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

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [2]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093, USA. [3]Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093, USA. [4]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA. [5]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK. [6]Department of Medicine, University of British Columbia, Vancouver, BC, Canada. [7]BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada. [8]Bioinformatics Programme, University of British Columbia, Vancouver, BC, Canada. [9]Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. [10]Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA 90095, USA. [11]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. [12]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA. [13]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 1×10^{-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 "superspreading" events (fig. S2)—and a fixed evolutionary rate of 0.8×10^{-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 nonsynonymous 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 1.6×10^{-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). 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. Gautom, 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
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 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 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 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 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.
9. D. A. Bolduc, "Webasto disputes link to Italy coronavirus outbreak." *Automotive News* (2020); 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. 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 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 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 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.
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 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 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 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 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/.
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.
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.
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.
25. M. Worobey, Epidemiology: Molecular mapping of Zika spread. *Nature* **546**, 355–357 (2017). doi: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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 Artic 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) and Zenodo (*26*).

SUPPLEMENTARY MATERIALS

science.sciencemag.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

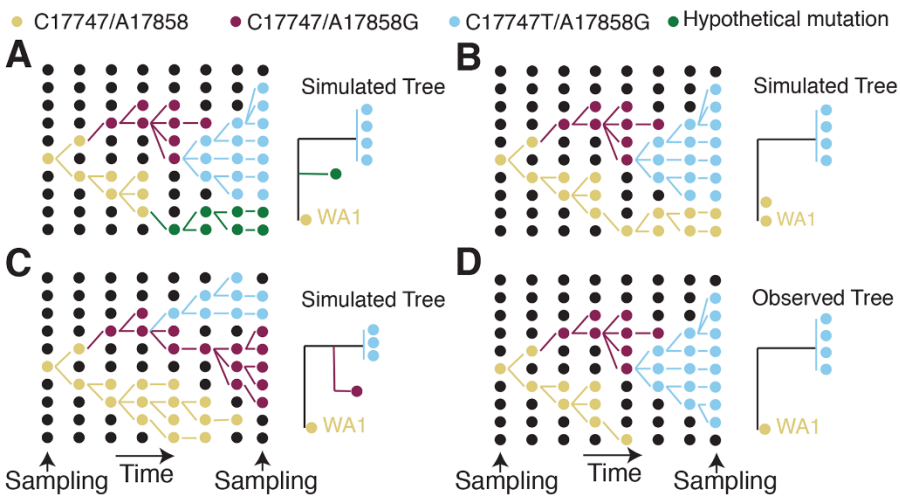

Fig. 1. Schematic showing a hypothetical path along which the key mutations in the WA outbreak could have taken in a susceptible population, alongside the inferred phylogeny. (**A**) Scenario where a hypothetical mutation occurs from WA1-like genomes. (**B**) A hypothetical phylogeny where A17747 and 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.

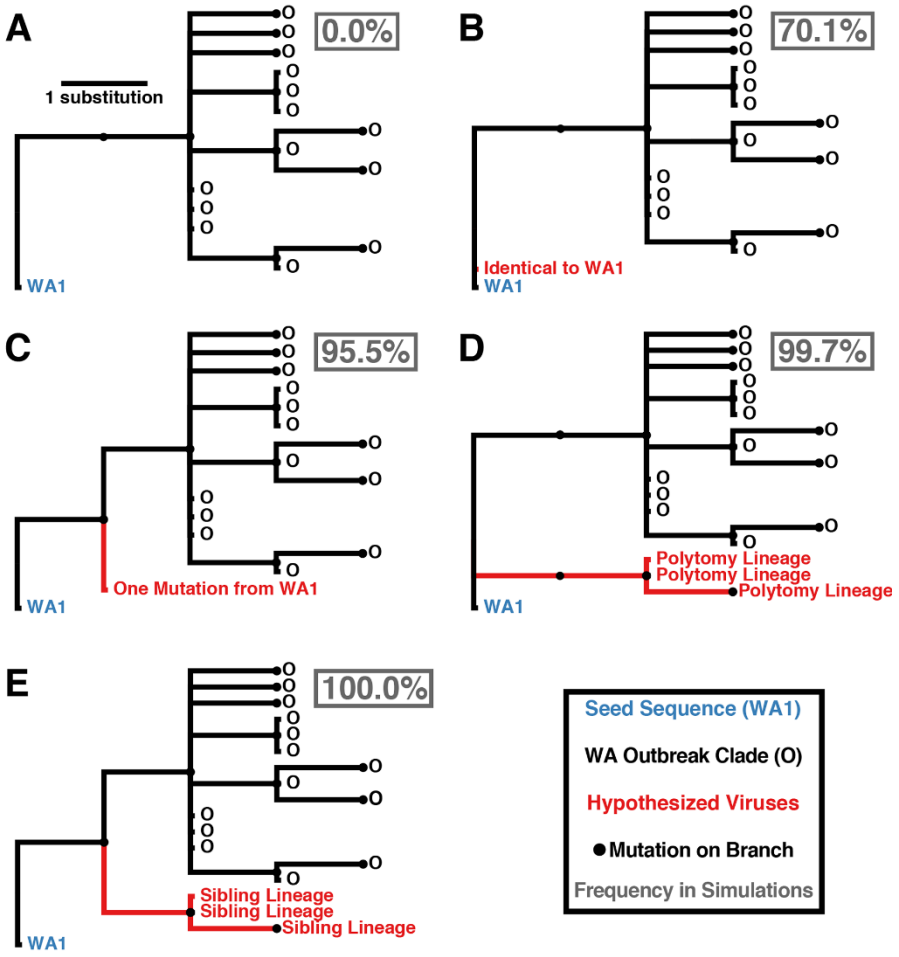

Fig. 2. Potential phylogenetic relationships between WA1 and the WA outbreak clade and their occurrence probabilities. (**A**) Observed pattern where the WA1 genome is the direct ancestor of the outbreak clade, separated by at least two mutations. (**B**) Identical sequence to WA1. (**C**) Sequence that 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.

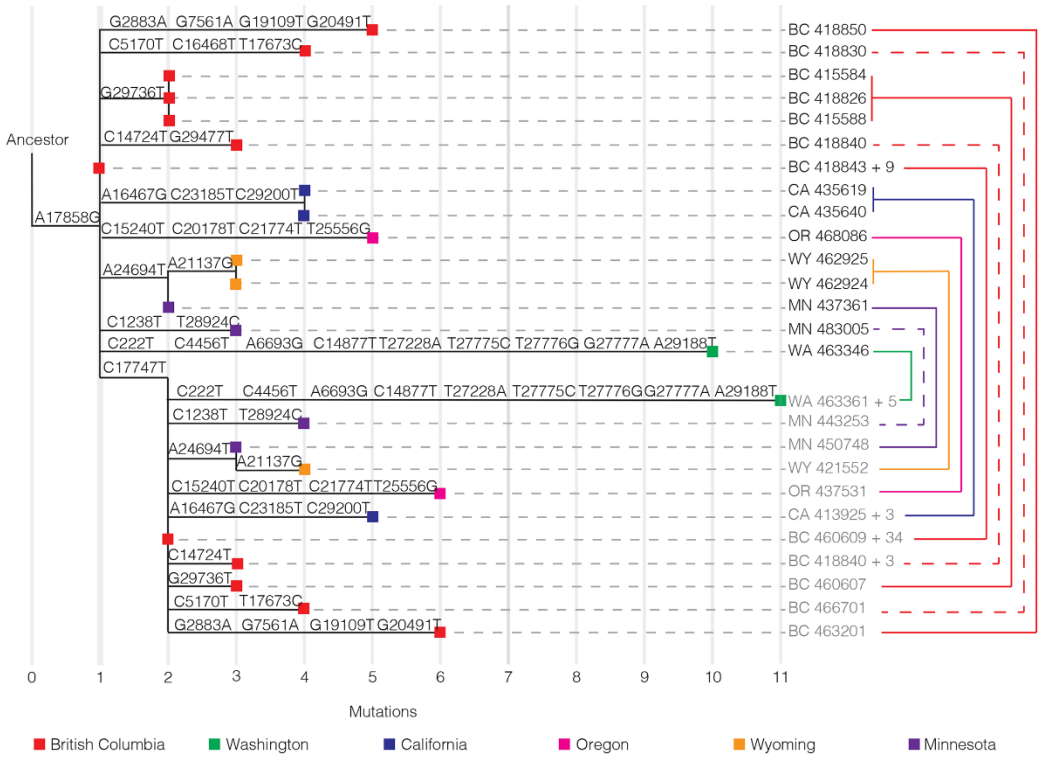

Fig. 3. Phylogeny of representative sequences showing connections between sequences that share derived mutations despite differences at the key site 17747. Derived mutations away from ancestral states are shown above each branch with position number (relative to the reference sequence hCoV-19/Wuhan/Hu-1/2019|EPI_ISL_402125). 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.

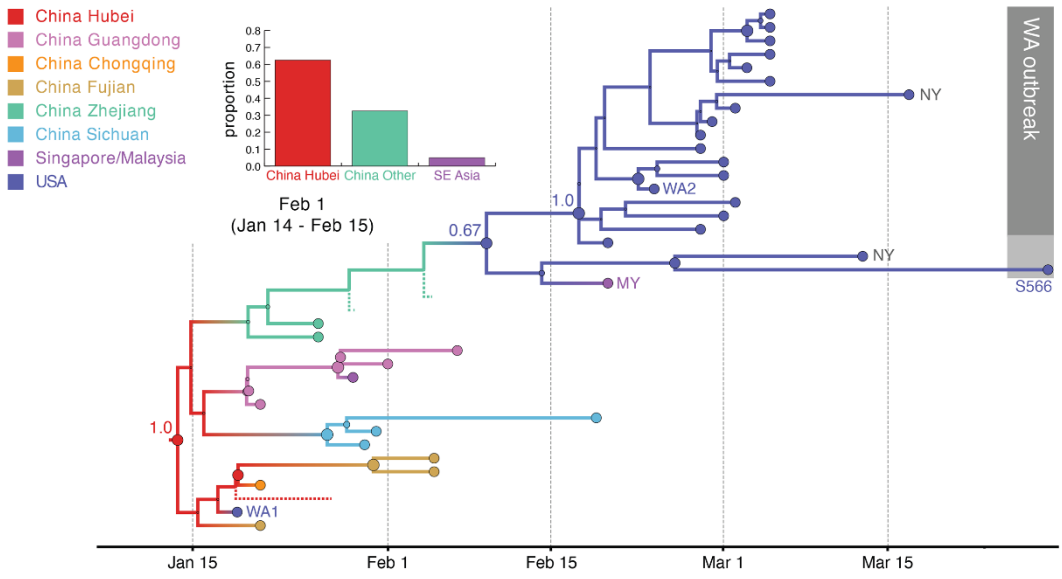

Fig. 4. Hypothesis of SARS-CoV-2 entry into Washington State. A subtree of the maximum clade credibility (MCC) tree depicting the evolutionary relationships inferred between (i) the first identified SARS-CoV-2 case in the US (WA1); (ii) the clade associated with the Washington State outbreak (including 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.

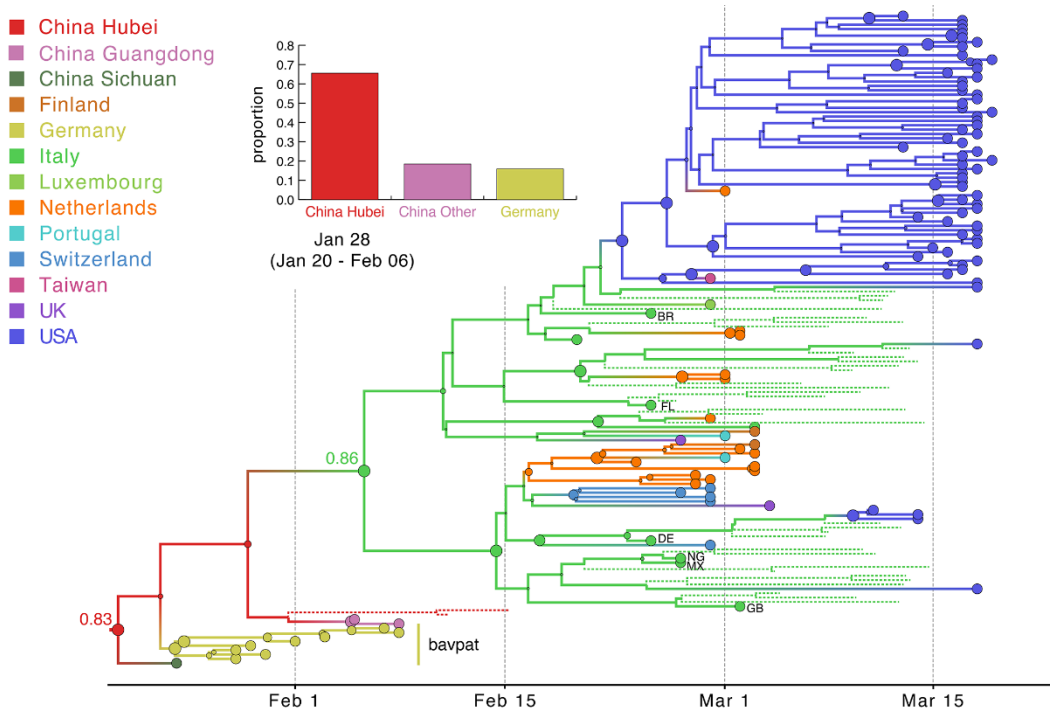

Fig. 5. MCC tree of SARS-CoV-2 entry into Europe. A subtree inferred for viruses from (i) the first outbreak in Europe (Germany, BatPat) and identical viruses from China, (ii) outbreaks in Italy and New York, and (iii) other locations in Europe. Dotted lines represent branches associated with unsampled 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.

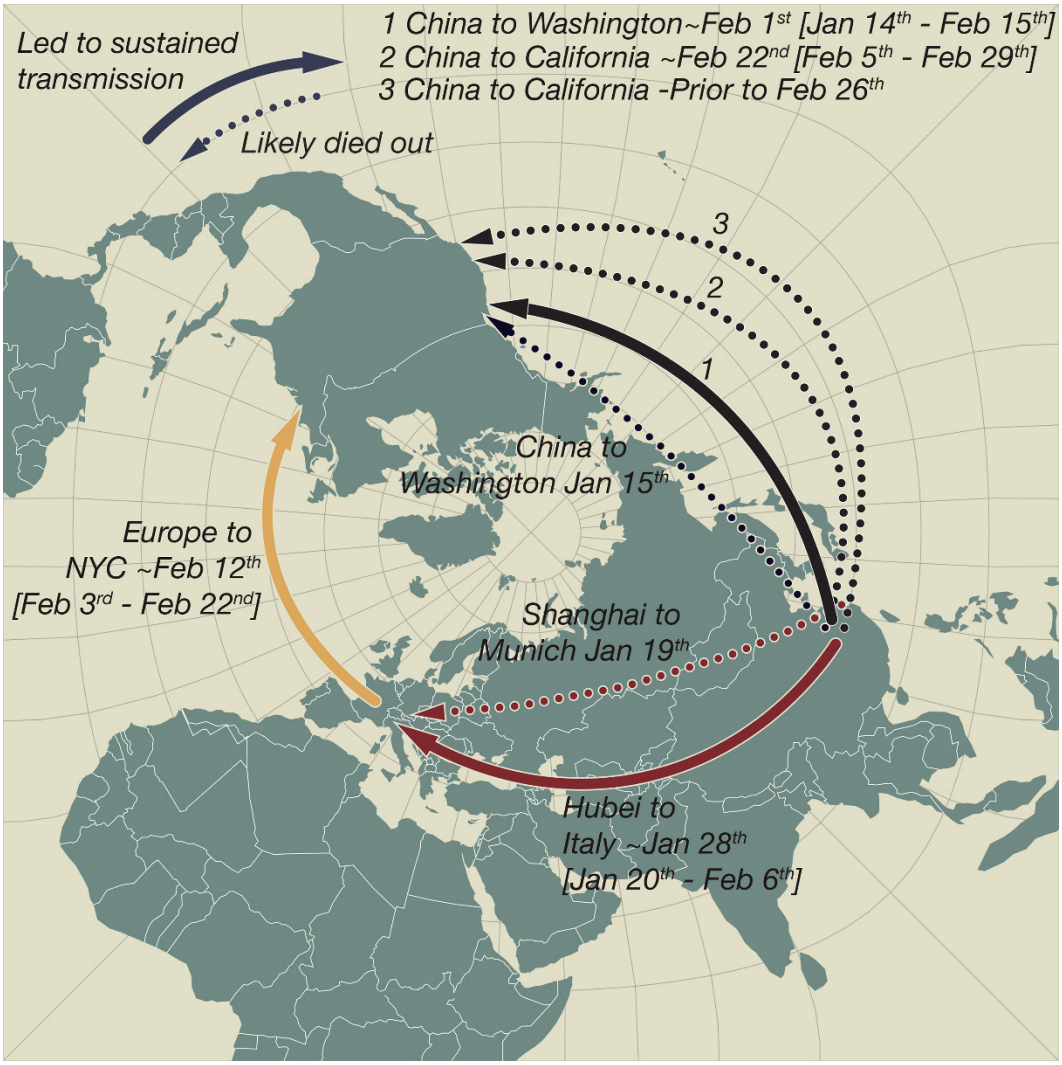

Fig. 6. SARS-CoV-2 introductions to Europe and the US. Pierce projection mapping early and apparently "dead-end" introductions of SARS-CoV-2 to Europe and the US (dashed arrows). Successful dispersals between late January and mid-February are shown with solid arrows: from Hubei Province, China to Northern 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, 2020originally published online September 10, 2020

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.