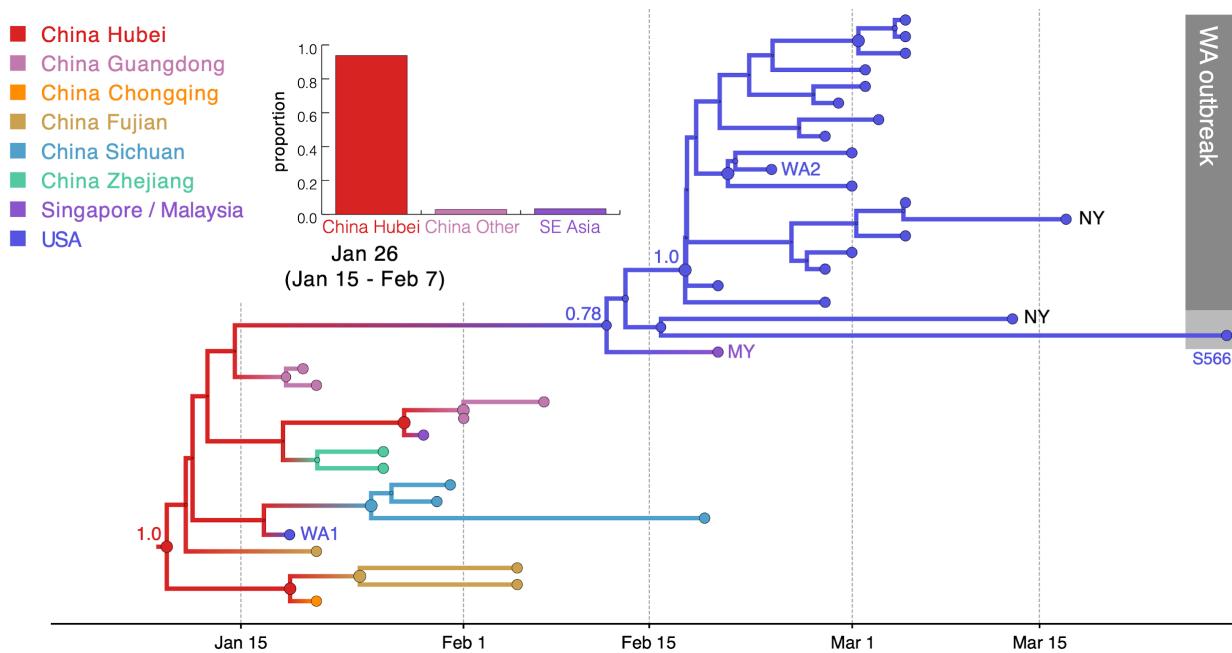
**Figure S2. Characterization of super-spreading events in 1000 FAVITES simulations of Washington outbreak clade.** (A) Maximum degree in each contact network. (B) Maximum degree in each transmission network. (C) Number of nodes with at least 14 transmission events in each transmission network. (D) High degree nodes in the simulated epidemics relative to expectation under a negative binomial distribution. A negative binomial distribution was fit to all the simulated transmissions combined and generated up to the maximum number of transmissions from the simulated data. A second negative binomial was generated using the parameters from the distribution fit to the simulated data, except with an adjusted  $k$  value of 0.15 to simulate overdispersion.



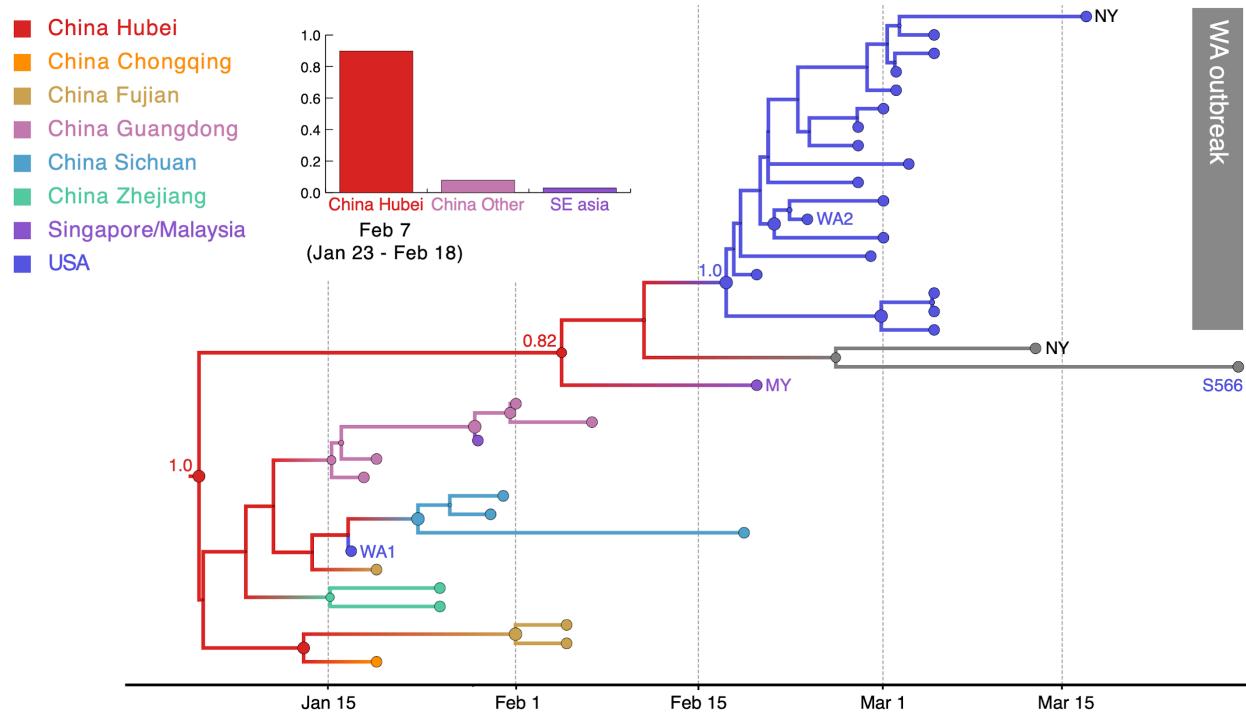
**Figure S3. Potential phylogenetic relationships between WA1, the WA outbreak clade, and strain S566 and their occurrence frequencies in 1000 epidemic simulations.** (A) At least 3 lineages forming basal polytomies with WA1, including the outbreak clade and S566. (B) At least 2 sibling lineages to the outbreak clade (3 lineages relative to WA1) experiencing fewer than two mutations from WA1 before divergence. These sibling lineages may have experienced a single mutation (as shown) or zero mutations (i.e., basal divergence). The frequency of each relationship across 1000 simulations is reported in the gray box.



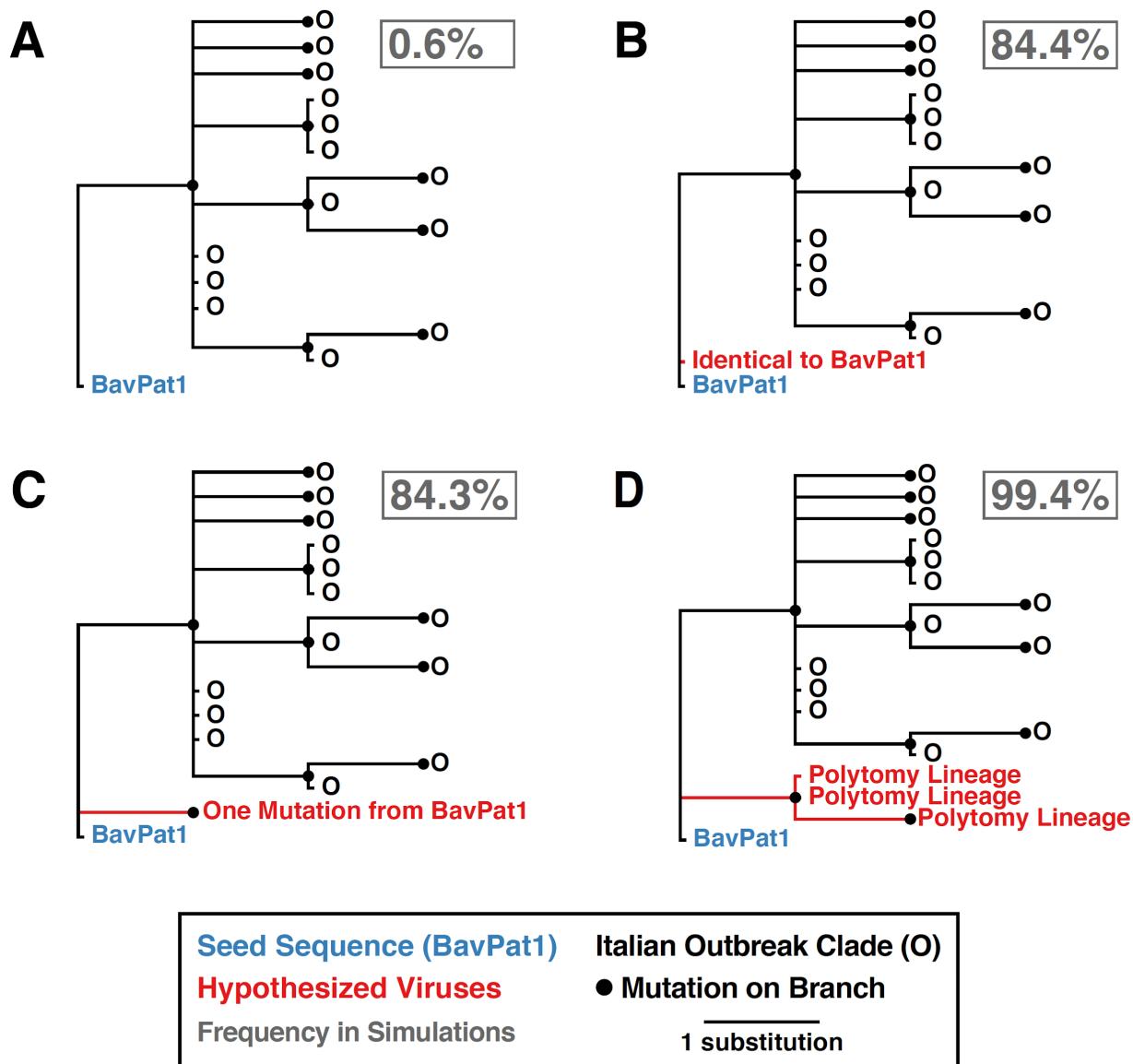
**Figure S4: Time-inhomogeneous and time-homogeneous phylogeographic estimates.** The upper panel summarizes the posterior log effect size estimates over time in the time-varying GLM-diffusion model fitted to the data set without unsampled diversity. The effect size estimates in the upper panel show that the contribution of air travel remains roughly constant through time whereas the flux out of Hubei varies strongly over time. The lower panel summarizes the posterior density of transitions out-of-Hubei estimated using Markov jumps in a time-homogeneous model fitted to the data set with unsampled diversity. This demonstrates that although the process is parameterized using a homogeneous model, the realizations of the process are strongly time-variable, providing reassurance that our estimates are sufficiently robust with respect to the assumption of time-homogeneity.



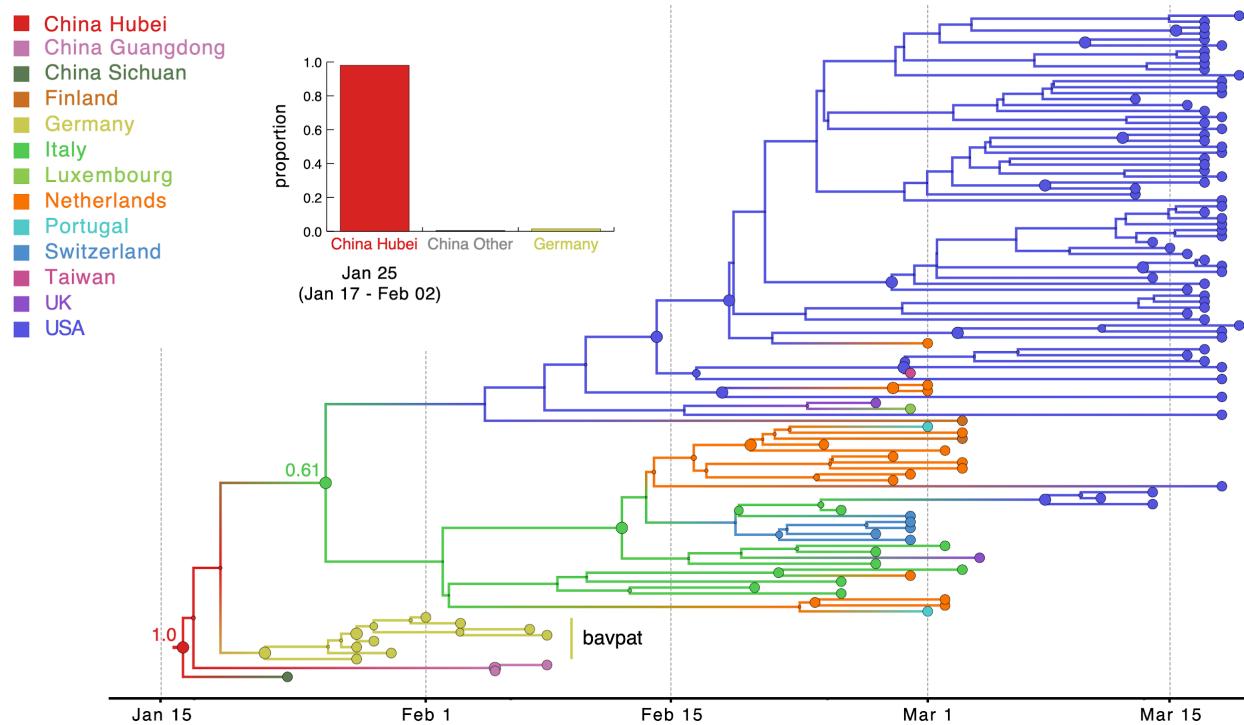
**Figure S5: MCC tree of SARS-CoV-2 entry into Washington State based on a reconstruction without unsampled diversity.** A subtree of the maximum clade credibility (MCC) tree depicting the evolutionary relationships inferred between (i) the first identified SARS-CoV-2 case in the US (WA1); (ii) the C17747T/A17858G clade associated with the Washington State outbreak (including WA2), plus closely related C17747/A17858 viruses (WA-S566 and NY); and (iii) closely related viruses that were identified in multiple locations in Asia. Genomes sampled at the tips of the phylogeny are represented by circles that are shaded according to location of sampling. Internal node circles, representing posterior clade support values, and branches are shaded similarly by modal posterior location state. Posterior location state probabilities are shown for two well-supported key nodes. The bar chart summarizes the probability by location for a second introduction giving rise to the WA outbreak clade. The mean date and 95% HPD intervals represent the estimate for the time of the introduction from Hubei.



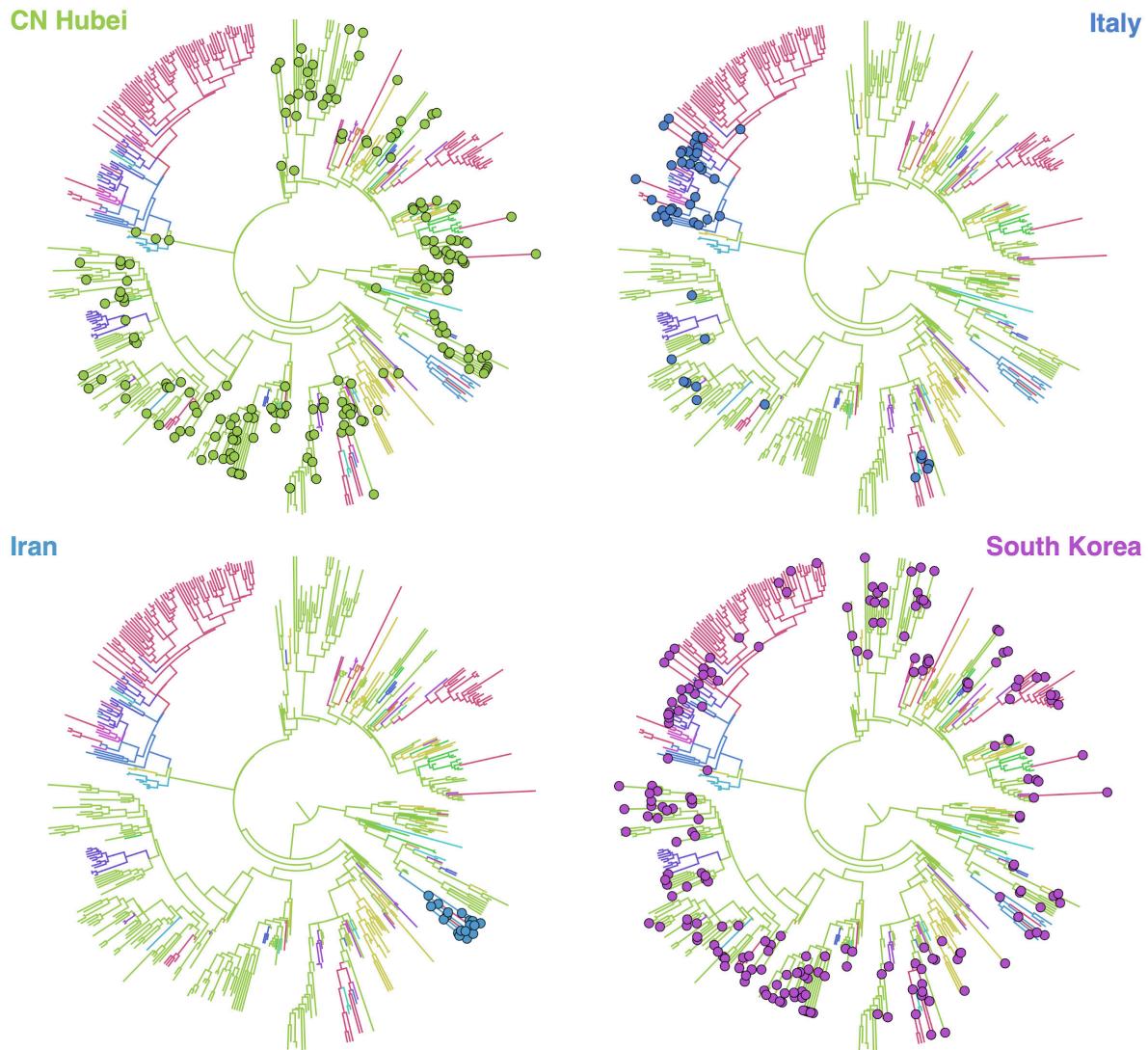
**Figure S6: MCC tree of SARS-CoV-2 entry into Washington State based on a reconstruction that does not assign a known location to two C17747/A17858 viruses (WA-S566 and NY).** A subtree of the maximum clade credibility (MCC) tree depicting the evolutionary relationships inferred between (i) the first identified SARS-CoV-2 case in the US (WA1); (ii) the C17747T/A17858G clade associated with the Washington State outbreak (including WA2), plus closely related C17747/A17858 viruses (WA-S566 and NY); and (iii) closely related viruses that were identified in multiple locations in Asia. Genomes sampled at the tips of the phylogeny are represented by circles that are shaded according to location of sampling. Internal node circles, representing posterior clade support values, and branches are shaded similarly by modal posterior location state. The grey color for WA-S566 and NY reflects the fact that they were not assigned to a known location. Posterior location state probabilities are shown for three well-supported key nodes. The bar chart summarizes the probability by location for a second introduction giving rise to the WA outbreak clade. The mean date and 95% HPD intervals represent the estimate for the time of the introduction from Hubei.



**Figure S7. Potential phylogenetic relationships between BavPat1 and the Italian outbreak and their occurrence frequencies in 1000 epidemic simulations.** (A) Observed pattern where the BavPat1 genome is the direct ancestor of the outbreak clade, separated by at least one mutation. (B) Identical sequence to BavPat1. (C) Sequences that are one mutation divergent from BavPat1. (D) Lineage forming a basal polytomy with BavPat1 and the outbreak clade. Frequency of each relationship was observed in 1000 simulations reported in gray box. Note that although BavPat1 was used as the seed sequence for the simulations to generate evolutionary patterns under the constraint that the Italian outbreak descended from the German one, that does not imply that it was in reality the seed sequence. Indeed, the simulations indicate that the observed patterns are inconsistent with those expected if it were.



**Figure S8: MCC tree of SARS-CoV-2 entry into Europe based on a reconstruction without unsampled diversity.** A subtree inferred for viruses from (i) the first outbreak in Europe (Germany, BatPat1) and identical viruses from China, (ii) outbreaks in Italy and New York, and (iii) other locations in Europe. Country codes are shown at tips for genomes sampled from travellers returning from Italy. The bar chart summarizes the probability distribution for the location state ancestral to the Italian clade. Other features as described in Fig. S5.



**Figure S9.** Bayesian phylogenetic placement of an unsampled taxon from Hubei, Italy, Iran and South Korea respectively. The colored circles indicate nodes that are visited by each unsampled taxon; visiting a node implies that the unsampled taxon is placed as a sister lineage of that node.

## References and Notes

1. T. B. (@trvrb), “The team at the @seattleflustudy have sequenced the genome the #COVID19 community case reported yesterday from Snohomish County, WA, and have posted the sequence publicly to <http://gisaid.org>. There are some enormous implications here.” Twitter (2020); <https://twitter.com/trvrb/status/1233970271318503426>.
2. T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M.-L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Fazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, The Seattle Flu Study Investigators, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautam, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, Cryptic transmission of SARS-CoV-2 in Washington state. *Science* 10.1126/science.abc0523 (2020). [doi:10.1126/science.abc0523](https://doi.org/10.1126/science.abc0523)
3. J. R. Fauver, M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, A. G. Watts, C. B. F. Vogels, A. F. Brito, T. Alpert, A. Muyombwe, J. Razeq, R. Downing, N. R. Cheemarla, A. L. Wyllie, C. C. Kalinich, I. M. Ott, J. Quick, N. J. Loman, K. M. Neugebauer, A. L. Greninger, K. R. Jerome, P. Roychoudhury, H. Xie, L. Shrestha, M.-L. Huang, V. E. Pitzer, A. Iwasaki, S. B. Omer, K. Khan, I. I. Bogoch, R. A. Martinello, E. F. Foxman, M. L. Landry, R. A. Neher, A. I. Ko, N. D. Grubaugh, Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990–996.e5 (2020). [doi:10.1016/j.cell.2020.04.021](https://doi.org/10.1016/j.cell.2020.04.021) [Medline](#)
4. A. S. Gonzalez-Reiche, M. M. Hernandez, M. J. Sullivan, B. Ciferri, H. Alshammary, A. Obla, S. Fabre, G. Kleiner, J. Polanco, Z. Khan, B. Alburquerque, A. van de Guchte, J. Dutta, N. Francoeur, B. S. Melo, I. Oussenko, G. Deikus, J. Soto, S. H. Sridhar, Y.-C. Wang, K. Twyman, A. Kasarskis, D. R. Altman, M. Smith, R. Sebra, J. Aberg, F. Krammer, A. García-Sastre, M. Luksza, G. Patel, A. Paniz-Mondolfi, M. Gitman, E. M. Sordillo, V. Simon, H. van Bakel, Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020). [doi:10.1126/science.abc1917](https://doi.org/10.1126/science.abc1917) [Medline](#)
5. X. Deng, W. Gu, S. Federman, L. du Plessis, O. G. Pybus, N. R. Faria, C. Wang, G. Yu, B. Bushnell, C.-Y. Pan, H. Guevara, A. Sotomayor-Gonzalez, K. Zorn, A. Gopez, V. Servellita, E. Hsu, S. Miller, T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, H. Y. Chu, J. Shendure, K. R. Jerome, C. Anderson, K. Gangavarapu, M. Zeller, E. Spencer, K. G. Andersen, D. MacCannell, C. R. Paden, Y. Li, J. Zhang, S. Tong, G. Armstrong, S. Morrow, M. Willis, B. T. Matyas, S. Mase, O. Kasirye, M. Park, G. Masinde, C. Chan, A. T. Yu, S. J. Chai, E. Villarino, B. Bonin, D. A. Wadford, C. Y. Chiu, Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582–587 (2020). [doi:10.1126/science.abb9263](https://doi.org/10.1126/science.abb9263) [Medline](#)
6. T. B. (@trvrb), “This separate introduction may have been to British Columbia or may have been elsewhere. Better resolving this introduction geographically would benefit from

additional sequencing of samples collected closer in time to the introduction event. 14/18.” Twitter (2020); <https://twitter.com/trvrb/status/1265063937663328256>.

7. M. L. Holshue, C. DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, A. Cohn, L. Fox, A. Patel, S. I. Gerber, L. Kim, S. Tong, X. Lu, S. Lindstrom, M. A. Pallansch, W. C. Weldon, H. M. Biggs, T. M. Uyeki, S. K. Pillai; Washington State 2019-nCoV Case Investigation Team, First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* **382**, 929–936 (2020). [doi:10.1056/NEJMoa2001191](https://doi.org/10.1056/NEJMoa2001191) [Medline](#)
8. A. Harmon, “Inside the Race to Contain America’s First Coronavirus Case.” *The New York Times* (2020); [www.nytimes.com/2020/02/05/us/corona-virus-washington-state.html](http://www.nytimes.com/2020/02/05/us/corona-virus-washington-state.html).
9. D. A. Bolduc, “Webasto disputes link to Italy coronavirus outbreak.” *Automotive News* (2020); [www.autonews.com/suppliers/webasto-disputes-link-italy-coronavirus-outbreak](http://www.autonews.com/suppliers/webasto-disputes-link-italy-coronavirus-outbreak).
10. M. M. Böhmer, U. Buchholz, V. M. Corman, M. Hoch, K. Katz, D. V. Marosevic, S. Böhm, T. Woudenberg, N. Ackermann, R. Konrad, U. Eberle, B. Treis, A. Dangel, K. Bengs, V. Fingerle, A. Berger, S. Hörmansdorfer, S. Ippisch, B. Wicklein, A. Grahl, K. Pörtner, N. Muller, N. Zeitlmann, T. S. Boender, W. Cai, A. Reich, M. An der Heiden, U. Rexroth, O. Hamouda, J. Schneider, T. Veith, B. Mühlmann, R. Wölfel, M. Antwerpen, M. Walter, U. Protzer, B. Liebl, W. Haas, A. Sing, C. Drosten, A. Zapf, Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: A case series. *Lancet Infect. Dis.* **20**, 920–928 (2020). [doi:10.1016/S1473-3099\(20\)30314-5](https://doi.org/10.1016/S1473-3099(20)30314-5) [Medline](#)
11. C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, T. Zimmer, V. Thiel, C. Janke, W. Guggemos, M. Seilmaier, C. Drosten, P. Vollmar, K. Zwirglmaier, S. Zange, R. Wölfel, M. Hoelscher, Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *N. Engl. J. Med.* **382**, 970–971 (2020). [doi:10.1056/NEJMc2001468](https://doi.org/10.1056/NEJMc2001468) [Medline](#)
12. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9241–9243 (2020). [doi:10.1073/pnas.2004999117](https://doi.org/10.1073/pnas.2004999117) [Medline](#)
13. A. Rambaut, E. C. Holmes, V. Hill, Á. O’Toole, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv (2020); <https://doi.org/10.1101/2020.04.17.046086>.
14. M. T. Maurano, S. Ramaswami, G. Westby, P. Zappile, D. Dimartino, G. Shen, X. Feng, A. M. Ribeiro-dos-Santos, N. A. Vulpescu, M. Black, M. Hogan, C. Marier, P. Meyn, Y. Zhang, J. Cadley, R. Ordonez, R. Luther, E. Huang, E. Guzman, A. Serrano, B. Belovarac, T. Gindin, A. Lytle, J. Pinnell, T. Vougiouklakis, L. Boytard, J. Chen, L. H. Lin, A. Rapkiewicz, V. Raabe, M. I. Samanovic-Golden, G. Jour, I. Osman, M. Aguero-Rosenfeld, M. J. Mulligan, P. Cotzia, M. Snuderl, A. Heguy, Sequencing identifies multiple, early introductions of SARS-CoV2 to New York City Region. medRxiv (2020); [doi:10.1101/2020.04.15.20064931](https://doi.org/10.1101/2020.04.15.20064931).
15. E. Minskaia, T. Hertzig, A. E. Gorbalenya, V. Campanacci, C. Cambillau, B. Canard, J. Ziebuhr, Discovery of an RNA virus 3'→5' exoribonuclease that is critically involved in

coronavirus RNA synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5108–5113 (2006).  
[doi:10.1073/pnas.0508200103](https://doi.org/10.1073/pnas.0508200103) [Medline](#)

16. Materials and methods are available as supplementary materials at the *Science* website.
17. P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, M. A. Suchard, Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLOS Pathog.* **10**, e1003932 (2014). [doi:10.1371/journal.ppat.1003932](https://doi.org/10.1371/journal.ppat.1003932) [Medline](#)
18. V. N. Minin, M. A. Suchard, Fast, accurate and simulation-free stochastic mapping. *Philos. Trans. R. Soc. B* **363**, 3985–3995 (2008). [doi:10.1098/rstb.2008.0176](https://doi.org/10.1098/rstb.2008.0176) [Medline](#)
19. H. Y. Chu, J. A. Englund, L. M. Starita, M. Famulare, E. Brandstetter, D. A. Nickerson, M. J. Rieder, A. Adler, K. Lacombe, A. E. Kim, C. Graham, J. Logue, C. R. Wolf, J. Heimonen, D. J. McCulloch, P. D. Han, T. R. Sibley, J. Lee, M. Ilcisin, K. Fay, R. Burstein, B. Martin, C. M. Lockwood, M. Thompson, B. Lutz, M. Jackson, J. P. Hughes, M. Boeckh, J. Shendure, T. Bedford; Seattle Flu Study Investigators, Early Detection of Covid-19 through a Citywide Pandemic Surveillance Platform. *N. Engl. J. Med.* **383**, 185–187 (2020). [doi:10.1056/NEJMc2008646](https://doi.org/10.1056/NEJMc2008646) [Medline](#)
20. The White House, “Proclamation on Suspension of Entry as Immigrants and Nonimmigrants of Persons who Pose a Risk of Transmitting 2019 Novel Coronavirus”; [www.whitehouse.gov/presidential-actions/proclamation-suspension-entry-immigrants-nonimmigrants-persons-pose-risk-transmitting-2019-novel-coronavirus/](https://www.whitehouse.gov/presidential-actions/proclamation-suspension-entry-immigrants-nonimmigrants-persons-pose-risk-transmitting-2019-novel-coronavirus/).
21. S. Eder, H. Fountain, M. H. Keller, M. Xiao, A. Stevenson, “430,000 People Have Traveled From China to U.S. Since Coronavirus Surfaced.” *The New York Times* (2020); [www.nytimes.com/2020/04/04/us/coronavirus-china-travel-restrictions.html](https://www.nytimes.com/2020/04/04/us/coronavirus-china-travel-restrictions.html).
22. “Preliminary analysis of SARS-CoV-2 importation and establishment of UK transmission lineages.” *Virological* (2020); <https://virological.org/t/preliminary-analysis-of-sars-cov-2-importation-establishment-of-uk-transmission-lineages/507>.
23. J. T. Ladner, B. B. Larsen, J. R. Bowers, C. M. Hepp, E. Bolyen, M. Folkerts, K. Sheridan, A. Pfeiffer, H. Yaglom, D. Lemmer, J. W. Sahl, E. A. Kaelin, R. Maqsood, N. A. Bokulich, G. Quirk, T. D. Watt, K. Komatsu, V. Waddell, E. S. Lim, J. G. Caporaso, D. M. Engelthaler, M. Worobey, P. Keim, Defining the Pandemic at the State Level: Sequence-Based Epidemiology of the SARS-CoV-2 virus by the Arizona COVID-19 Genomics Union (ACGU). *medRxiv* (2020); doi:[10.1101/2020.05.08.20095935](https://doi.org/10.1101/2020.05.08.20095935).
24. D. Stadlbauer, J. Tan, K. Jiang, M. Hernandez, S. Fabre, F. Amanat, C. Teo, G. Asthagiri Arunkumar, M. McMahon, J. Jhang, M. Nowak, V. Simon, E. Sordillo, H. van Bakel, F. Krammer, Seroconversion of a city: Longitudinal monitoring of SARS-CoV-2 seroprevalence in New York City. *medRxiv* (2020); doi:[10.1101/2020.06.28.20142190](https://doi.org/10.1101/2020.06.28.20142190).
25. M. Worobey, Epidemiology: Molecular mapping of Zika spread. *Nature* **546**, 355–357 (2017). [doi:10.1038/nature22495](https://doi.org/10.1038/nature22495) [Medline](#)
26. Zenodo, <http://doi.org/10.5281/zenodo.3979896>.
27. Z. Dezső, A.-L. Barabási, Halting viruses in scale-free networks. *Phys. Rev. E* **65**, 055103 (2002). [doi:10.1103/PhysRevE.65.055103](https://doi.org/10.1103/PhysRevE.65.055103) [Medline](#)

28. J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, W. J. Edmunds, Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Med.* **5**, e74 (2008). [doi:10.1371/journal.pmed.0050074](https://doi.org/10.1371/journal.pmed.0050074) [Medline](#)
29. J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, W. M. Getz, Superspreaders and the effect of individual variation on disease emergence. *Nature* **438**, 355–359 (2005). [doi:10.1038/nature04153](https://doi.org/10.1038/nature04153) [Medline](#)
30. S. J. Spielman, C. O. Wilke, Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLOS ONE* **10**, e0139047 (2015). [doi:10.1371/journal.pone.0139047](https://doi.org/10.1371/journal.pone.0139047) [Medline](#)
31. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020). [doi:10.1093/molbev/msaa015](https://doi.org/10.1093/molbev/msaa015) [Medline](#)
32. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016). [doi:10.1093/molbev/msw046](https://doi.org/10.1093/molbev/msw046) [Medline](#)
33. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). [doi:10.1093/molbev/mst010](https://doi.org/10.1093/molbev/mst010) [Medline](#)
34. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016). [doi:10.1093/ve/vew007](https://doi.org/10.1093/ve/vew007) [Medline](#)
35. N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, S. Mirarab, FAVITES: Simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics* **35**, 1852–1861 (2019). [doi:10.1093/bioinformatics/bty921](https://doi.org/10.1093/bioinformatics/bty921) [Medline](#)
36. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018). [doi:10.1093/ve/vey016](https://doi.org/10.1093/ve/vey016) [Medline](#)
37. D. L. Ayres, M. P. Cummings, G. Baele, A. E. Darling, P. O. Lewis, D. L. Swofford, J. P. Huelsenbeck, P. Lemey, A. Rambaut, M. A. Suchard, BEAGLE 3: Improved Performance, Scaling, and Usability for a High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019). [doi:10.1093/sysbio/syz020](https://doi.org/10.1093/sysbio/syz020) [Medline](#)
38. P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, H.-D. Chen, J. Chen, Y. Luo, H. Guo, R.-D. Jiang, M.-Q. Liu, Y. Chen, X.-R. Shen, X. Wang, X.-S. Zheng, K. Zhao, Q.-J. Chen, F. Deng, L.-L. Liu, B. Yan, F.-X. Zhan, Y.-Y. Wang, G.-F. Xiao, Z.-L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020). [doi:10.1038/s41586-020-2012-7](https://doi.org/10.1038/s41586-020-2012-7) [Medline](#)

39. F. Bielejec, P. Lemey, G. Baele, A. Rambaut, M. A. Suchard, Inferring heterogeneous evolutionary processes through time: From sequence substitution to phylogeography. *Syst. Biol.* **63**, 493–504 (2014). [doi:10.1093/sysbio/syu015](https://doi.org/10.1093/sysbio/syu015) [Medline](#)
40. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018). [doi:10.1093/sysbio/syy032](https://doi.org/10.1093/sysbio/syy032) [Medline](#)