Cite as: M. Worobey *et al.*, *Science* 10.1126/science.abp8715 (2022).

The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic

Michael Worobey[1]*, Joshua I. Levy[2], Lorena Malpica Serrano[1], Alexander Crits-Christoph[3], Jonathan E. Pekar[4,5], Stephen A. Goldstein[6], Angela L. Rasmussen[7,8], Moritz U. G. Kraemer[9], Chris Newman[10], Marion P. G. Koopmans[11,12], Marc A. Suchard[13,14,15], Joel O. Wertheim[16], Philippe Lemey[17,18], David L. Robertson[19], Robert F. Garry[18,20,21], Edward C. Holmes[22], Andrew Rambaut[23], Kristian G. Andersen[2,24]*

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [2]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA. [3]W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA. [4]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093, USA. [5]Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093, USA. [6]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84112, USA. [7]Vaccine and Infectious Disease Organization, University of Saskatchewan, Saskatoon SK S7N 5E3, Canada. [8]Center for Global Health Science and Security, Georgetown University, Washington, DC 20057, USA. [9]Department of Zoology, University of Oxford, Oxford OX1 3SZ, UK. [10]Wildlife Conservation Research Unit, Department of Zoology, The Recanati-Kaplan Centre, University of Oxford, Oxford OX13 5QL, UK. [11]Pandemic and Disaster Preparedness Centre, Erasmus University Medical Center, 3015 CE Rotterdam, Netherlands. [12]Department of Viroscience, Erasmus University Medical Center, 3015 CE Rotterdam, Netherlands. [13]Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA 90095, USA. [14]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. [15]Department of Computational Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. [16]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA. [17]Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, KU Leuven, 3000 Leuven, Belgium. [18]Global Virus Network (GVN), Baltimore, MD 21201, USA. [19]MRC-University of Glasgow Center for Virus Research, Glasgow G61 1QH, UK. [20]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA 70112, USA. [21]Zalgen Labs, Frederick, MD 21703, USA. [22]Sydney Institute for Infectious Diseases, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia. [23]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK. [24]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

*Corresponding author. Email: worobey@arizona.edu (MW); andersen@scripps.edu (KGA)

Understanding how severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in 2019 is critical to preventing zoonotic outbreaks before they become the next pandemic. The Huanan Seafood Wholesale Market in Wuhan, China, was identified as a likely source of cases in early reports but later this conclusion became controversial. We show the earliest known COVID-19 cases from December 2019, including those without reported direct links, were geographically centered on this market. We report that live SARS-CoV-2 susceptible mammals were sold at the market in late 2019 and, within the market, SARS-CoV-2-positive environmental samples were spatially associated with vendors selling live mammals. While there is insufficient evidence to define upstream events, and exact circumstances remain obscure, our analyses indicate that the emergence of SARS-CoV-2 occurred via the live wildlife trade in China, and show that the Huanan market was the epicenter of the COVID-19 pandemic.

On 31 December 2019, the Chinese government notified the World Health Organization (WHO) of an outbreak of severe pneumonia of unknown etiology in Wuhan, Hubei province (*1*–*4*), a city of approximately 11 million people. Of the initial 41 people hospitalized with unknown pneumonia by 2 January 2020, 27 (66%) had direct exposure to the Huanan Wholesale Seafood Market (hereafter, "Huanan market") (*2*, *5*, *6*). These first cases were confirmed to be infected with a novel coronavirus, subsequently named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and were suffering from a disease later named coronavirus disease 2019 (COVID-19). The initial diagnoses of COVID-19 were made in several hospitals independently between 18 and 29 December 2019 (*5*). These early reports were free from ascertainment bias as they were based on signs and symptoms before the Huanan market was identified as a shared risk factor (*5*). A subsequent systematic review of all cases notified to China's National Notifiable Disease Reporting System by hospitals in Wuhan as part of the joint WHO-Chinese "WHO-convened global study of origins of SARS-CoV-2: China Part" (hereafter, "WHO mission report") (*7*) showed that 55 of 168 of the earliest known COVID-19 cases were associated with this market. However, the observation that the preponderance of early cases were linked to the Huanan market does not establish that the pandemic originated there.

Sustained live mammal sales during 2019 occurred at the Huanan and three other markets in Wuhan, including wild and farmed wild-life (*8*). Several of these species are known to be experimentally susceptible to SARS-related coronaviruses (SARSr-CoVs), such as SARS-CoV (hereafter, "SARS-CoV-1") and SARS-CoV-2 (*9*–*11*). During the early stages of the COVID-19 pandemic, animals sold at the Huanan market

were hypothesized to be the source of the unexplained pneumonia cases (*12–19*) (data S1), consistent with the emergence of SARS-CoV-1 from 2002-2004 (*20*), as well as other viral zoonoses (*21–23*). This led to the decision to close and sanitize the Huanan market on 1 January 2020, with environmental samples also being collected from vendors' stalls (*7*, *12*, *24*) (data S1).

Determining the epicenter of the COVID-19 pandemic at a neighborhood- rather than city-level could help resolve if SARS-CoV-2 had a zoonotic origin, similar to SARS-CoV-1 (*20*). In this study, we obtained data from a range of sources to test the hypothesis that the COVID-19 pandemic began at the Huanan market. Despite limited testing of live wildlife sold at the market, collectively, our results provide evidence that the Huanan market was the early epicenter of the COVID-19 pandemic and suggest that SARS-CoV-2 likely emerged from the live wildlife trade in China. However, events upstream of the market, as well as exact circumstances at the market, remain obscure, highlighting the need for further studies to understand and lower the risk of future pandemics.

Results

Early cases lived near to and centered on the Huanan market

The 2021 WHO mission report identified 174 COVID-19 cases in Hubei province in December 2019 after careful examination of reported case histories (*7*). Although geographical coordinates of the residential locations of the 164 cases who lived within Wuhan were unavailable, we were able to reliably extract the latitude and longitude coordinates of 155 cases from maps in the report (figs. S1 to S8).

While early COVID-19 cases occurred across Wuhan, the majority clustered in central Wuhan near the west bank of the Yangtze River, with a high density of cases near to, and surrounding, the Huanan market (Fig. 1A). We used a kernel density estimate (KDE) to reconstruct an underlying probability density function from which the home locations for each case were drawn (*25*). Using all 155 December 2019 cases, the location of the Huanan market lies within the highest density contour that contains 1% of the probability mass (Fig. 1B). For a KDE estimated using the 120 cases with no known linkage to the market, the market remains within the highest density 1% contour (Fig. 1C). The clustering of COVID-19 cases in December around the Huanan market (Fig. 1, B and C, insets) contrasts with the pattern of widely dispersed cases across Wuhan by early January through mid-February 2020 (Fig. 1, D and E), which we mapped using location data from individuals using a COVID-19 assistance app on Sina Weibo (*26*). Weibo-based data analyses show that, unlike early COVID-19 cases, by January and February many of the sick who sought help resided in highly populated areas of the city, and particularly in areas with a high density of older people (Fig. 1E and figs. S9 and S10).

We also investigated whether the December COVID-19 cases were closer to the market than expected based on an empirical null distribution of Wuhan's population density (data from worldpop.org (*27*, *28*)), with its median distance to the Huanan market of 16.11km (*25*). To account for older individuals being more likely to be hospitalized and sick with COVID-19 (*29*), we age-matched the population data to the December 2019 COVID-19 case data. We considered three categories of cases, and they were all significantly closer to the Huanan market than expected: (*i*) all cases (median 4.28km: $p<0.001$), (*ii*) cases linked directly to the Huanan market (median 5.74km; $p<0.001$), and (*iii*) cases with no evidence of a direct link to the Huanan market (median 4.00km: $p<0.001$) (Fig. 2A). The cases with no known link to the market on average resided closer to the market than the cases with links to the market ($p=0.029$). Furthermore, the distances between the center-points (Fig. 2B) and the Huanan market were shorter than expected for all categories of December cases compared with the empirical null distribution of Wuhan's population density (Fig. 2A). For all the December cases the center-point was located 1.02km away ($p=0.007$); the center-point for cases with market links was 2.28km away ($p=0.034$), and the center-point for the cases with no reported link to the market was 0.91km away ($p=0.006$). In comparison, the center-point of age-matched samples drawn from the empirical null distribution was 4.65km away from the market (Fig. 2A).

We tested the robustness of our results to the possibility of ascertainment bias (*25*). For all mapped cases ($n=155$), under the 'center-point distance to the Huanan market' test, the 38 cases residing closest to the market (within a radius of 1.6km) could be removed from the data set before losing significance at the $\alpha=0.05$ level (fig. S12). For the 'median distance to Huanan market' test, we could remove 98 (63%) ($r=5.8$km). For cases not directly linked to the Huanan market (n=120), we could remove 36 (30%) ($r=1.5$km) and 81 (68%) ($r=4.3$km) for the two tests, respectively, before losing significance at the $\alpha=0.05$ level (fig. S12).

We performed a spatial relative risk analysis (*25*) to compare December 2019 COVID-19 cases with January-February 2020 cases, reported via Weibo (Fig. 2C). The Huanan market is located within a well-defined area with high case density that would be expected to be observed in fewer than one in 100,000 samplings of the Weibo data empirical distribution (relative risk analysis in Fig. 2C, control distribution in Fig. 1D). No other regions in Wuhan showed a comparable case density.

Both early lineages of SARS-CoV-2 were geographically associated with the market

Two lineages of SARS-CoV-2 designated A and B (*30*) have co-circulated globally since early in the COVID-19 pandemic (*31*). Until a report in a recent preprint (*24*), only lineage B

sequences had been sampled at the Huanan market. The eleven lineage B cases from December 2019, for which we have location information, resided closer than expected to the Huanan market compared to the age-matched Wuhan population distribution (median 8.30km; p=0.017) (*25*). The center-point of the eleven lineage B cases was 1.95km from the Huanan market, also closer than expected (p=0.026). The two lineage A cases for which we have location information involved the two earliest lineage A genomes known to date. Neither case reported any contact to the Huanan market (*7*). The first case was detected before any knowledge of a possible association of unexplained pneumonia in Wuhan with the Huanan market (*5*) and therefore could not have been a product of ascertainment bias in favor of cases residing near the market. The second had stayed in a hotel near the market (*32*) for the five days preceding symptom onset (*25*). Relative to the age-matched Wuhan population distribution, the first individual resided closer to the Huanan market (2.31km) than expected (p=0.034). While the exact location of the hotel near the market was not reported (*32*), there are at least 20 hotels within 500 m (table S1). Under the conservative assumption that the hotel could have been located as far as 2.31km from the Huanan market (as was the residence of the other lineage A case), and assuming this location is comparable to a residential location given the timing of the stay prior to symptom onset (*25*), it would be unlikely to observe both the earliest lineage A cases this near to the Huanan market (p=0.001 or less). That both identified lineage A cases had a geographical connection to the market, in combination with the detection of lineage A within the market (*24*), support the likelihood that during the early epidemic lineage A was, like lineage B, disseminating outward from the Huanan market into the surrounding neighborhoods.

Our statistical results were robust to a range of factors, for example, the use of an empirical control distribution based on presumptive COVID-19 cases locations later in the Wuhan epidemic (Weibo data); laboratory-confirmed versus clinically-diagnosed cases; and uncertainty in case location or missing data (figs. S13 to S15) (*25*). For instance, we artificially introduced location uncertainty ('noise') in each case location in our data set by randomly re-sampling each point within a circle of radius 1000m centered on its original center-point; the conclusions were unaffected (fig. S13). The extraction method we employed actually introduced up to about 50m of noise in each case location estimate (fig. S7), ruling out the possibility that our overall results were affected by this source of error. The results were also robust when corrected for multiple hypothesis testing (table S4).

Wild animal trading in Wuhan markets

In addition to selling seafood, poultry, and other commodities, the Huanan market was among four markets in Wuhan reported to consistently sell a variety of live, wild-captured or farmed, mammal species in the years and months leading up to the COVID-19 pandemic (*8*). There are, however, no prior reports of which species, if any, were sold at the Huanan market in the months leading up to the pandemic. Here, we report that multiple plausible intermediate wildlife hosts of SARS-CoV-2 progenitor viruses, including red foxes (*Vulpes vulpes*), hog badgers (*Arctonyx albogularis*) and common raccoon dogs (*Nyctereutes procyonoides*), were sold live at the Huanan market up until at least November of 2019 (Table 1 and table S5). No reports are known to be available for SARS-CoV-2 test results from these mammals at the Huanan market. Despite a general slow-down in live animal sales during the winter months, we report that raccoon dogs that are sold for both meat and fur were consistently available for sale throughout the year, including at the Huanan market in November 2019 (Table 1 and table S5).

There were potentially many locations in Wuhan, a city of 11 million, that would have been equally or more likely than the Huanan market to sustain the first recognized cluster of a new respiratory pathogen had its introduction not been linked to a live animal market, including other shopping venues, hospitals, elder care facilities, workplaces, universities, and places of worship. To investigate possible sites, we compared the relative extent of intra-urban human traffic to the Huanan market versus other locations within the city of Wuhan using a location-specific data set of social media check-ins in the Sina Visitor System (*25*, *33*). We found at least 70 other markets throughout the city of Wuhan that received more social media visitors than the Huanan market (Fig. 3). To extend this analysis beyond only markets, we also used a subsequently published list of known SARS-CoV-2 superspreader locations (*34*) to identify 430 locations in Wuhan that may have been at high risk for superspreader events and which received more check-ins than the Huanan market (Fig. 3, inset). The Huanan market accounted for 0.12% (120 of 98,146) of social media check-ins to markets in the data set that received at least as many check-ins as the Huanan market. The market accounted for 0.04% (120 of 262,233) of all social media check-ins to the >400 sites in Wuhan identified as especially likely to be potential superspreader locations and which received at least as many social media visits as the Huanan market. Considering the number of check-ins to all four markets selling live, wild animals in Wuhan (combined), they accounted for 0.21% (206 of 98,146) of market visits and 0.079% (206 of 262,233) of visits to the 430 potential superspreader sites, where a new respiratory disease might first be noticed in a large city.

A data set from the Chinese Center for Disease Prevention and Control (CCDC) report dated 22 January 2020 (data S1) (*12*, *13*, *15*, *16*) was made publicly available in June 2020 (*24*, *35*). 585 environmental samples were initially taken from various surfaces in the Huanan market on 1 and 12 January 2020 by the

CCDC (tables S6 and S7 and data S1) (*12*, *13*, *15*, *16*, *24*, *35*), with further samples taken through the market during January and February (*24*). We extended the analysis in the WHO mission report (*7*) by integrating public online maps and photographic evidence, data from public business registries (table S8 and data S2), information about which live mammal species were sold at the Huanan market in late 2019 (Table 1 and table S5), and the CCDC report (data S1). We reconstructed the floor plan of the market and integrated information from business registries of vendors at the market (fig. S16 and table S8), as well as an official report (*36*) recording fines to three business owners for illegal sale of live mammals (data S2) (*36*). From this, we identified an additional five stalls that were likely selling live or freshly butchered mammals or other unspecified meat products in the southwest corner of the western section of the market (Fig. 4A, figs. S16 and S17, and table S6).

Five of the SARS-CoV-2-positive environmental samples were taken from a single stall selling live mammals in late 2019 (table S6). Further, the objects sampled showed an association with animal sales, including a metal cage, two carts (of the kind frequently used to transport mobile animal cages) and a hair/feather remover (table S6). No human COVID-19 cases were reported there (*7*, *12*). The same stall was visited by one of us (ECH) in 2014, who then observed live raccoon dogs housed in a metal cage stacked on top of a cage with live birds (Fig. 4A) (*37*). A recent report (*24*) identified that the grates outside of this stall, upon which animal cages were stacked (*37*), were positive for SARS-CoV-2.

Positive environmental samples linked both to live mammal sales and to human cases at the Huanan market

We used a spatial relative risk analysis to identify potential regions of the market with an increased density of positive environmental samples (*25*). We found evidence ($p<0.05$) of a region in the southwest area of the market where live mammals were on sale (Fig. 4B). Although environmental sampling of the market was incomplete and spatially heterogeneous (data S1 and table S6), our analysis accounts for the empirical environmental sampling distribution, which was biased toward 'stalls related to December cases' as well as 'stalls that sold livestock, poultry, farmed wildlife' (*7*) (Fig. 4, C and D). The 'distance to the nearest vendor selling live mammals' and 'distance to the nearest human case' were independently predictive of environmental sample positivity ($p=0.004$ and 0.014, respectively for N=6; table S9). To further investigate the robustness of these findings to possible sampling biases, we considered three scenarios: (*i*) oversampling of live mammal and unknown meat stalls, (*ii*) overcounting of positive samples, and (*iii*) exclusion of the seafood stand near the wildlife area of the market (with five positive samples) from our analysis (table S10). In each case, the distance to live mammal vendors remained predictive of environmental sample positivity, and the region of increased positive sample density in the southwest corner of the western section of the market remained consistent (fig. S18).

Finally, to analyze the spatial patterning of human cases within the Huanan market, we plotted cases as a function of symptom onset from the WHO mission report (*7*) (Fig. 5A and table S11) (*25*). All eight COVID-19 cases detected prior to 20 December were from the western side of the market, where mammal species were also sold (Fig. 5, B and C). Unlike SARS-CoV-2 positive environmental samples (Fig. 4, A and C), we found that COVID-19 cases were more diffuse throughout the building (Fig. 5).

Study limitations

There are several limitations to our study. We have been able to recover location data for most of the December-onset COVID-19 cases identified by the WHO mission (*7*) and have been able to do so with sufficient precision to support our conclusions. However, we do not have access to the precise latitude and longitude coordinates of all these cases. Should such data exist, they may be accompanied by additional metadata, some of which we have reconstructed, but some of which, including the date of onset of each case, would be valuable for ongoing studies. We also lack direct evidence of an intermediate animal infected with a SARS-CoV-2 progenitor virus either at the Huanan market or at a location connected to its supply chain, like a farm. Additionally, no line list of early COVID-19 cases is available and we do not have complete details of environmental sampling, though compared to many other outbreaks, we have more comprehensive information on early cases, hospitalizations and environmental sampling (*7*).

Discussion

Several lines of evidence support the hypothesis that the Huanan market was the epicenter of the COVID-19 pandemic and that SARS-CoV-2 emerged from activities associated with live wildlife trade. Spatial analyses within the market show that SARS-CoV-2-positive environmental samples, including cages, carts, and freezers, were associated with activities concentrated in the southwest corner of the market. This is the same section where vendors were selling live mammals, including raccoon dogs, hog badgers, and red foxes, immediately prior to the COVID-19 pandemic. Multiple positive samples were taken from one stall known to have sold live mammals, and the water drain proximal to this stall, as well as other sewerages and a nearby wildlife stall on the southwest side of the market, tested positive for SARS-CoV-2 (*24*). These findings suggest that infected animals were present at the Huanan market at the beginning of the COVID-19

pandemic; however, we do not have access to any live animal samples from relevant species. Additional information, including sequencing data and detailed sampling strategy, would be invaluable to test this hypothesis comprehensively.

In a related study, we infer separate introductions of SARS-CoV-2 lineages A and B into humans from likely infected animals at the Huanan market (*38*). We estimate the first COVID-19 case to have occurred in November 2019, with few human cases and hospitalizations occurring through mid-December (*38*). A recent preprint (*24*) confirms the authenticity of the CCDC report (data S1) and records additional positive environmental samples in the southwestern area of the market selling live animals. This report also documents the early presence of the A lineage of SARS-CoV-2 in a Huanan market environmental sample. This, along with the lineage A cases we report in close geographical proximity to the market in December, challenges the suggestion that the market was simply a superspreading event, which would be lineage-specific. Rather, it adds to the evidence presented here that lineage A, like lineage B, may have originated at the Huanan market then spread from this epicenter into the neighborhoods surrounding the market and then beyond.

Several observations suggest that the geographic association of early COVID-19 cases with the Huanan market is unlikely to have been the result of ascertainment bias (supplementary text and tables S2 and S3) (*39*). These include: (*i*) few, if any, cases among Huanan market-unlinked individuals are likely to have been detected by active searching in the neighborhoods around the market – only in hospitals – since all cases analyzed here were hospitalized (*7*), (*ii*) public health officials simultaneously became aware of Huanan-linked cases near and far from the Huanan market, not just ones near it (fig. S11) (*5*), (*iii*) Huanan-unlinked cases would not be expected to live significantly closer to the market than linked cases if they had been ascertained as contacts traced from those market-linked cases, and (*iv*) seroprevalence in Wuhan was highest in the districts around the market (*40*, *41*). It is also noteworthy that the December 2019 COVID-19 cases we consider here were identified based on reviews of clinical signs and symptoms, not epidemiological factors such as where they resided or links to the Huanan market (*7*) and that excess deaths from pneumonia rose first in the districts surrounding the market (*42*). Moreover, the spatial relationship with the Huanan market remains after removing the two-thirds of the unlinked cases residing nearest the market.

One of the key findings of our study is that 'unlinked' early COVID-19 patients, those who neither worked at the market or knew someone who did, nor had recently visited the market, resided significantly closer to the market than patients with a direct link to the market. The observation that a substantial proportion of early cases had no known epidemiological link had previously been used as an argument against a Huanan market epicenter of the pandemic. However, this group of cases resided significantly closer to the market than those who worked there, indicating that they had been exposed to the virus at, or near, the Huanan market. For market workers, the exposure risk was their place of work not their residential locations, which were significantly further afield than those cases not formally linked to the market.

Our spatial analyses show how patterns of COVID-19 cases shifted between late 2019, when the outbreak began (*43*), and early 2020, as the epidemic spread widely across Wuhan. COVID-19 cases in December 2019 were associated with the Huanan market in a manner unrelated to Wuhan population density or demographic patterns, unlike the wide spatial distribution of cases observed during later stages of the epidemic in January and February. This observation fits with the evidence from other sources that SARS-CoV-2 was not widespread in Wuhan at the end of 2019. For example, no SARS-CoV-2-positive sera or influenza-like illness (ILI) reports were recorded among more 40,000 blood donor samples collected up to December 2019 (*44*, *45*), and none of thousands of samples taken from ILI patients at Wuhan hospitals in October-December 2019 tested for SARS-CoV-2 RNA was positive (*7*).

The sustained presence of a potential source of virus transmission into the human population in late 2019, plausibly from infected live mammals sold at the Huanan market, offers an explanation of our findings and the origins of SARS-CoV-2. The pattern of COVID-19 cases reported for the Huanan market, with the earliest cases in the same part of the market as the wildlife sales and evidence of at least two introductions (*38*), resembles the multiple cross-species transmissions of SARS-CoV-2 subsequently observed during the pandemic from animals to humans on mink farms (*46*), and from infected hamsters to humans in the pet trade (*47*). There was an extensive network of wildlife farms in western Hubei province, including hundreds of thousands of raccoon dogs on farms in Enshi prefecture, which supplied the Huanan market (*48*). This region of Hubei contains extensive cave complexes housing *Rhinolophus* bats, which carry SARSr-CoVs (*49*). SARS-CoV-1 was recovered from farmed masked palm civets from Hubei in 2003 and 2004 (*20*). The animals on these farms (nearly 1 million) were rapidly released, sold, or killed in early 2020 (*48*), apparently without testing for SARS-CoV-2 (*7*). Live animals sold at the market (Table 1) were apparently not sampled either. By contrast, during the SARS-CoV-1 outbreaks farms and markets remained open for over a year after the first human cases occurred, allowing sampling of viruses from infected animals (*20*).

The live animal trade and live animal markets are a common theme in virus spillover events (*21–23*, *50*), with markets such as the Huanan market selling live mammals being in the highest risk category (*51*). The events leading up to the

COVID-19 pandemic mirror the SARS-CoV-1 outbreaks from 2002-2004, which were traced to infected animals in Guangdong, Jiangxi, Henan, Hunan, and Hubei provinces in China (*20*). Maximum effort must now be applied to elucidate the upstream events that might have brought SARS-CoV-2 into the Huanan market, culminating in the COVID-19 pandemic. To reduce the risk of future pandemics we must understand, and then limit, the routes and opportunities for virus spillover.

Methods summary

Ethics statement

This research was reviewed by the Human Subject Protection Program at the University of Arizona and the Institutional Review Board at The Scripps Research Institute and determined to be exempt from IRB approval because it constitutes secondary research for which consent is not required.

Data sources

COVID-19 case data from December 2019 was obtained from the WHO mission report (*7*) and our previous analyses (*5*). Location information was extracted and sensitivity analyses performed to confirm accuracy and assess potential ascertainment bias. Geotagged January/February 2020 data from Weibo COVID-19 help seekers was obtained from the authors (*26*). Population density data was obtained from worldpop.org (*27*). Sequencing- or qPCR-based environmental sample SARS-CoV-2 positivity from the Huanan market was obtained from a January 2020 China CDC report (data S1) (*24*).

Wildlife trading at the Huanan market

Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic was previously reported (*8*) and in this study we report details about animals for sale at the Huanan market up until November 2019.

Spatial analyses of COVID-19 cases

Haversine distances to the Huanan market were calculated for each of the geolocated December 2019 cases. Centerpoints and median distances from cases to the Huanan market were calculated separately for (1) all 155 cases, (2) the 35 cases epidemiologically linked to the Huanan market, (3) for the 120 cases not epidemiologically linked to the market, (4) the eleven lineage B cases, and (5) the earliest lineage A case. These distances were also calculated for the 737 Weibo help seekers from 8 January to 10 February 2020 (*26*). Empirical null distributions were generated from the population density data and the Weibo data. The population density null distributions were age-matched to the December 2019 cases.. Kernel density estimates were also generated for the market-linked cases, unlinked-cases and all cases, to infer a probability density function from which the cases could have been drawn. Highest-density contours representing specific probability masses (0.5, 0.25, 0.1, 0.05, and 0.01) were inferred and the location of the market compared to these.

Mobility analyses

To estimate the relative amount of intra-urban human traffic to the Huanan market compared to other locations within the city of Wuhan, we utilized a location-specific dataset of social media check-ins in the Sina Visitor System as shared by Li *et al.* 2015 (*33*). This dataset is based on 1,491,499 individual check-in events across the city of Wuhan from the years 2013-2014 (5-6 years before the start of the COVID-19 pandemic), and 770,521 visits are associated with 312,190 unique user identifiers. Location names and categories were translated using a Python API for Google Translate.

Spatial analyses of environmental samples at the Huanan market

We used the official maps from the China CDC (*12*) (data S1) and WHO map (*7*), as well as satellite photographs (Google Maps, Google Earth, Baidu Maps), aerial photographs, and images of the market in the public domain to reconstruct the floorplan of the market. Market stalls were assigned by categories of the types of goods sold using official reports and data from the TianYanCha.com business directory (table S8). Final maps of the Huanan market were converted into geojson format for spatial analyses. Significance testing of live animal vendors and/or human SARS-CoV-2 cases on the number of positive environmental samples was performed using a binomial GLM. Distances between businesses were defined as the distance between their respective centerpoints and spatial relative risk analysis was performed using the 'sparr' package in R, using linear boundary kernels for edge correction (*52*), with bandwidth selection performed using least squares cross-validation.

REFERENCES AND NOTES

1. Sina Finance, "Wuhan pneumonia of unknown cause cases isolated, test results to be announced ASAP" (Sina Finance, 2019); https://finance.sina.cn/2019-12-31/detail-iihnzahk1074832.d.html?from=wap.
2. Wuhan Municipal Health Commission, "Wuhan Municipal Health Commission's briefing on the current situation of pneumonia in our city" (Wuham Municipal Health Commission, 2019); https://web.archive.org/web/20200131202951/http://wjw.wuhan.gov.cn/front/web/showDetail/2019123108989.
3. World Health Organization, "COVID-19 – China" (WHO, 2020); https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229.
4. The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) – China, 2020. *China CDC Wkly* **2**, 113–122 (2020). doi:10.46234/ccdcw2020.032 Medline
5. M. Worobey, Dissecting the early COVID-19 cases in Wuhan. *Science* **374**, 1202–1204 (2021). doi:10.1126/science.abm4454 Medline
6. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L.

Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020). doi:10.1016/S0140-6736(20)30183-5 Medline
7. World Health Organization, "WHO-convened global study of origins of SARS-CoV-2: China Part" (WHO, 2021); https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.
8. X. Xiao, C. Newman, C. D. Buesching, D. W. Macdonald, Z.-M. Zhou, Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. *Sci. Rep.* **11**, 11898 (2021). doi:10.1038/s41598-021-91470-2 Medline
9. C. M. Freuling, A. Breithaupt, T. Müller, J. Sehl, A. Balkema-Buschmann, M. Rissmann, A. Klein, C. Wylezich, D. Höper, K. Wernike, A. Aebischer, D. Hoffmann, V. Friedrichs, A. Dorhoi, M. H. Groschup, M. Beer, T. C. Mettenleiter, Susceptibility of raccoon dogs for experimental SARS-CoV-2 infection. *Emerg. Infect. Dis.* **26**, 2982–2985 (2020). doi:10.3201/eid2612.203733 Medline
10. W. K. Jo, E. F. de Oliveira-Filho, A. Rasche, A. D. Greenwood, K. Osterrieder, J. F. Drexler, Potential zoonotic sources of SARS-CoV-2 infections. *Transbound. Emerg. Dis.* **68**, 1824–1834 (2021). doi:10.1111/tbed.13872 Medline
11. I. R. Fischhoff, A. A. Castellanos, J. P. G. L. M. Rodrigues, A. Varsani, B. A. Han, Predicting the zoonotic capacity of mammals to transmit SARS-CoV-2. *Proc. Biol. Sci.* **288**, 20211651 (2021). doi:10.1098/rspb.2021.1651 Medline
12. W. Guizhen, "Chinese CDC disease control report" (see data S1).
13. Xinhua News, "Good news! Phased progress made in tracing the origin of the coronavirus" (Xinhua News, 2020); http://www.xinhuanet.com/politics/2020-01/26/c_1125503792.htm.
14. Beijing News, "Huanan Seafood Market in the pneumonia of unexplained incident" (Beijing News, 2020); http://www.bjnews.com.cn/feature/2020/01/02/669054.html.
15. Chinese Center for Disease Control and Prevention, "Chinese Center for Disease Control and Prevention detects large quantity of novel coronavirus in Wuhan Huanan Seafood Market" (Chinese CDC, 2020); https://www.chinacdc.cn/yw_9324/202001/t20200127_211469.html.
16. Yicai Global, "China detects large quantity of novel coronavirus at Wuhan Seafood Market" (Yicai Global, 2020); https://www.yicaiglobal.com/opinion/yicai.global/china-detects-large-quantity-of-novel-coronavirus-at-wuhan-seafood-market.
17. Chinese Center for Disease Control and Prevention, "China CDC calls on the public to protect themselves" (Chinese CDC, 2020); https://www.chinacdc.cn/yw_9324/202001/t20200128_211498.html.
18. Chinese Center for Disease Control and Prevention, "On the front line, disease control warriors race against the new coronavirus" (Chinese CDC, 2020); https://www.chinacdc.cn/yw_9324/202002/t20200201_212137.html.
19. Xinhua News, "China detects large quantity of novel coronavirus at Wuhan seafood market" (Xinhua News, 2020); https://web.archive.org/web/20200126230041/http://www.xinhuanet.com/english/2020-01/27/c_138735677.htm.
20. Z. Shi, Z. Hu, A review of studies on animal reservoirs of the SARS coronavirus. *Virus Res.* **133**, 74–87 (2008). doi:10.1016/j.virusres.2007.03.012 Medline
21. W. B. Karesh, R. A. Cook, E. L. Bennett, J. Newcomb, Wildlife trade and global disease emergence. *Emerg. Infect. Dis.* **11**, 1000–1002 (2005). doi:10.3201/eid1107.050194 Medline
22. N. D. Wolfe, P. Daszak, A. M. Kilpatrick, D. S. Burke, Bushmeat hunting, deforestation, and prediction of zoonoses emergence. *Emerg. Infect. Dis.* **11**, 1822–1827 (2005). doi:10.3201/eid1112.040789 Medline
23. C. K. Johnson, P. L. Hitchens, P. S. Pandit, J. Rushmore, T. S. Evans, C. C. W. Young, M. M. Doyle, Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proc. Biol. Sci.* **287**, 20192736 (2020). doi:10.1098/rspb.2019.2736 Medline
24. G. Gao, W. Liu, P. Liu, W. Lei, Z. Jia, X. He, L.-L. Liu, W. Shi, Y. Tan, S. Zou, X. Zhao, G. Wong, J. Wang, F. Wang, G. Wang, K. Qin, R. Gao, J. Zhang, M. Li, W. Xiao, Y. Guo, Z. Xu, Y. Zhao, J. Song, J. Zhang, W. Zhen, W. Zhou, B. Ye, J. Song, M. Yang, W. Zhou, Y. Bi, K. Cai, D. Wang, W. Tan, J. Han, W. Xu, G. Wu, "Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market" [Preprint] (Research Square, 2022); https://www.researchsquare.com/article/rs-1370392/v1.
25. Material and methods are available as supplementary materials.
26. Z. Peng, R. Wang, L. Liu, H. Wu, Exploring urban spatial features of COVID-19 transmission in Wuhan based on social media data. *ISPRS Int. J. Geoinf.* **9**, 402 (2020). doi:10.3390/ijgi9060402
27. WorldPop, "WorldPop: Open spatial demographic data and research" (2020); http://worldpop.org.
28. A. J. Tatem, WorldPop, open data for spatial demography. *Sci. Data* **4**, 170004 (2017). doi:10.1038/sdata.2017.4 Medline
29. M. O'Driscoll, G. Ribeiro Dos Santos, L. Wang, D. A. T. Cummings, A. S. Azman, J. Paireau, A. Fontanet, S. Cauchemez, H. Salje, Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature* **590**, 140–145 (2021). doi:10.1038/s41586-020-2918-0 Medline
30. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020). doi:10.1038/s41564-020-0770-5 Medline
31. outbreak.info, "SARS-CoV-2 (hCoV-19) mutation reports: Lineage/mutation tracker" (outbreak.info, 2022); https://outbreak.info/situation-reports.
32. R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020). doi:10.1016/S0140-6736(20)30251-8 Medline
33. L. Li, L. Yang, H. Zhu, R. Dai, Explorative analysis of Wuhan Intra-urban human mobility using social media check-in data. *PLOS ONE* **10**, e0135286 (2015). doi:10.1371/journal.pone.0135286 Medline
34. D. Majra, J. Benson, J. Pitts, J. Stebbing, SARS-CoV-2 (COVID-19) superspreader events. *J. Infect.* **82**, 36–40 (2021). doi:10.1016/j.jinf.2020.11.021 Medline
35. Epoch Times, "[Exclusive] The secret of Wuhan Huanan Seafood Market testing" (Epoch Times, 2020); https://www.epochtimes.com/gb/20/5/31/n12150755.htm.
36. Wuhan Municipal Bureau of Landscape Architecture and Forestry, "Administrative penalties in 2019" (Wuhan Municipal Bureau of Landscape Architecture and Forestry, 2019); https://web.archive.org/web/20211117124950/http://ylj.wuhan.gov.cn/zwgk/zwxxgkzl_12298/cfqz/xzcf/202011/t20201110_1499879.shtml.
37. Y.-Z. Zhang, E. C. Holmes, A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* **181**, 223–227 (2020). doi:10.1016/j.cell.2020.03.035 Medline
38. J. E. Pekar, A. Magee, E. Parker, N. Moshiri, K. Izhikevich, J. L. Havens, K. Gangavarapu, L. M. Malpica Serrano, A. Crits-Christoph, N. L. Matteson, M. Zeller, J. I. Levy, J. C. Wang, S. Hughes, J. Lee, H. Park, M.-S. Park, K. Ching Zi Yan, R. T. Pin Lin, M. N. Mat Isa, Y. M. Noor, T. I. Vasylyeva, R. F. Garry, E. C. Holmes, A. Rambaut, M. A. Suchard, K. G. Andersen, M. Worobey, J. O. Wertheim, "SARS-CoV-2 emergence very likely resulted from at least two zoonotic events" (Zenodo, 2022); https://zenodo.org/record/6291628#.YtbMInbMKUk.
39. N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, J. Xia, T. Yu, X. Zhang, L. Zhang, Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **395**, 507–513 (2020). doi:10.1016/S0140-6736(20)30211-7 Medline
40. Z. Li, X. Guan, N. Mao, H. Luo, Y. Qin, N. He, Z. Zhu, J. Yu, Y. Li, J. Liu, Z. An, W. Gao, X. Wang, X. Sun, T. Song, X. Yang, M. Wu, X. Wu, W. Yao, Z. Peng, J. Sun, L. Wang, Q. Guo, N. Xiang, J. Liu, B. Zhang, X. Su, L. Rodewald, L. Li, W. Xu, H. Shen, Z. Feng, G. F. Gao, Antibody seroprevalence in the epicenter Wuhan, Hubei, and six selected provinces after containment of the first epidemic wave of COVID-19 in China. *Lancet Reg Health West Pac* **8**, 100094 (2021). doi:10.1016/j.lanwpc.2021.100094 Medline
41. Z. He, L. Ren, J. Yang, L. Guo, L. Feng, C. Ma, X. Wang, Z. Leng, X. Tong, W. Zhou, G. Wang, T. Zhang, Y. Guo, C. Wu, Q. Wang, M. Liu, C. Wang, M. Jia, X. Hu, Y. Wang, X. Zhang, R. Hu, J. Zhong, J. Yang, J. Dai, L. Chen, X. Zhou, J. Wang, W. Yang, C. Wang, Seroprevalence and humoral immune durability of anti-SARS-CoV-2 antibodies in Wuhan, China: A longitudinal, population-level, cross-sectional study. *Lancet* **397**, 1075–1084 (2021). doi:10.1016/S0140-6736(21)00238-5 Medline

42. E. C. Holmes, S. A. Goldstein, A. L. Rasmussen, D. L. Robertson, A. Crits-Christoph, J. O. Wertheim, S. J. Anthony, W. S. Barclay, M. F. Boni, P. C. Doherty, J. Farrar, J. L. Geoghegan, X. Jiang, J. L. Leibowitz, S. J. D. Neil, T. Skern, S. R. Weiss, M. Worobey, K. G. Andersen, R. F. Garry, A. Rambaut, The origins of SARS-CoV-2: A critical review. *Cell* **184**, 4848–4856 (2021). doi:10.1016/j.cell.2021.08.017 Medline
43. J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, J. O. Wertheim, Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**, 412–417 (2021). doi:10.1126/science.abf8003 Medline
44. L. Chang, W. Hou, L. Zhao, Y. Zhang, Y. Wang, L. Wu, T. Xu, L. Wang, J. Wang, J. Ma, L. Wang, J. Zhao, J. Xu, J. Dong, Y. Yan, R. Yang, Y. Li, F. Guo, W. Cheng, Y. Su, J. Zeng, W. Han, T. Cheng, J. Zhang, Q. Yuan, N. Xia, L. Wang, The prevalence of antibodies to SARS-CoV-2 among blood donors in China. *Nat. Commun.* **12**, 1383 (2021). doi:10.1038/s41467-021-21503-x Medline
45. L. Chang, L. Zhao, Y. Xiao, T. Xu, L. Chen, Y. Cai, X. Dong, C. Wang, X. Xiao, L. Ren, L. Wang, Serosurvey for SARS-CoV-2 among blood donors in Wuhan, China from September to December 2019. *Protein Cell* pwac013 (2019). 10.1093/procel/pwac013
46. L. Lu, R. S. Sikkema, F. C. Velkers, D. F. Nieuwenhuijse, E. A. J. Fischer, P. A. Meijer, N. Bouwmeester-Vincken, A. Rietveld, M. C. A. Wegdam-Blans, P. Tolsma, M. Koppelman, L. A. M. Smit, R. W. Hakze-van der Honing, W. H. M. van der Poel, A. N. van der Spek, M. A. H. Spierenburg, R. J. Molenaar, J. Rond, M. Augustijn, M. Woolhouse, J. A. Stegeman, S. Lycett, B. B. Oude Munnink, M. P. G. Koopmans, Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. *Nat. Commun.* **12**, 6802 (2021). doi:10.1038/s41467-021-27096-9 Medline
47. H.-L. Yen, T. H. C. Sit, C. J. Brackman, S. S. Y. Chuk, H. Gu, K. W. S. Tam, P. Y. T. Law, G. M. Leung, M. Peiris, L. L. M. Poon, S. M. S. Cheng, L. D. J. Chang, P. Krishnan, D. Y. M. Ng, G. Y. Z. Liu, M. M. Y. Hui, S. Y. Ho, W. Su, S. F. Sia, K.-T. Choy, S. S. Y. Cheuk, S. P. N. Lau, A. W. Y. Tang, J. C. T. Koo, L. Yung, Transmission of SARS-CoV-2 (Variant Delta) from pet hamsters to humans and onward human propagation of the adapted strain: A case study. *Lancet* **399**, 1070–1078 (2022). doi:10.1016/S0140-6736(22)00326-9 Medline
48. M. Standaert, E. Dou, "In search for coronavirus origins, Hubei caves and wildlife farms draw new scrutiny," *The Washington Post*, 11 October 2021; https://www.washingtonpost.com/world/asia_pacific/china-covid-bats-caves-hubei/2021/10/10/082eb8b6-1c32-11ec-bea8-308ea134594f_story.html.
49. X.-D. Lin, W. Wang, Z.-Y. Hao, Z.-X. Wang, W.-P. Guo, X.-Q. Guan, M.-R. Wang, H.-W. Wang, R.-H. Zhou, M.-H. Li, G.-P. Tang, J. Wu, E. C. Holmes, Y.-Z. Zhang, Extensive diversity of coronaviruses in bats from China. *Virology* **507**, 1–10 (2017). doi:10.1016/j.virol.2017.03.019 Medline
50. Q. Li, L. Zhou, M. Zhou, Z. Chen, F. Li, H. Wu, N. Xiang, E. Chen, F. Tang, D. Wang, L. Meng, Z. Hong, W. Tu, Y. Cao, L. Li, F. Ding, B. Liu, M. Wang, R. Xie, R. Gao, X. Li, T. Bai, S. Zou, J. He, J. Hu, Y. Xu, C. Chai, S. Wang, Y. Gao, L. Jin, Y. Zhang, H. Luo, H. Yu, J. He, Q. Li, X. Wang, L. Gao, X. Pang, G. Liu, Y. Yan, H. Yuan, Y. Shu, W. Yang, Y. Wang, F. Wu, T. M. Uyeki, Z. Feng, Epidemiology of human infections with avian influenza A(H7N9) virus in China. *N. Engl. J. Med.* **370**, 520–532 (2014). doi:10.1056/NEJMoa1304617 Medline
51. B. Lin, M. L. Dietrich, R. A. Senior, D. S. Wilcove, A better classification of wet markets is key to safeguarding human health and biodiversity. *Lancet Planet. Health* **5**, e386–e394 (2021). doi:10.1016/S2542-5196(21)00112-1 Medline
52. T. M. Davies, J. C. Marshall, M. L. Hazelton, Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk. *Stat. Med.* **37**, 1191–1221 (2018). doi:10.1002/sim.7577 Medline
53. Data and code for: M. Worobey, J. I. Levy, L. Malpica Serrano, A. Crits-Christoph, J. E. Pekar, S. A. Goldstein, A. L. Rasmussen, M. U. G. Kraemer, C. Newman, M. P. G. Koopmans, M. A. Suchard, J. O. Wertheim, P. Lemey, D. L. Robertson, R. F. Garry, E. C. Holmes, A. Rambaut, K. G. Andersen, The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19, Zenodo (2022); http://doi.org/10.5281/zenodo.6786454.
54. M. Bondarenko, D. Kerr, A. Sorichetta, A. Tatem, "Census/projection-disaggregated gridded population datasets for 189 countries in 2020 using Built-Settlement Growth Model (BSGM) outputs" (WorldPop, 2020).
55. M. L. Hazelton, T. M. Davies, Inference based on kernel estimates of the relative risk function in geographical epidemiology. *Biom. J.* **51**, 98–109 (2009). doi:10.1002/bimj.200810495 Medline
56. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020). doi:10.1038/s41591-020-0820-9 Medline
57. W. Wang, J.-H. Tian, X. Chen, R.-X. Hu, X.-D. Lin, Y.-Y. Pei, J.-X. Lv, J.-J. Zheng, F.-H. Dai, Z.-G. Song, Y.-M. Chen, Y.-Z. Zhang, Coronaviruses in wild animals sampled in and around Wuhan at the beginning of COVID-19 emergence. *Virus Evol.* **8**, veac046 (2022). doi:10.1093/ve/veac046 Medline
58. E. C. Holmes, A. Rambaut, K. G. Andersen, Pandemics: Spend on surveillance, not prediction. *Nature* **558**, 180–182 (2018). doi:10.1038/d41586-018-05373-w Medline
59. W.-H. Kong, Y. Li, M.-W. Peng, D.-G. Kong, X.-B. Yang, L. Wang, M.-Q. Liu, SARS-CoV-2 detection in patients with influenza-like illness. *Nat. Microbiol.* **5**, 675–678 (2020). doi:10.1038/s41564-020-0713-1 Medline
60. J. Tao, H. Gao, S. Zhu, L. Yang, D. He, Influenza versus COVID-19 cases among influenza-like illness patients in travelers from Wuhan to Hong Kong in January 2020. *Int. J. Infect. Dis.* **101**, 323–325 (2020). doi:10.1016/j.ijid.2020.09.1474 Medline
61. J. Bai, F. Shi, J. Cao, H. Wen, F. Wang, S. Mubarik, X. Liu, Y. Yu, J. Ding, C. Yu, The epidemiological characteristics of deaths with COVID-19 in the early stage of epidemic in Wuhan, China. *Glob. Health Res. Policy* **5**, 54 (2020). doi:10.1186/s41256-020-00183-y Medline
62. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, Z. Feng, Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi:10.1056/NEJMoa2001316 Medline
63. Y. Jia, Z. Zheng, Q. Zhang, M. Li, X. Liu, Associations of spatial aggregation between neighborhood facilities and the population of age groups based on points-of-interest data. *Sustainability (Basel)* **12**, 1692 (2020). doi:10.3390/su12041692
64. F. Maussion, TimoRoth, R. Bell, F. Li, J. Landmann, M. Dusch, "fmaussion/salem: v0.3.7" (Zenodo, 2021); https://zenodo.org/record/596573.
65. D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, Z. Peng, Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020). doi:10.1001/jama.2020.1585 Medline
66. D. Wang, J. Cai, T. Shi, Y. Xiao, X. Feng, M. Yang, W. Li, W. Liu, L. Yu, Z. Ye, T. Xu, J. Ma, M. Li, W. Chen, Epidemiological characteristics and the entire evolution of coronavirus disease 2019 in Wuhan, China. *Respir. Res.* **21**, 257 (2020). doi:10.1186/s12931-020-01525-7 Medline
67. F. Li, Y.-Y. Li, M.-J. Liu, L.-Q. Fang, N. E. Dean, G. W. K. Wong, X.-B. Yang, I. Longini, M. E. Halloran, H.-J. Wang, P.-L. Liu, Y.-H. Pang, Y.-Q. Yan, S. Liu, W. Xia, X.-X. Lu, Q. Liu, Y. Yang, S.-Q. Xu, Household transmission of SARS-CoV-2 and risk factors for susceptibility and infectivity in Wuhan: A retrospective observational study. *Lancet Infect. Dis.* **21**, 617–628 (2021). doi:10.1016/S1473-3099(20)30981-6 Medline
68. K. Wernike, A. Aebischer, A. Michelitsch, D. Hoffmann, C. Freuling, A. Balkema-Buschmann, A. Graaf, T. Müller, N. Osterrieder, M. Rissmann, D. Rubbenstroth, J. Schön, C. Schulz, J. Trimpert, L. Ulrich, A. Volz, T. Mettenleiter, M. Beer, Multi-species ELISA for the detection of antibodies against SARS-CoV-2 in animals. *Transbound. Emerg. Dis.* **68**, 1779–1785 (2021). doi:10.1111/tbed.13926 Medline
69. X. Zhao, D. Chen, R. Szabla, M. Zheng, G. Li, P. Du, S. Zheng, X. Li, C. Song, R. Li, J.-T. Guo, M. Junop, H. Zeng, H. Lin, Broad and differential animal angiotensin-converting enzyme 2 receptor usage by SARS-CoV-2. *J. Virol.* **94**, e00940-20 (2020). doi:10.1128/JVI.00940-20 Medline
70. A. Z. Mykytyn, M. M. Lamers, N. M. A. Okba, T. I. Breugem, D. Schipper, P. B. van den Doel, P. van Run, G. van Amerongen, L. de Waal, M. P. G. Koopmans, K. J. Stittelaar, J. M. A. van den Brand, B. L. Haagmans, Susceptibility of rabbits to SARS-CoV-2. *Emerg. Microbes Infect.* **10**, 1–7 (2021). doi:10.1080/22221751.2020.1868951 Medline
71. P. Chen, J. Wang, X. Xu, Y. Li, Y. Zhu, X. Li, M. Li, P. Hao, Molecular dynamic simulation analysis of SARS-CoV-2 spike mutations and evaluation of ACE2 from pets and wild animals for infection risk. *Comput. Biol. Chem.* **96**, 107613 (2022). doi:10.1016/j.compbiolchem.2021.107613 Medline

72. V. L. Hale, P. M. Dennis, D. S. McBride, J. M. Nolting, C. Madden, D. Huey, M. Ehrlich, J. Grieser, J. Winston, D. Lombardi, S. Gibson, L. Saif, M. L. Killian, K. Lantz, R. M. Tell, M. Torchetti, S. Robbe-Austerman, M. I. Nelson, S. A. Faith, A. S. Bowman, SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature* **602**, 481–486 (2022). doi:10.1038/s41586-021-04353-x Medline
73. S. M. Porter, A. E. Hartwig, H. Bielefeldt-Ohmann, A. M. Bosco-Lauth, J. J. Root, Susceptibility of wild canids to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). bioRxiv 478082 [Preprint] (2022); https://doi.org/10.1101/2022.01.27.478082.
74. L. Jemeršić, I. Lojkić, N. Krešić, T. Keros, T. A. Zelenika, L. Jurinović, D. Skok, I. Bata, J. Boras, B. Habrun, D. Brnić, Investigating the presence of SARS CoV-2 in free-living and captive animals. *Pathogens* **10**, 635 (2021). doi:10.3390/pathogens10060635 Medline
75. C. S. Lupala, V. Kumar, X.-D. Su, C. Wu, H. Liu, Computational insights into differential interaction of mammalian angiotensin-converting enzyme 2 with the SARS-CoV-2 spike receptor binding domain. *Comput. Biol. Med.* **141**, 105017 (2022). doi:10.1016/j.compbiomed.2021.105017 Medline
76. C. D. Eckstrand, T. J. Baldwin, K. A. Rood, M. J. Clayton, J. K. Lott, R. M. Wolking, D. S. Bradway, T. Baszler, An outbreak of SARS-CoV-2 with high mortality in mink (Neovison vison) on multiple Utah farms. *PLOS Pathog.* **17**, e1009952 (2021). doi:10.1371/journal.ppat.1009952 Medline
77. N. Oreshkova, R. J. Molenaar, S. Vreman, F. Harders, B. B. Oude Munnink, R. W. Hakze-van der Honing, N. Gerhards, P. Tolsma, R. Bouwstra, R. S. Sikkema, M. G. Tacken, M. M. de Rooij, E. Weesendorp, M. Y. Engelsma, C. J. Bruschke, L. A. Smit, M. Koopmans, W. H. van der Poel, A. Stegeman, SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. *Euro Surveill.* **25**, (2020). doi:10.2807/1560-7917.ES.2020.25.23.2001005 Medline
78. A. S. Hammer, M. L. Quaade, T. B. Rasmussen, J. Fonager, M. Rasmussen, K. Mundbjerg, L. Lohse, B. Strandbygaard, C. S. Jørgensen, A. Alfaro-Núñez, M. W. Rosenstierne, A. Boklund, T. Halasa, A. Fomsgaard, G. J. Belsham, A. Bøtner, SARS-CoV-2 transmission between mink (*Neovison vison*) and humans, Denmark. *Emerg. Infect. Dis.* **27**, 547–551 (2021). doi:10.3201/eid2702.203794 Medline
79. Z. Song, L. Bao, W. Deng, J. Liu, E. Ren, Q. Lv, M. Liu, F. Qi, T. Chen, R. Deng, F. Li, Y. Liu, Q. Wei, H. Gao, P. Yu, Y. Han, W. Zhao, J. Zheng, X. Liang, F. Yang, C. Qin, Integrated histopathological, lipidomic, and metabolomic profiles reveal mink is a useful animal model to mimic the pathogenicity of severe COVID-19 patients. *Signal Transduct. Target. Ther.* **7**, 29 (2022). doi:10.1038/s41392-022-00891-6 Medline
80. H.-L. Zhang, Y.-M. Li, J. Sun, Y.-Y. Zhang, T.-Y. Wang, M.-X. Sun, M.-H. Wang, Y.-L. Yang, X.-L. Hu, Y.-D. Tang, J. Zhao, X. Cai, Evaluating angiotensin-converting enzyme 2-mediated SARS-CoV-2 entry across species. *J. Biol. Chem.* **296**, 100435 (2021). doi:10.1016/j.jbc.2021.100435 Medline
81. K. L. Stout, "'Wuhan SARS': Tracing the origin of the new virus to China's wild animal markets" (YouTube, 2020); https://www.youtube.com/watch?v=Je0_U2ym_r0.

ACKNOWLEDGMENTS

We thank the researchers who generated the geospatial and environmental sample data and the members of the China team involved in producing the WHO mission report for the maps that made this work possible. We thank Michael Standaert, Bonnie LaFleur, @babarlelephant, Maciej Boni, Florence Débarre and Benjamin Pierce for comments and assistance. We thank worldpop.org for making population density and demographic data from Wuhan freely available. We thank both the patients and the clinicians and researchers whose data made this research possible. We thank five reviewers for insightful comments and feedback. **Funding:** This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00015 (MW). JIL acknowledges support from the NIH (5T32AI007244-38). SAG acknowledges support from the NIH (F32AI152341). JEP acknowledges support from the NIH (T15LM011271). JOW acknowledges support from NIH (AI135992 and AI136056). DLR acknowledges support of the Medical Research Council (MC_UU_12014/12) and the Wellcome Trust (220977/Z/20/Z). MAS, PL and AR acknowledge the support of the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network), the European Research Council (grant agreement no. 725422 – ReservoirDOCS) and NIH grant R01AI153044. ALR is supported by the Canadian Institutes of Health Research as part of the Coronavirus Variants Rapid Response Network (CoVaRR-Net; CIHR FRN#175622) and acknowledges that VIDO receives operational funding from the Canada Foundation for Innovation – Major Science Initiatives Fund and from the Government of Saskatchewan through Innovation Saskatchewan and the Ministry of Agriculture. MK receives funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 874735 (VEO, Versatile Emerging infectious disease Observatory). RFG is supported by the NIH (R01AI132223, R01AI132244, U19AI142790, U54CA260581, U54HG007480, OT2HL158260), the Coalition for Epidemic Preparedness Innovation, the Wellcome Trust Foundation, Gilead Sciences, and the European and Developing Countries Clinical Trials Partnership Programme. ECH is supported by an Australian Research Council Laureate Fellowship (FL170100022). KGA is supported by the NIH (U19AI135995, U01AI151812, and UL1TR002550). **Contributions:** Conceptualization: MW, KGA; Methodology: MW, JIL, AC-C, LM, JEP, MUGK, MAS, ALR, DLR, SAG, AR, JOW, RFG, PL, ECH, KGA; Software: LM, JIL, JEP, JOW, PL, AR; Validation: MW, LM, JIL, JEP, PL, JOW, KGA; Formal analysis: MW, JIL, AC-C, LM, JEP, MUGK, MAS, ALR, DLR, SAG, AR, JOW, RFG, PL, ECH, KGA; Investigation: MW, JIL, AC-C, LM, JEP, MUGK, MAS, MK, ALR, DLR, CN, SAG, AR, JOW, RFG, PL, ECH, KGA; Resources: MW, JOW, KGA; Data Curation: MW, AR, KGA; Writing - original draft preparation: MW, RFG; Writing - review and editing: MW, JIL, AC-C, LM, JEP, MUGK, MAS, MK, ALR, CN, DLR, SAG, AR, JOW, RFG, PL, ECH, KGA; Visualization: MW, JIL, LM, JEP, ALR, AR, JOW, RFG, PL, ECH, KGA; Supervision: MW, JOW, KGA; Project administration: MW, KGA. Funding acquisition: MW, JIL, AC-C, LM, JEP, MUGK, MAS, ALR, DLR, SAG, AR, JOW, RFG, PL, ECH, KGA. **Competing interests:** JOW receives funding from the CDC via contracts to his institution unrelated to this research. MAS receives funding from Janssen Research & Development, US Food & Drug Administration, and the US Department of Veterans Affairs via contracts and grants unrelated to this research. RFG is a co-founder of Zalgen Labs, a biotechnology company developing countermeasures for emerging viruses. MW, ALR, AR, MAS, ECH, SAG, JOW, and KGA have received consulting fees and/or provided compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic. MK has participated in the second WHO mission to China to study the origins of the pandemic and has served as scientific advisor on emerging disease preparedness to the Guangdong CDC prior to 2020. **Data and materials availability:** Data and code for this manuscript are available from (*53*). We acquired the Weibo data set from (*26*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abp8715
Materials and Methods
Supplementary Text
Figs. S1 to S18
Tables S1 to S12
References (*54*–*81*)
Data S1 and S2
MDAR Reproducibility Checklist

Submitted 2 March 2022; accepted 18 July 2022
Published online 26 July 2022
10.1126/science.abp8715

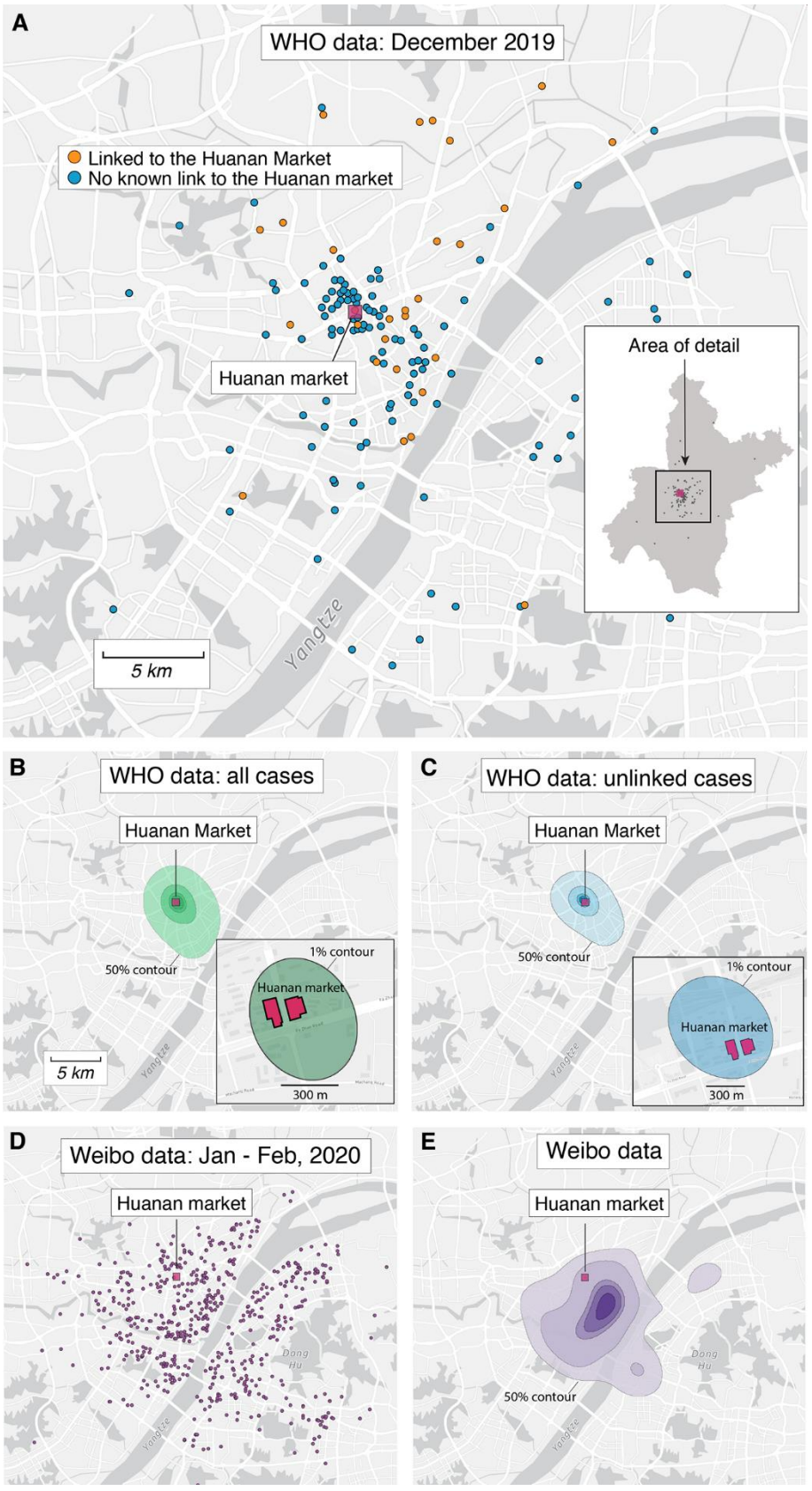
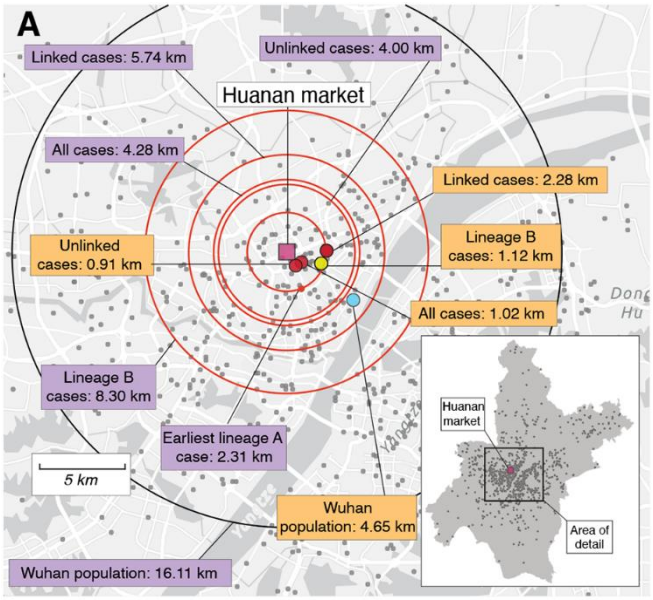
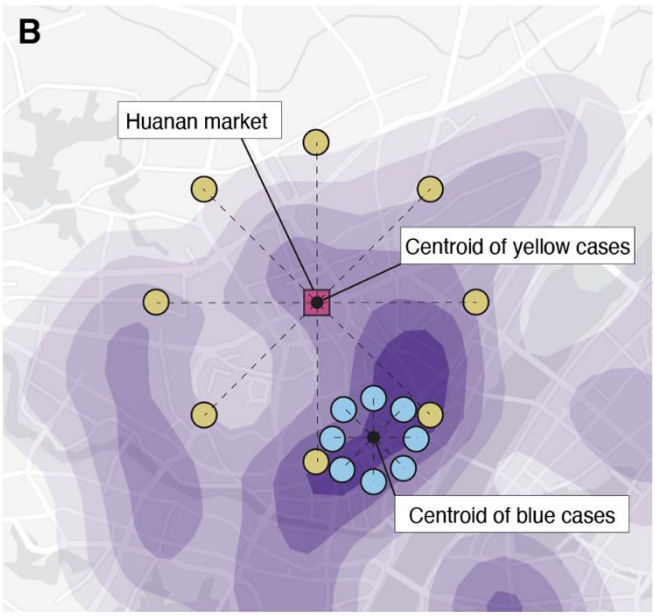


Fig. 1. Spatial patterns of COVID-19 cases in Wuhan in December 2019 and January-February 2020. (**A**) Locations of the 155 cases we extracted from the WHO mission report (*7*). Inset: map of Wuhan with December 2019 case indicated with gray dots. (No cases are obscured by the inset.) In both the inset and the main panel the location of the Huanan market is indicated with a red square. (**B**) Probability density contours reconstructed by a kernel density estimate (KDE) using all 155 COVID-19 cases locations from December 2019. The highest density 50% contour marked is the area for which cases drawn from the probability distribution are as likely to lie inside as outside. Also shown are the highest density 25%, 10%, 5%, and 1% contours. Inset showing an expanded view and the highest density 1% probability density contour. (**C**) Probability density contours reconstructed using the 120 COVID-19 cases locations from December 2019 that were unlinked to the Huanan market. (**D**) Locations of 737 COVID-19 cases from Weibo data dating to January and February of 2020. (**E**) The same highest probability density contours (50% through 1%) for 737 COVID-19 case locations from Weibo data.





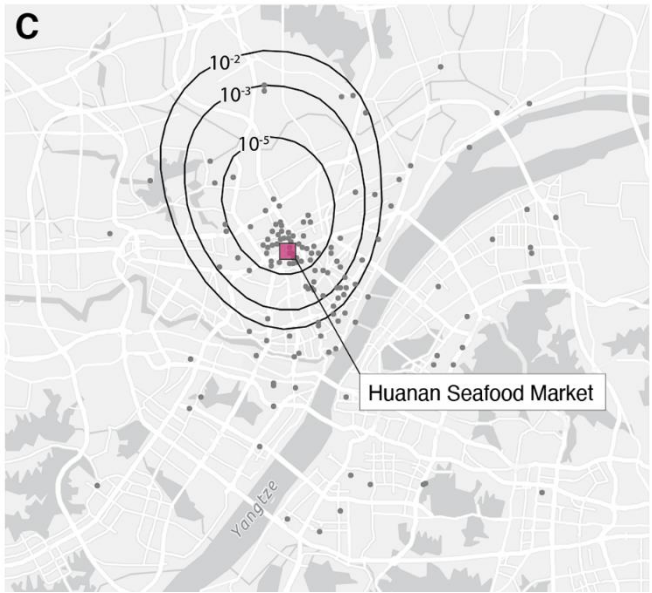


Fig. 2. Spatial analyses. (**A**) Inset: map of Wuhan, with gray dots indicating 1000 random samples from worldpop.com null distribution. Main panel: median distance between Huanan market and (1) worldpop.org null distribution shown with a black circle and (2) December cases shown by red circles (distance to Huanan market depicted in purple boxes). Center-point of Wuhan population density data shown by blue dot. Center-points of December case locations shown by red dots ('all', 'linked' and 'unlinked' cases); dark blue dot (lineage A cases); and yellow dot (lineage B cases). Distance from center-points to Huanan market depicted in orange boxes. (**B**) Schematic showing how cases can be near to, but not centered on, a specific location. We hypothesized that if the Huanan market epicenter of the pandemic then early cases should fall not just unexpectedly near to it but should also be unexpectedly centered on it (see Methods). The blue cases show how cases quite near the Huanan market could nevertheless not be centered on it. (**C**) Tolerance contours based on relative risk of COVID-19 cases in December, 2019 versus data from January-February 2020. The dots show the December case locations. The contours represent the probability of observing that density of December cases within the bounds of the given contour if the December cases had been drawn from the same spatial distribution as the January-February data.

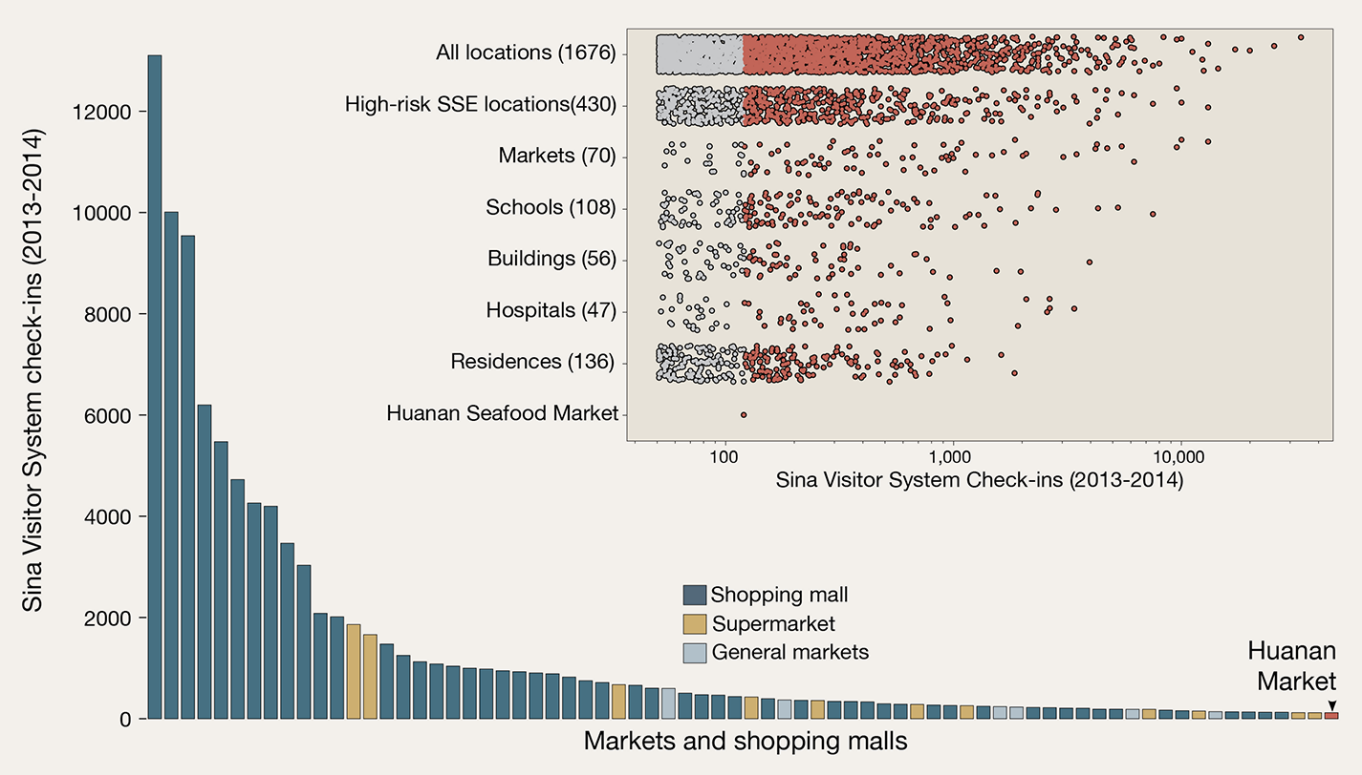


Fig. 3. Visitors to locations throughout Wuhan. Number of social media check-ins in the Sina Visitor System from 2013-2014 as shared by (*33*). Number of visitors to individual markets throughout the city are shown in comparison to the Huanan market. Inset: the total number of check-ins to all individual locations across the city of Wuhan, grouped by category. Locations with more than 50 visitor check-ins are shown, and the locations which received more check-ins than the Huanan market in the same period are shown in red.

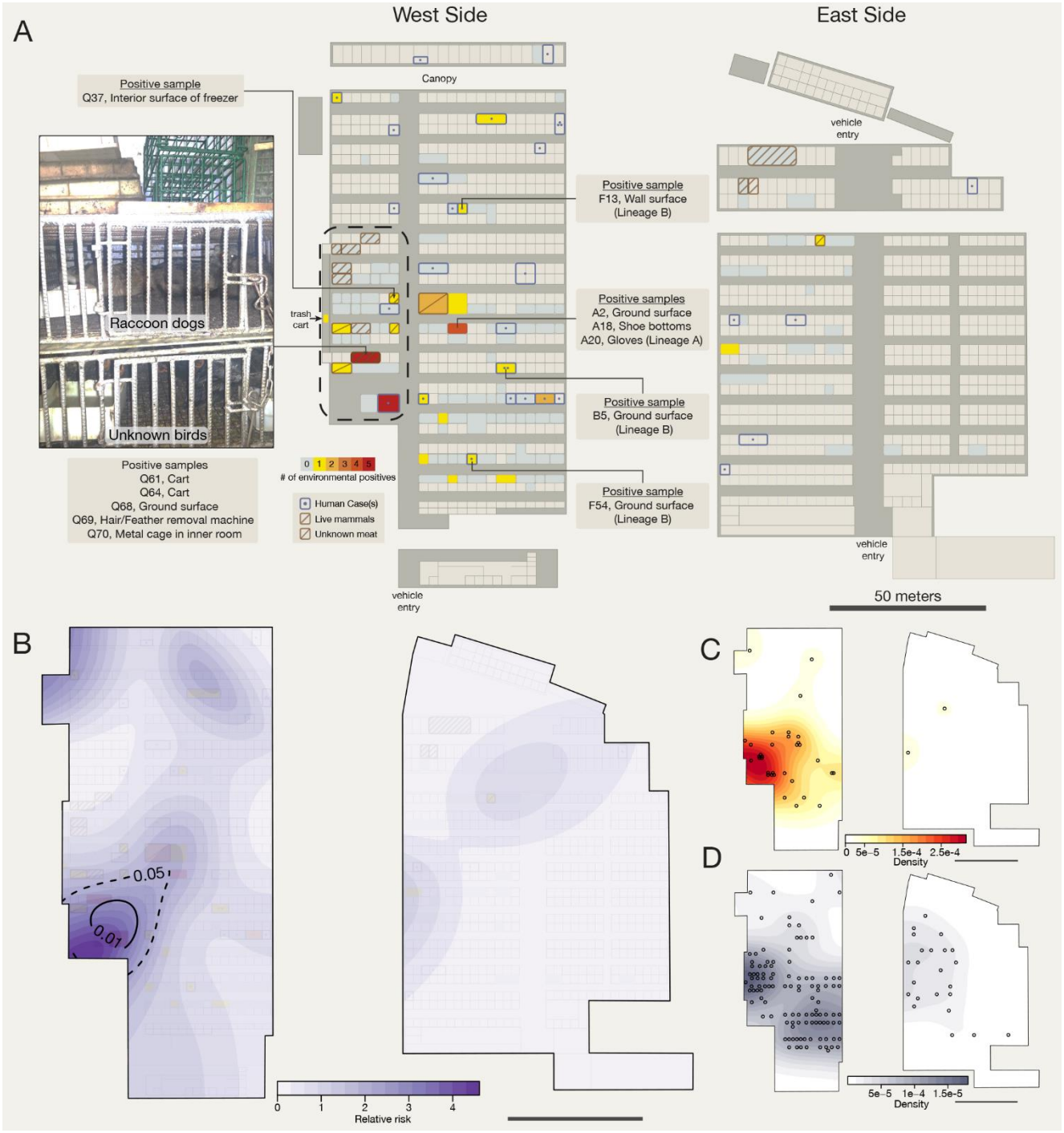


Fig. 4. Map of the Huanan Wholesale Seafood Market. (**A**) Aggregated environmental sampling and human case data from Huanan Market. Captions (left) describe the types of SARS-CoV-2 positive environmental samples obtained from known live animal vendors and (center) from stalls with samples with known virus lineage. Lineage is unknown unless noted; sequencing data has not been released for some samples and many samples were PCR-positive but not sequenced. Image (left) of raccoon dogs in a metal cage, on top of caged birds, taken in business with five positive environmental samples (photo credit: E.C.H.). Rectangle with dashed outline is used to denote the 'wildlife' section of the market. (**B**) Relative risk analysis of positive environmental samples. Tolerance contours enclose regions with statistically significant elevation in density of positive environmental samples relative to the distribution of sampled stalls. (**C**) Distribution of positive environmental samples. Sample locations (centroid of corresponding business) and quantity are shown as black circles. (**D**) Control distribution for relative risk analysis. All businesses investigated with environmental sampling are shown as black circles (one per business, whether or not a positive sample was found). See table S12 for details on stalls that were SARS-CoV-2-negative.

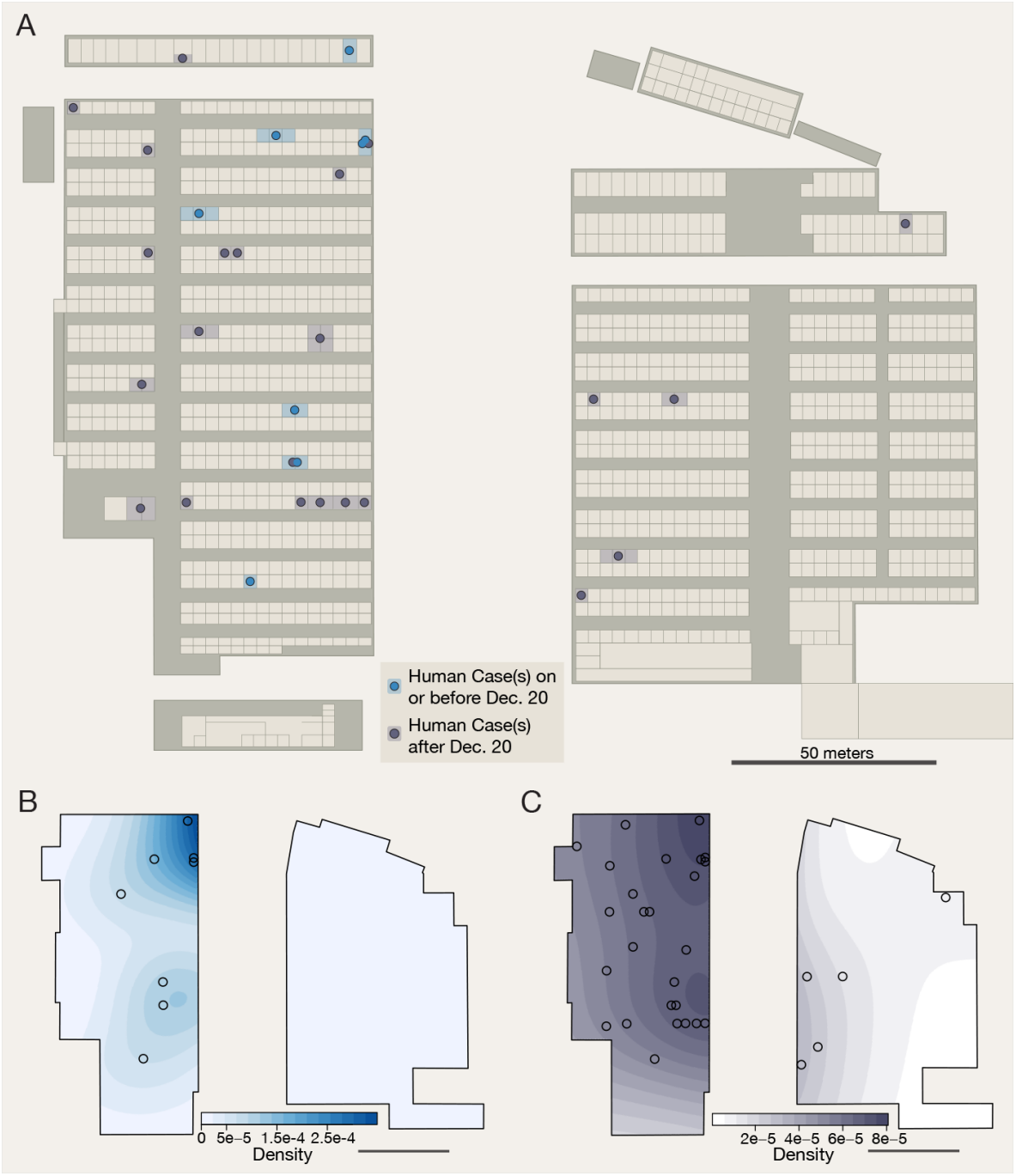


Fig. 5. Location and timing of human cases in Huanan market. (**A**) Outline colors correspond to the timing of the first known case in each business. Individual case timing is denoted by marker color and shown within the outlined business. (**B**) Distribution of known cases on or before December 20th, 2019. Locations of each case are shown as a black circle. (**C**) Distribution of all known human cases in Huanan Market. See table S11 for details on SARS-CoV-2 positive human cases with the Huanan market.

Table 1. Live mammals traded at the Huanan market in November and December 2019

Species (susceptibility*)	Family (susceptibility*)	Order (susceptibility*)	Observed at Huanan market, November 2019
Raccoon dog (*Nyctereutes procyonoides*) (Y)	Canidae (Y)	Carnivora (Y)	Y
Amur hedgehog (*Erinaceus amurensis*)	Erinaceidae	Eulipotyphla	Y
Hog badger (*Arctonyx albogularis*) (Y)	Mustelidae (Y)	Carnivora (Y)	Y
Asian badger (*Meles leucurus*)	Mustelidae (Y)	Carnivora (Y)	Y
Chinese hare (*Lepus sinensis*)	Leporidae (Y)	Lagomorpha (Y)	Y
Chinese bamboo rat (*Rhizomys sinensis*) (Y)	Spalacidae (Y)	Rodentia (Y)	Y
Malayan porcupine (*Hystrix brachyura*)	Hystricidae	Rodentia (Y)	Y
Chinese muntjac (*Muntiacus reevesi*)	Cervidae (Y)	Artiodactyla (Y)	Y
Marmot (*Marmota himalayana*)	Sciuridae	Rodentia (Y)	Y
Red fox (*Vulpes vulpes*) (Y)	Canidae (Y)	Carnivora (Y)	Y
Siberian weasel (*Mustela sibirica*)	Mustelidae (Y)	Carnivora (Y)	N†
Pallas's squirrel (*Callosciurus erythraeus*)	Sciuridae	Rodentia (Y)	N
Masked palm civet (*Paguma larvata*) (Y)	Viverridae (Y)	Carnivora (Y)	N
Coypu (*Myocastor coypus*)	Echimyidae	Rodentia (Y)	N
Mink (*Neovison vison*) (Y)	Mustelidae (Y)	Carnivora (Y)	N
Red squirrel (*Sciurus vulgaris*)	Sciuridae	Rodentia (Y)	N
Wild boar (*Sus scrofa*) (Y)	Suidae (Y)	Artiodactyla (Y)	N
Complex-toothed flying squirrel (*Trogopterus xanthipes*)	Sciuridae	Rodentia (Y)	N

*Based on live susceptibility findings, serological findings, or ACE2-binding assays. See table S5 for details and associated references.
†Animals listed as "No" were, however, present at Wuhan markets during the 2017–2019 study period (*8*).

The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic

Michael WorobeyJoshua I. LevyLorena Malpica SerranoAlexander Crits-ChristophJonathan E. PekarStephen A. GoldsteinAngela L. RasmussenMoritz U. G. KraemerChris NewmanMarion P. G. KoopmansMarc A. SuchardJoel O. WertheimPhilippe LemeyDavid L. RobertsonRobert F. GarryEdward C. HolmesAndrew RambautKristian G. Andersen

Science, **Ahead of Print** • DOI: 10.1126/science.abp8715

Cite as: J. E. Pekar *et al.*, *Science* 10.1126/science.abp8337 (2022).

The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2

Jonathan E. Pekar[1,2]*, Andrew Magee[3], Edyth Parker[4], Niema Moshiri[5], Katherine Izhikevich[5,6], Jennifer L. Havens[1], Karthik Gangavarapu[3], Lorena Mariana Malpica Serrano[7], Alexander Crits-Christoph[8], Nathaniel L. Matteson[4], Mark Zeller[4], Joshua I. Levy[4], Jade C. Wang[9], Scott Hughes[9], Jungmin Lee[10], Heedo Park[10,11], Man-Seong Park[10,11], Katherine Ching Zi Yan[12], Raymond Tzer Pin Lin[12], Mohd Noor Mat Isa[13], Yusuf Muhammad Noor[13], Tetyana I. Vasylyeva[14], Robert F. Garry[15,16,17], Edward C. Holmes[18], Andrew Rambaut[19], Marc A. Suchard[3,20,21]*, Kristian G. Andersen[4,22]*, Michael Worobey[7]*, Joel O. Wertheim[14]*

[1]Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093, USA. [2]Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093, USA. [3]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. [4]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA. [5]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA 92093, USA. [6]Department of Mathematics, University of California San Diego, La Jolla, CA 92093, USA. [7]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [8]W. Harry Feinstone Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA. [9]New York City Public Health Laboratory, New York City Department of Health and Mental Hygiene, New York, NY 11101, USA. [10]Department of Microbiology, Institute for Viral Diseases, Biosafety Center, College of Medicine, Korea University, Seoul, South Korea. [11]BK21 Graduate Program, Department of Biomedical Sciences, Korea University College of Medicine, Seoul, 02841, Republic of Korea. [12]National Public Health Laboratory, National Centre for Infectious Diseases, Singapore. [13]Malaysia Genome and Vaccine Institute, Jalan Bangi, 43000 Kajang, Selangor, Malaysia. [14]Department of Medicine, University of California San Diego, La Jolla, CA 92093, USA. [15]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA 70112, USA. [16]Zalgen Labs, LCC, Frederick, MD 21703 USA. [17]Global Virus Network (GVN), Baltimore, MD 21201, USA. [18]Sydney Institute for Infectious Diseases, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, NSW 2006, Australia. [19]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3FL, UK. [20]Department of Biomathematics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA. [21]Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA 90095, USA. [22]Scripps Research Translational Institute, La Jolla, CA 92037, USA.

***Corresponding author. Email: jepekar@ucsd.edu (J.E.P.); msuchard@ucla.edu (M.A.S.); andersen@scripps.edu (K.G.A.); worobey@arizona.edu (M.W.); jwertheim@health.ucsd.edu (J.O.W.)**

Understanding the circumstances that lead to pandemics is important for their prevention. Here, we analyze the genomic diversity of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) early in the coronavirus disease 2019 (COVID-19) pandemic. We show that SARS-CoV-2 genomic diversity before February 2020 likely comprised only two distinct viral lineages, denoted A and B. Phylodynamic rooting methods, coupled with epidemic simulations, reveal that these lineages were the result of at least two separate cross-species transmission events into humans. The first zoonotic transmission likely involved lineage B viruses around 18 November 2019 (23 October–8 December), while the separate introduction of lineage A likely occurred within weeks of this event. These findings indicate that it is unlikely that SARS-CoV-2 circulated widely in humans prior to November 2019 and define the narrow window between when SARS-CoV-2 first jumped into humans and when the first cases of COVID-19 were reported. As with other coronaviruses, SARS-CoV-2 emergence likely resulted from multiple zoonotic events.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the coronavirus disease 19 (COVID-19) pandemic that caused more than 5 million confirmed deaths in the two years following its detection at the Huanan Seafood Wholesale Market (hereafter the 'Huanan market') in December 2019 in Wuhan, China (*1*–*3*). As the original outbreak spread to other countries, the diversity of SARS-CoV-2 quickly increased and led to the emergence of multiple variants of concern, but the beginning of the pandemic was marked by two major lineages denoted 'A' and 'B' (*4*).

Lineage B has been the most common throughout the pandemic and includes all eleven sequenced genomes from humans directly associated with the Huanan market, including the earliest sampled genome, Wuhan/IPBCAMS-WH-01/2019, and the reference genome, Wuhan/Hu-1/2019 (hereafter 'Hu-1') (*5*), sampled on 24 and 26 December 2019, respectively. The earliest lineage A viruses, Wuhan/IME-WH01/2019 and Wuhan/WH04/2020, were sampled on 30 December 2019 and 5 January 2020, respectively (*6*). Lineage A differs from lineage B by two nucleotide substitutions, C8782T and T28144C, which are also found in related coronaviruses from *Rhinolophus* bats (*4*), the presumed host reservoir (*7*). Lineage B viruses have a 'C/T' pattern at these key sites (C8782, T28144), whereas lineage A viruses have a 'T/C' pattern (C8782T, T28144C). The earliest lineage A genomes from humans lack a direct epidemiological connection to the

Huanan market, but were sampled from individuals who lived or had recently stayed close to the market (*8*). It has been hypothesized that lineages A and B emerged separately (*9*), but 'C/C' and 'T/T' genomes intermediate to lineages A and B present a challenge to that hypothesis, as their existence suggests within-human evolution of one lineage toward the other via a transitional form.

Questions about these lineages remain: if lineage B viruses are more distantly related to sarbecoviruses from *Rhinolophus* bats, (i) why were lineage B viruses detected earlier than lineage A viruses and (ii) why did lineage B predominate early in the pandemic?

Answering these questions requires determining the ancestral haplotype, the genomic sequence characteristics of the most recent common ancestor (MRCA) at the root of the SARS-CoV-2 phylogeny. In this study, we combined genomic and epidemiological data from early in the COVID-19 pandemic with phylodynamic models and epidemic simulations. We eliminated many of the haplotypes previously suggested as the MRCA of SARS-CoV-2 and show that the pandemic most likely began with at least two separate zoonotic transmissions starting in November 2019.

Results

Erroneous assignment of haplotypes intermediate to lineages A and B

There are 787 near-full length genomes available from lineages A and B sampled by 14 February 2020 (data S1 and S2). However, there are also 20 genomes of intermediate haplotypes from this period containing either T28144C or C8782T but not both mutations: C/C or T/T, respectively.

We identified numerous instances of C/C and T/T genomes sharing rare mutations with lineage A or lineage B viruses, often sequenced in the same laboratory, indicating these intermediate genomes are likely artifacts of contamination or bioinformatics (*10*), similar to findings from our analysis of the emergence of SARS-CoV-2 in North America (*11*) (fig. S1 and supplementary text). We confirmed that a C/C genome from South Korea sharing three such mutations had low sequencing depth at position 28144 (≤10x), a T/T genome sampled in Singapore had low coverage at both 8782 and 28144 (≤10x), and three T/T genomes sampled in Wuhan had low sequencing depth and indeterminate nucleotide assignment at position 8782 (table S1). Further, the authors of eleven C/C genomes sampled in Wuhan and Sichuan confirmed that low sequencing depth at position 8782 led to the erroneous assignment of intermediate haplotypes.

C/C and T/T genomes continue to be observed throughout the pandemic as a result of convergent evolution, including T/T aboard the Diamond Princess cruise ship outbreak and subsequent COVID-19 waves in New York City and San Diego (fig. S2 to S5 and supplementary text). Instances of convergent evolution are identifiable because SARS-CoV-2 phylogenies exist in 'near-perfect' tree space where topology can be inferred with high accuracy (*12*). These findings cast doubt on the claim that transitional C/C or T/T haplotypes between lineages A and B circulated in humans, reopening the door to the hypothesis that lineages A and B represent separate zoonotic introductions.

Progenitor genome reconstruction

To better understand SARS-CoV-2 mutational patterns, we reconstructed the genome of a hypothetical progenitor of SARS-CoV-2. Using maximum likelihood ancestral state reconstruction across 15 non-recombinant regions of SARS-CoV-2 and closely related sarbecovirus genomes sampled from bats and pangolins (*13*), we inferred the genome of this recombinant common ancestor ("recCA") (figs. S6 and S7 and supplementary text). The recCA differed from Hu-1 by just 381 substitutions, including C8782T and T28144C. It is more informative than an outgroup sarbecovirus because it accounts for the closest relative across all recombinant segments (figs. S8 to S14 and supplementary text) (*14*), and, as an internal node on the phylogeny, is more genetically similar to SARS-CoV-2 than any extant sarbecovirus.

Reversions across the early pandemic phylogeny

The ubiquity of SARS-CoV-2 reversions (*i.e.*, mutations from Hu-1 toward the recCA) indicates that genetic similarity to related viruses is a poor proxy for the ancestral haplotype. We observe 23 unique reversions and 631 unique substitutions (excluding reversions) across the SARS-CoV-2 phylogeny from the COVID-19 pandemic up to 14 February 2020 (Fig. 1). Substitutions were overrepresented at the 381 sites separating the recCA from Hu-1 (23/381 = 6.04%), compared with substitutions at all other sites (631/29,134 = 2.17%).

Most reversions were C-to-T mutations (19/23 = 82.6%), matching the mutational bias of SARS-CoV-2 (*15–17*). Genomes with C-to-T reversions can be found within lineage A, including C18060T (lineage A.1; *e.g.*, WA1) and C29095T (*e.g.*, 20SF012), as well as C24023T, C25000T, C4276T, and C22747T in mid-late January and February 2020. Hence, triple revertant genomes, like WA1 and 20SF012, are neither unique nor rare. We also identified a lineage A genome (Malaysia/MKAK-CL-2020-6430/2020), sampled on 4 February 2020 from a Malaysian citizen traveling from Wuhan whose only four mutations from Hu-1 are all reversions (lineage A.1+T6025C) (Fig. 1). Therefore, no highly revertant haplotype can automatically be assumed to represent the MRCA of SARS-CoV-2, especially when these reversions are most often the result of C-to-T mutations. In fact, we continue to observe these reversion patterns throughout the pandemic, including in the emergence of WHO-named variants (figs. S15 and S16).

Inferring the MRCA of SARS-CoV-2

To infer the ancestral SARS-CoV-2 haplotype, we developed a non-reversible, random-effects substitution process model in a Bayesian phylodynamic framework that simultaneously reconstructs the underlying coalescent processes and the sequence of the MRCA of the SARS-CoV-2 phylogeny. The random-effects substitution model captures the C-to-T transition and G-to-T transversion biases (fig. S17 and supplementary text). Using this model, referred to as the unconstrained rooting (fig. S18A), we inferred the ancestral haplotype of the 787 lineage A and B genomes sampled by 14 February 2020.

Our unconstrained rooting strongly favors a lineage B or C/C ancestral haplotype and shows that a lineage A ancestral haplotype is inconsistent with the molecular clock [Bayes factor (BF) = 48.1] (Table 1). Lineage B exhibits more divergence from the root of the tree than would be expected if lineage A were the ancestral virus in humans (figs. S19 and S20). The T/T ancestral haplotype was also disfavored (BF>10), likely because of the C-to-T transition bias (fig. S17). We acknowledge that the timing of the earliest sampled lineage B genomes associated with the Huanan market could bias rooting inference toward lineage B haplotypes; however, lineage A was still disfavored after excluding all market-associated genomes (BF=11.0).

Even though sequence similarity to closely related sarbecoviruses alone is insufficient to determine the SARS-CoV-2 ancestral haplotype, this similarity can inform phylodynamic inference. Rather than rely on outgroup rooting [fig. S18B and (*18*)], we developed a rooting method that assigns the recCA as the progenitor of the inferred SARS-CoV-2 MRCA (fig. S18C). As opposed to the unconstrained rooting, the recCA root favored a lineage A haplotype over lineage B, although support for C/C was unchanged (Table 1). Our results were insensitive to the method of breakpoint identification in the recCA (supplementary text).

The A.1 and A+C29095T proposed ancestral haplotypes were strongly rejected by all the phylodynamic analyses, even when rooting with recCA or bat sarbecovirus outgroups, which include both C18060T and C29095T (Table 1 and data S3). Hence, WA1-like and 20SF012-like haplotypes cannot plausibly represent the MRCA of SARS-CoV-2 as previously suggested (*19–21*): the similarity of these genomes to the recCA is due to C-to-T reversions. Haplotypes not reported in Table 1 were similarly rejected (data S3).

We inferred the tMRCA for SARS-CoV-2 to be 11 December 2019 (95% HPD: 25 November–12 December) using unconstrained rooting. It has been suggested that a phylogenetic root in lineage A would produce an older time of most recent common ancestor (tMRCA) than a lineage B rooting (*21*). Therefore, we developed an approach to assign a haplotype as the SARS-CoV-2 MRCA and inferred the tMRCA (*i.e.*, A, B, C/C, A.1 or A+C29095T) (fig. S18D). The tMRCA was consistent with the recCA-rooted and fixed ancestral haplotype analyses (table S2 and supplementary text).

We infer only three plausible ancestral haplotypes: lineage A, lineage B, and C/C. However, the inability to reconcile the molecular clock at the outset of the COVID-19 pandemic with a lineage A ancestor without information from related sarbecoviruses (*e.g.*, the recCA) requires us to question the assumption that both lineages A and B resulted from a single introduction.

Separate introductions of lineages A and B

We next sought to determine whether a single introduction from one of the plausible ancestral haplotypes (lineage A, lineage B, or C/C) is consistent with the SARS-CoV-2 phylogeny. We simulated SARS-CoV-2-like epidemics (*22*, *23*) with a doubling time of 3.47 days [95% highest density interval (HDI) across simulations: 1.35–5.44] (*24–26*) to account for the rapid spread of SARS-CoV-2 before it was identified as the etiological agent of COVID-19 (figs. S21 and S22, tables S3 and S4, and supplementary text). We then simulated coalescent processes and viral genome evolution across these epidemics to determine how frequently we recapitulated the observed SARS-CoV-2 phylogeny.

Lineages A and B comprise 35.2% and 64.8% of the early SARS-CoV-2 genomes, and each lineage is characterized by a large polytomy (*i.e.*, many sampled lineages descending from a single node on the phylogenetic tree), with the base of lineages A and B being the two largest polytomies observed in the early pandemic (Fig. 1). Furthermore, large polytomies are characteristic of SARS-CoV-2 introductions into geographical regions at the start of the pandemic (e.g., fig. S23) (*11*, *27–29*) and would similarly be expected to occur after a successful introduction of SARS-CoV-2 into humans. Congruently, the most common topology in our simulations is a large basal polytomy (with ≥100 descendent lineages), present in 47.5% of simulated epidemics (Fig. 2A).

In contrast, a topology corresponding to a single introduction of an ancestral C/C haplotype, characterized by two clades, each comprising ≥30% of the taxa, possessing a large polytomy at the base, and separated from the MRCA by one mutation (Fig. 2B), was only observed in 0.1% of our simulations. Further, a topology corresponding to a single introduction of an ancestral lineage A or lineage B haplotype, characterized by a large basal polytomy and a large clade, comprising between 30% and 70% of taxa, two mutations from the root with no intermediate genomes, was observed in only 0.5% of our simulations (Fig. 2C, see supplementary text for details).

Our epidemic simulations do not support a single introduction of SARS-CoV-2 giving rise to the observed phylogeny. We therefore quantified the relative support for two introductions resulting in the empirical topology. By synthesizing

posterior probabilities of inferred ancestral haplotypes, frequencies of topologies in epidemic simulations, and the expected relationships between these haplotypes and topologies, we infer strong support favoring separate introductions of lineages A and B (BF=61.6 and BF=60.0 using the recCA and unconstrained rooting, respectively; see Methods). This support is robust across shorter and longer doubling times, varying ascertainment rates, and minimum polytomy size (tables S4 and S5).

If lineages A and B arose from separate introductions, then the MRCA of SARS-CoV-2 was not in humans, and it is the tMRCAs of lineages A and B that are germane to the origins of SARS-CoV-2 (i.e, not the timing of their shared ancestor). Rooting with the recCA, we inferred the median tMRCA of lineage B to be 15 December (95% HPD: 5 December to 23 December) and the median tMRCA of lineage A to be 20 December (95% HPD: 5 December to 29 December) (Fig. 3A). The tMRCA of lineage B consistently predates the tMRCA of lineage A (Fig. 3B). These results are robust to using unconstrained rooting, fixing the ancestral haplotype, and excluding market-associated genomes (Fig. 3, A and B; table S2; and supplementary text).

Timing the introductions of lineages A and B

The primary case, the first human infected with a virus in an outbreak, could precede the tMRCA if basal lineages went extinct during cryptic transmission (*23*, *30*, *31*). The index case, the first identified case, is rarely also the primary case (*32*, *33*). We next used an extension of our previously published framework combining epidemic simulations and phylodynamic tMRCA inference [see Methods; (*23*, *30*, *31*)] to infer the timing of the lineage B and lineage A primary cases, accounting for both the index case symptom onset date and earliest documented COVID-19 hospitalization date.

The earliest unambiguous case of COVID-19, with symptom onset on 10 December and hospitalization on 16 December, was a seafood vendor at the Huanan market. Unfortunately no published genome is available for this case (*8*). Nonetheless, we can reasonably assume this individual had a lineage B virus (supplementary text), as an environmental sample (EPI_ISL_408512) from the stall this vendor operated was lineage B. The earliest lineage A genome (IME-WH01) is from a familial cluster where the earliest symptom onset is 15 December and earliest hospitalization is 25 December (*34*). Accounting for these dates and using the recCA rooting, we inferred the infection date of the lineage B primary case to be 18 November (95% HPD: 23 October to 8 December) and the infection date of the primary case of lineage A to be 25 November (95% HPD: 29 October to 14 December). The lineage B primary case predated that of lineage A in 64.6% of the posterior sample, by a median of 7 days (Fig. 3D and table S6).

Our lineage A and B primary case inference is robust to rooting on the recCA and fixing the plausible ancestral haplotype to lineage A, lineage B, or C/C, as well as different index case dates, accounting for only hospitalization dates, and varying growth rates and ascertainment rates (tables S7 to S10 and supplementary text). Therefore, our results indicate that lineage B was introduced into humans no earlier than late-October and likely in mid-November 2019, and the introduction of lineage A occurred within days to weeks of this event.

We then inferred the number of ascertained infections and hospitalizations arising from these separate introductions. We find that an earlier introduction of lineage B leads to a faster rise in lineage B-associated infections, dominating the simulated epidemics (Fig. 4) and recapitulating the predominance of lineage B observed in China in early 2020 (*35*). Similarly, simulated lineage B hospitalizations are more common than those from lineage A through January 2020 (fig. S24). We observe these patterns regardless of rooting strategy (unconstrained or recCA), ancestral haplotype (B, A, or C/C) (Fig. 4 and tables S11 and S12), and doubling time (figs. S25 to S28).

Minimal cryptic circulation of SARS-CoV-2

We do not see evidence for substantial cryptic circulation before December 2019 (Fig. 4), even if we assume a single introduction (fig. S29 and supplementary text). Our simulated epidemics have a median of three (95% HPD 1-18) cumulative infections at the tMRCA, with 99% of simulated epidemics resulting in at most 33 infections (table S13 and supplementary text). Further, it is unlikely there were any COVID-19 related hospitalizations before December (*36*), as the simulated epidemics show a median of zero (95% HPD: 0–2) hospitalizations by 1 December 2019. These results are in accordance with the lack of a single SARS-CoV-2-positive sample among tens of thousands of serology samples from healthy blood donors from September to December 2019 (*37*) and thousands of specimens obtained from influenza-like illness patients at Wuhan hospitals from October to December 2019 (*34*). Therefore, there was likely extremely low prevalence of SARS-CoV-2 in Wuhan before December 2019. Even when we simulated epidemics with a longer doubling time, resulting in an earlier timing of the primary cases (tables S8 and S10), there were still few infections prior to December 2019 (table S13).

Additional introductions

The extinction rate of our simulated epidemics (*i.e.*, simulations that did not produce self-sustaining transmission chains) indicate there were likely multiple failed introductions of SARS-CoV-2. Similar to our previous findings (*23*), 77.8% of simulated epidemics went extinct. These failed introductions produced a mean of 2.06 infections and 0.10

hospitalizations; hence, failed introductions could easily go unnoticed. If we treat each SARS-CoV-2 introduction, failed or successful, as a Bernoulli trial and simulate introductions until we see two successful introductions, we estimate that eight (95% HPD: 2–23) introductions led to the establishment of both lineage A and B in humans.

Limitations

Our analysis of the putative intermediate haplotypes suggests there remain lineage assignment errors between lineages A and B, particularly of genomes sampled in January and February of 2020, which could influence the precision of the phylogenetic topology and tMRCA inference. Importantly, we lack direct evidence of a virus closely related to SARS-CoV-2 in non-human mammals at the Huanan market or its supply chain. The genome sequence of a virus directly ancestral to SARS-CoV-2 would provide more precision regarding the timing of the introductions of SARS-CoV-2 into humans and the epidemiological dynamics prior to its discovery. Although we simulated epidemics across a range of plausible epidemiological dynamics, our models represent a timeframe prior to the ascertainment of COVID-19 cases and sequencing of SARS-CoV-2 genomes and thus prior to when these models could be empirically validated.

Discussion

The genomic diversity of SARS-CoV-2 during the early pandemic presents a paradox. Lineage A viruses are at least two mutations closer to bat coronaviruses, indicating that the ancestor of SARS-CoV-2 arose from this lineage. However, lineage B viruses predominated early in the pandemic, particularly at the Huanan market, indicating that this lineage began spreading earlier in humans. Further complicating this matter is the molecular clock of SARS-CoV-2 in humans, which rejects a single-introduction origin of the pandemic from a lineage A virus. Here, we resolve this paradox by showing that early SARS-CoV-2 genomic diversity and epidemiology is best explained by at least two separate zoonotic transmissions, in which lineage A and B progenitor viruses were both circulating in non-human mammals prior to their introduction into humans (figs. S30 and S31).

The most probable explanation for the introduction of SARS-CoV-2 into humans involves zoonotic jumps from as-yet undetermined, intermediate host animals at the Huanan market (*34*, *38*, *39*). Through late-2019 the Huanan market sold animals that are known to be susceptible to SARS-CoV-2 infection and capable of intra-species transmission (*40–42*). The presence of potential animal reservoirs, coupled with the timing of the lineage B primary case and the geographic clustering of early cases around the Huanan market (*39*), support the hypothesis that SARS-CoV-2 lineage B jumped into humans at the Huanan market in mid-November 2019.

In a related study (*39*), we show that the two earliest lineage A cases are more closely positioned geographically to the Huanan market than expected compared with other COVID-19 cases in Wuhan in early 2020, despite having no known association with the market. This geographic proximity is consistent with a separate and subsequent origin of lineage A at the Huanan market in late-November 2019. The presence of lineage A virus at the Huanan market was confirmed by Gao *et al.* (*43*) from a sample taken from discarded gloves.

The high extinction rate of SARS-CoV-2 transmission chains, observed in both our simulations and real-world data (*44*), indicates that the two zoonotic events establishing lineages A and B may have been accompanied by additional, cryptic introductions. However, such introductions could easily be missed, particularly if their subsequent transmission chains quickly went extinct or the introduced viruses had a lineage A or B haplotype. Failed introductions of intermediate haplotypes are also possible. Critically, we have no evidence of subsequent zoonotic introductions in late-December leading up to the closure of the Huanan market on 1 January 2020. By then, the susceptible host animals that had been documented at the market during the previous months were no longer found in the Huanan market (*34*).

Other coronavirus epidemics and outbreaks in humans, including SARS-CoV-1, MERS-CoV, and, most recently, porcine deltacoronavirus in Haiti, have been the result of repeated introductions from animal hosts (*45–47*). These repeated introductions were easily identifiable because human viruses in these outbreaks were more closely related to viruses sampled in the animal reservoirs than to other human viruses. However, the genomic diversity within the putative SARS-CoV-2 animal reservoir at the Huanan market was likely shallower than that seen in SARS-CoV-1 and MERS-CoV reservoirs (*45*, *46*, *48*). Hence, even though lineages A and B had nearly identical haplotypes, their MRCA likely existed in an animal reservoir. The ability to disentangle repeated introductions of SARS-CoV-2 from a shallow genetic reservoir has previously been shown in the early SARS-CoV-2 epidemic in Washington state, where two viruses, separated by two mutations, were independently introduced from, and shared an MRCA in, China (figs. S23 and S30 and supplementary text) (*11*).

Successful transmission of both lineage A and B viruses after independent zoonotic events indicates that evolutionary adaptation within humans was not needed for SARS-CoV-2 to spread (*49*). We now know that SARS-CoV-2 can readily spread after reverse-zoonosis to Syrian hamsters (*Mesocricetus auratus*), American mink (*Neovison vison*), and white-tailed deer (*Odocoileus virginianus*), indicating its host generalist capacity (*50–55*). Furthermore, once an animal virus acquires the capacity for human infection and transmission,

the only remaining barrier to spillover is contact between humans and the pathogen. Thereafter, a single zoonotic transmission event indicates the conditions necessary for spillovers have been met, which portends additional jumps. For example, there were at least two zoonotic jumps of SARS-CoV-2 into humans from pet hamsters in Hong Kong (*56*) and dozens from minks to humans on Dutch fur farms (*52*, *53*).

We show that it is highly unlikely that SARS-CoV-2 circulated widely in humans earlier than November 2019 and that there was limited cryptic spread, with, at most, dozens of SARS-CoV-2 infections in the weeks leading up to the inferred tMRCA, but likely far fewer. By late-December, when SARS-CoV-2 was identified as the etiological agent of COVID-19 (*8*), the virus had likely been introduced into humans multiple times as a result of persistent contact with a viral reservoir.

Materials and methods summary

Materials and methods described in full detail can be found in the supplementary materials.

Sequence data

We queried the GISAID database (*57*), GenBank, and National Genomics Data Center of the China National Center for Bioinformatics (CNCB), for complete high-coverage SARS-CoV-2 genomes collected by 14 February 2020, resulting in a dataset of 787 taxa belonging to lineages A and B and 20 taxa with C/C or T/T haplotypes. Genomes were aligned using MAFFT v7.453 (*58*) to the SARS-CoV-2 reference genome (Wuhan/Hu-1/2019) and 388 sites were masked at the 5′ and 3′ ends and at sites based on De Maio *et al.* (*59*). All genome accessions are available in data S1 and S2.

Progenitor genome reconstruction and reversion analysis

We reconstructed the progenitor of SARS-CoV-2, the recombinant common ancestor (the recCA). We (i) inferred a maximum likelihood tree of 31 sarbecovirus genomes (SARS-CoV-2 and 30 closely related sarbecoviruses sampled from bats and pangolins) across 15 predefined non-recombinant regions (*13*) with IQ-TREE v2.0.7 (*60*), (ii) inferred the sequence of the ancestor of SARS-CoV-2 in each tree with TreeTime v0.8.1 (*61*), and (iii) concatenated the resulting sequences. We next inferred a maximum likelihood tree of the 787 SARS-CoV-2 taxa with IQ-TREE and performed ancestral state reconstruction with TreeTime to identify substitutions that were reversions from Wuhan-Hu-1 to the recCA across the SARS-CoV-2 phylogeny.

Phylodynamic inference and epidemic simulations

We performed phylodynamic inference using BEAST v1.10.5 (*62*) with the 787-taxa dataset to infer the ancestral haplotype and the tMRCA of SARS-CoV-2 (and the tMRCAs of lineages A and B), employing a non-reversible random-effects substitution model and exploring unconstrained rooting, recCA-rooting, fixing the ancestral haplotype as a root, and outgroup rooting. SARS-CoV-2-like epidemics were simulated with FAVITES-COVID-Lite v0.0.1 (*22*, *63*) using a scale-free network of 5 million individuals and a customized extension of the SAPHIRE model (*64*), producing coalescent trees on which we simulated mutations. We calculated the Bayes factor comparing the support of two introductions of SARS-CoV-2 to one introduction by considering the posterior probabilities of the four most likely ancestral haplotypes from the phylodynamic inference (Lineage A, Lineage B, C/C, and T/T), the frequencies of the phylogenetic structures associated with introductions of these haplotypes in the epidemic simulations, and equal prior probabilities for each ancestral haplotype and one versus two introductions.

We connected the phylodynamic inference and epidemic simulations via a rejection sampling-based approach (*23*), accounting for the tMRCAs of lineages A and B and the earliest documented COVID-19 illness onset and hospitalization dates. We then inferred the timing of the introductions of lineages A and B and the infections and hospitalizations for each lineage. The proportion of epidemic simulations that went extinct (*i.e.*, no onward transmission by the end of the simulation) was used to approximate the number of SARS-CoV-2 introductions needed to result in two introductions with sustained onward transmission.

REFERENCES AND NOTES

1. E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020). doi:10.1016/S1473-3099(20)30120-1 Medline
2. L.-L. Ren, Y.-M. Wang, Z.-Q. Wu, Z.-C. Xiang, L. Guo, T. Xu, Y.-Z. Jiang, Y. Xiong, Y.-J. Li, X.-W. Li, H. Li, G.-H. Fan, X.-Y. Gu, Y. Xiao, H. Gao, J.-Y. Xu, F. Yang, X.-M. Wang, C. Wu, L. Chen, Y.-W. Liu, B. Liu, J. Yang, X.-R. Wang, J. Dong, L. Li, C.-L. Huang, J.-P. Zhao, Y. Hu, Z.-S. Cheng, L.-L. Liu, Z.-H. Qian, C. Qin, Q. Jin, B. Cao, J.-W. Wang, Identification of a novel coronavirus causing severe pneumonia in human: A descriptive study. *Chin. Med. J. (Engl.)* **133**, 1015–1024 (2020). doi:10.1097/CM9.0000000000000722 Medline
3. H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, S. Beltekian, X. Roser, Coronavirus Pandemic (COVID-19). *Our World in Data* (2022); https://ourworldindata.org/covid-deaths.
4. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020). doi:10.1038/s41564-020-0770-5 Medline
5. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). doi:10.1038/s41586-020-2008-3 Medline
6. R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020). doi:10.1016/S0140-6736(20)30251-8 Medline

7. S. Lytras, J. Hughes, D. Martin, P. Swanepoel, A. de Klerk, R. Lourens, S. L. Kosakovsky Pond, W. Xia, X. Jiang, D. L. Robertson, Exploring the natural origins of SARS-CoV-2 in the light of recombination. *Genome Biol. Evol.* **14**, evac018 (2022). doi:10.1093/gbe/evac018 Medline
8. M. Worobey, Dissecting the early COVID-19 cases in Wuhan. *Science* **374**, 1202–1204 (2021). doi:10.1126/science.abm4454 Medline
9. R. F. Garry, Early appearance of two distinct genomic lineages of SARS-CoV-2 in different Wuhan wildlife markets suggests SARS-CoV-2 has a natural origin. *Virological* (2021); https://virological.org/t/early-appearance-of-two-distinct-genomic-lineages-of-sars-cov-2-in-different-wuhan-wildlife-markets-suggests-sars-cov-2-has-a-natural-origin/691.
10. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, N. Goldman, Issues with SARS-CoV-2 sequencing data. *Virological* (2020); https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
11. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020). doi:10.1126/science.abc8169 Medline
12. J. O. Wertheim, M. Steel, M. J. Sanderson, Accuracy in Near-Perfect Virus Phylogenies. *Syst. Biol.* **71**, 426–438 (2022). doi:10.1093/sysbio/syab069 Medline
13. S. Temmam, K. Vongphayloth, E. Baquero, S. Munier, M. Bonomi, B. Regnault, B. Douangboubpha, Y. Karami, D. Chrétien, D. Sanamxay, V. Xayaphet, P. Paphaphanh, V. Lacoste, S. Somlor, K. Lakeomany, N. Phommavanh, P. Pérot, O. Dehan, F. Amara, F. Donati, T. Bigot, M. Nilges, F. A. Rey, S. van der Werf, P. T. Brey, M. Eloit, Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* **604**, 330–336 (2022). doi:10.1038/s41586-022-04532-4 Medline
14. J. B. Pease, M. W. Hahn, More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution* **67**, 2376–2384 (2013). doi:10.1111/evo.12118 Medline
15. J. Ratcliff, P. Simmonds, Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. *Virology* **556**, 62–72 (2021). doi:10.1016/j.virol.2020.12.018 Medline
16. P. Simmonds, Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *MSphere* **5**, e00408-20 (2020). doi:10.1128/mSphere.00408-20 Medline
17. P. Simmonds, M. A. Ansari, Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. *PLOS Pathog.* **17**, e1009596 (2021). doi:10.1371/journal.ppat.1009596 Medline
18. 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
19. J. D. Bloom, Recovery of Deleted Deep Sequencing Data Sheds More Light on the Early Wuhan SARS-CoV-2 Epidemic. *Mol. Biol. Evol.* **38**, 5211–5224 (2021). doi:10.1093/molbev/msab246 Medline
20. M. A. Caraballo-Ortiz, S. Miura, M. Sanderford, T. Dolker, Q. Tao, S. Weaver, S. L. K. Pond, S. Kumar, TopHap: Rapid inference of key phylogenetic structures from common haplotypes in large genome collections with limited diversity. *Bioinformatics* **38**, 2719–2726 (2022). doi:10.1093/bioinformatics/btac186 Medline
21. S. Kumar, Q. Tao, S. Weaver, M. Sanderford, M. A. Caraballo-Ortiz, S. Sharma, S. L. K. Pond, S. Miura, An Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic. *Mol. Biol. Evol.* **38**, 3046–3059 (2021). doi:10.1093/molbev/msab118 Medline
22. 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
23. J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, J. O. Wertheim, Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**, 412–417 (2021). doi:10.1126/science.abf8003 Medline
24. S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, L. Y. Huang, A. Hultgren, E. Krasovich, P. Lau, J. Lee, E. Rolf, J. Tseng, T. Wu, The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* **584**, 262–267 (2020). doi:10.1038/s41586-020-2404-8 Medline
25. A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, D. Sledge, The challenges of modeling and forecasting the spread of COVID-19. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 16732–16738 (2020). doi:10.1073/pnas.2006520117 Medline
26. S. Sanche, Y. T. Lin, C. Xu, E. Romero-Severson, N. Hengartner, R. Ke, High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerg. Infect. Dis.* **26**, 1470–1477 (2020). doi:10.3201/eid2607.200282 Medline
27. 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, 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, H. Y. Chu, M. Boeckh, J. A. Englund, M. Famulare, B. R. Lutz, D. A. Nickerson, M. J. Rieder, L. M. Starita, M. Thompson, J. Shendure, T. Bedford, A. Adler, E. Brandstetter, S. Cho, C. D. Frazar, D. Giroux, P. D. Han, J. Hadfield, S. Huang, M. L. Jackson, A. Kiavand, L. E. Kimball, K. Lacombe, J. Logue, V. Lyon, K. L. Newman, M. Richardson, T. R. Sibley, M. L. Zigman Suchsland, M. Truong, C. R. Wolf, Seattle Flu Study Investigators, Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020). doi:10.1126/science.abc0523 Medline
28. M. Zeller, K. Gangavarapu, C. Anderson, A. R. Smither, J. A. Vanchiere, R. Rose, D. J. Snyder, G. Dudas, A. Watts, N. L. Matteson, R. Robles-Sikisaka, M. Marshall, A. K. Feehan, G. Sabino-Santos Jr., A. R. Bell-Kareem, L. D. Hughes, M. Alkuzweny, P. Snarski, J. Garcia-Diaz, R. S. Scott, L. I. Melnik, R. Klitting, M. McGraw, P. Belda-Ferre, P. DeHoff, S. Sathe, C. Marotz, N. D. Grubaugh, D. J. Nolan, A. C. Drouin, K. J. Genemaras, K. Chao, S. Topol, E. Spencer, L. Nicholson, S. Aigner, G. W. Yeo, L. Farnaes, C. A. Hobbs, L. C. Laurent, R. Knight, E. B. Hodcroft, K. Khan, D. N. Fusco, V. S. Cooper, P. Lemey, L. Gardner, S. L. Lamers, J. P. Kamil, R. F. Garry, M. A. Suchard, K. G. Andersen, Emergence of an early SARS-CoV-2 epidemic in the United States. *Cell* **184**, 4939–4952.e15 (2021). doi:10.1016/j.cell.2021.07.030 Medline
29. C. Alteri, V. Cento, A. Piralla, V. Costabile, M. Tallarita, L. Colagrossi, S. Renica, F. Giardina, F. Novazzi, S. Gaiarsa, E. Matarazzo, M. Antonello, C. Vismara, R. Fumagalli, O. M. Epis, M. Puoti, C. F. Perno, F. Baldanti, Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat. Commun.* **12**, 434 (2021). doi:10.1038/s41467-020-20688-x Medline
30. L. du Plessis, O. Pybus, Further musings on the tMRCA. *Virological* (2020); https://virological.org/t/further-musings-on-the-tmrca/340.
31. J. Giesecke, Primary and index cases. *Lancet* **384**, 2024 (2014). doi:10.1016/S0140-6736(14)62331-X Medline
32. Centers for Disease Control and Prevention (CDC), Prevalence of IgG antibody to SARS-associated coronavirus in animal traders—Guangdong Province, China, 2003. *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003). Medline
33. A. Marí Saéz, S. Weiss, K. Nowak, V. Lapeyre, F. Zimmermann, A. Düx, H. S. Kühl, M. Kaba, S. Regnaut, K. Merkel, A. Sachse, U. Thiesen, L. Villányi, C. Boesch, P. W. Dabrowski, A. Radonić, A. Nitsche, S. A. J. Leendertz, S. Petterson, S. Becker, V. Krähling, E. Couacy-Hymann, C. Akoua-Koffi, N. Weber, L. Schaade, J. Fahr, M. Borchert, J. F. Gogarten, S. Calvignac-Spencer, F. H. Leendertz, Investigating the zoonotic origin of the West African Ebola epidemic. *EMBO Mol. Med.* **7**, 17–23 (2015). doi:10.15252/emmm.201404792 Medline
34. WHO Headquarters, WHO-convened global study of origins of SARS-CoV-2: China Part (2021); https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

35. X. Zhang, Y. Tan, Y. Ling, G. Lu, F. Liu, Z. Yi, X. Jia, M. Wu, B. Shi, S. Xu, J. Chen, W. Wang, B. Chen, L. Jiang, S. Yu, J. Lu, J. Wang, M. Xu, Z. Yuan, Q. Zhang, X. Zhang, G. Zhao, S. Wang, S. Chen, H. Lu, Viral and host factors related to the clinical outcome of COVID-19. *Nature* **583**, 437–440 (2020). doi:10.1038/s41586-020-2355-0 Medline
36. E. O. Nsoesie, B. Rader, Y. L. Barnoon, L. Goodwin, J. Brownstein, Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019. *Dig. Acc. Scholar. Harv.* **2**, 019 (2020).
37. L. Chang, L. Zhao, Y. Xiao, T. Xu, L. Chen, Y. Cai, X. Dong, C. Wang, X. Xiao, L. Ren, L. Wang, Serosurvey for SARS-CoV-2 among blood donors in Wuhan, China from September to December 2019. *Protein Cell* 10.1093/procel/pwac013 (2022).
38. E. C. Holmes, S. A. Goldstein, A. L. Rasmussen, D. L. Robertson, A. Crits-Christoph, J. O. Wertheim, S. J. Anthony, W. S. Barclay, M. F. Boni, P. C. Doherty, J. Farrar, J. L. Geoghegan, X. Jiang, J. L. Leibowitz, S. J. D. Neil, T. Skern, S. R. Weiss, M. Worobey, K. G. Andersen, R. F. Garry, A. Rambaut, The origins of SARS-CoV-2: A critical review. *Cell* **184**, 4848–4856 (2021). doi:10.1016/j.cell.2021.08.017 Medline
39. M. Worobey, J. I. Levy, L. M. Malpica Serrano, A. Crits-Christoph, J. E. Pekar, S. A. Goldstein, A. L. Rasmussen, M. U. G. Kraemer, C. Newman, M. P. G. Koopmans, M. A. Suchard, J. O. Wertheim, P. Lemey, D. L. Robertson, R. F. Garry, E. C. Holmes, A. Rambaut, K. G. Andersen, The Huanan market was the epicenter of SARS-CoV-2 emergence. Zenodo (2022); https://zenodo.org/record/6299116.
40. X. Xiao, C. Newman, C. D. Buesching, D. W. Macdonald, Z.-M. Zhou, Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. *Sci. Rep.* **11**, 11898 (2021). doi:10.1038/s41598-021-91470-2 Medline
41. C. M. Freuling, A. Breithaupt, T. Müller, J. Sehl, A. Balkema-Buschmann, M. Rissmann, A. Klein, C. Wylezich, D. Höper, K. Wernike, A. Aebischer, D. Hoffmann, V. Friedrichs, A. Dorhoi, M. H. Groschup, M. Beer, T. C. Mettenleiter, Susceptibility of Raccoon Dogs for Experimental SARS-CoV-2 Infection. *Emerg. Infect. Dis.* **26**, 2982–2985 (2020). doi:10.3201/eid2612.203733 Medline
42. S. M. Porter, A. E. Hartwig, H. Bielefeldt-Ohmann, A. M. Bosco-Lauth, J. Root, Susceptibility of wild canids to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). bioRxiv 478082 [Preprint] (2022). https://doi.org/10.1101/2022.01.27.478082.
43. G. Gao, W. Liu, P. Liu, W. Lei, Z. Jia, X. He, L.-L. Liu, W. Shi, Y. Tan, S. Zou, X. Zhao, G. Wong, J. Wang, F. Wang, G. Wang, K. Qin, R. Gao, J. Zhang, M. Li, W. Xiao, Y. Guo, Z. Xu, Y. Zhao, J. Song, J. Zhang, W. Zhen, W. Zhou, B. Ye, J. Song, M. Yang, W. Zhou, Y. Bi, K. Cai, D. Wang, W. Tan, J. Han, W. Xu, G. Wu, Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market. Research Square (2022). https://doi.org/10.21203/rs.3.rs-1370392/v1.
44. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O'Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus; COVID-19 Genomics UK (COG-UK) Consortium, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021). doi:10.1126/science.abf2946 Medline
45. Chinese SARS Molecular Epidemiology Consortium, Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666–1669 (2004). doi:10.1126/science.1092002 Medline
46. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, MERS-CoV spillover at the camel-human interface. *eLife* **7**, e31257 (2018). doi:10.7554/eLife.31257 Medline
47. J. A. Lednicky, M. S. Tagliamonte, S. K. White, M. A. Elbadry, M. M. Alam, C. J. Stephenson, T. S. Bonny, J. C. Loeb, T. Telisma, S. Chavannes, D. A. Ostrov, C. Mavian, V. M. Beau De Rochars, M. Salemi, J. G. Morris Jr., Independent infections of porcine deltacoronavirus among Haitian children. *Nature* **600**, 133–137 (2021). doi:10.1038/s41586-021-04111-z Medline
48. B. Kan, M. Wang, H. Jing, H. Xu, X. Jiang, M. Yan, W. Liang, H. Zheng, K. Wan, Q. Liu, B. Cui, Y. Xu, E. Zhang, H. Wang, J. Ye, G. Li, M. Li, Z. Cui, X. Qi, K. Chen, L. Du, K. Gao, Y.-T. Zhao, X.-Z. Zou, Y.-J. Feng, Y.-F. Gao, R. Hai, D. Yu, Y. Guan, J. Xu, Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J. Virol.* **79**, 11892–11900 (2005). doi:10.1128/JVI.79.18.11892-11900.2005 Medline
49. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020). doi:10.1038/s41591-020-0820-9 Medline
50. V. L. Hale, P. M. Dennis, D. S. McBride, J. M. Nolting, C. Madden, D. Huey, M. Ehrlich, J. Grieser, J. Winston, D. Lombardi, S. Gibson, L. Saif, M. L. Killian, K. Lantz, R. M. Tell, M. Torchetti, S. Robbe-Austerman, M. I. Nelson, S. A. Faith, A. S. Bowman, SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature* **602**, 481–486 (2022). doi:10.1038/s41586-021-04353-x Medline
51. J. C. Chandler, S. N. Bevins, J. W. Ellis, T. J. Linder, R. M. Tell, M. Jenkins-Moore, J. J. Root, J. B. Lenoch, S. Robbe-Austerman, T. J. DeLiberto, T. Gidlewski, M. Kim Torchetti, S. A. Shriner, SARS-CoV-2 exposure in wild white-tailed deer (*Odocoileus virginianus*). *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2114828118 (2021). doi:10.1073/pnas.2114828118 Medline
52. L. Lu, R. S. Sikkema, F. C. Velkers, D. F. Nieuwenhuijse, E. A. J. Fischer, P. A. Meijer, N. Bouwmeester-Vincken, A. Rietveld, M. C. A. Wegdam-Blans, P. Tolsma, M. Koppelman, L. A. M. Smit, R. W. Hakze-van der Honing, W. H. M. van der Poel, A. N. van der Spek, M. A. H. Spierenburg, R. J. Molenaar, J. Rond, M. Augustijn, M. Woolhouse, J. A. Stegeman, S. Lycett, B. B. Oude Munnink, M. P. G. Koopmans, Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. *Nat. Commun.* **12**, 6802 (2021). doi:10.1038/s41467-021-27096-9 Medline
53. B. B. Oude Munnink, R. S. Sikkema, D. F. Nieuwenhuijse, R. J. Molenaar, E. Munger, R. Molenkamp, A. van der Spek, P. Tolsma, A. Rietveld, M. Brouwer, N. Bouwmeester-Vincken, F. Harders, R. Hakze-van der Honing, M. C. A. Wegdam-Blans, R. J. Bouwstra, C. GeurtsvanKessel, A. A. van der Eijk, F. C. Velkers, L. A. M. Smit, A. Stegeman, W. H. M. van der Poel, M. P. G. Koopmans, Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **371**, 172–177 (2021). doi:10.1126/science.abe5901 Medline
54. S. V. Kuchipudi, M. Surendran-Nair, R. M. Ruden, M. Yon, R. H. Nissly, K. J. Vandegrift, R. K. Nelli, L. Li, B. M. Jayarao, C. D. Maranas, N. Levine, K. Willgert, A. J. K. Conlan, R. J. Olsen, J. J. Davis, J. M. Musser, P. J. Hudson, V. Kapur, Multiple spillovers from humans and onward transmission of SARS-CoV-2 in white-tailed deer. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2121644119 (2022). doi:10.1073/pnas.2121644119 Medline
55. H.-L. Yen, T. H. C. Sit, C. J. Brackman, S. S. Y. Chuk, S. M. S. Cheng, H. Gu, L. D. J. Chang, P. Krishnan, D. Y. M. Ng, G. Y. Z. Liu, M. M. Y. Hui, S. Y. Ho, K. W. S. Tam, P. Y. T. Law, W. Su, S. F. Sia, K.-T. Choy, S. S. Y. Cheuk, S. P. N. Lau, A. W. Y. Tang, J. C. T. Koo, L. Yung, G. Leung, J. S. M. Peiris, L. L. M. Poon, Transmission of SARS-CoV-2 delta variant (AY.127) from pet hamsters to humans, leading to onward human-to-human transmission: A case study. *Lancet* **399**, 1070–1078 (2022). doi:10.1016/S0140-6736(22)00326-9 Medline
56. H.-L. Yen, T. H. C. Sit, C. J. Brackman, S. S. Y. Chuk, H. Gu, K. W. S. Tam, P. Y. T. Law, G. M. Leung, M. Peiris, L. L. M. Poon, S. M. S. Cheng, L. D. J. Chang, P. Krishnan, D. Y. M. Ng, G. Y. Z. Liu, M. M. Y. Hui, S. Y. Ho, W. Su, S. F. Sia, K.-T. Choy, S. S. Y. Cheuk, S. P. N. Lau, A. W. Y. Tang, J. C. T. Koo, L. Yung; HKU-SPH study team, Transmission of SARS-CoV-2 delta variant (AY.127) from pet hamsters to humans, leading to onward human-to-human transmission: A case study. *Lancet* **399**, 1070–1078 (2022). doi:10.1016/S0140-6736(22)00326-9 Medline
57. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). doi:10.2807/1560-7917.ES.2017.22.13.30494 Medline
58. 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
59. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, N. Goldman, Masking strategies for SARS-CoV-2 alignments. *Virological* (2020); https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480.
60. 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
61. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018). doi:10.1093/ve/vex042 Medline

62. 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
63. N. Moshiri, *FAVITES-COVID-Lite: A simplified (and much faster) simulation pipeline specifically for COVID-19 contact + transmission + phylogeny + sequence simulation* (Github, 2022); https://github.com/niemasd/FAVITES-COVID-Lite.
64. X. Hao, S. Cheng, D. Wu, T. Wu, X. Lin, C. Wang, Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* **584**, 420–424 (2020). doi:10.1038/s41586-020-2554-8 Medline
65. J. E. Pekar, A. Rambaut, sars-cov-2-origins/multi-introduction: v1.0.0. Zenodo (2022); doi:10.5281/zenodo.6585475.
66. J. E. Pekar, J. O. Wertheim, Data 1 for: The molecular epidemiology of multiple zoonotic transmissions of SARS-CoV-2. Zenodo (2022); 10.5281/zenodo.6887187.
67. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, R. A. Neher, Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018). doi:10.1093/bioinformatics/bty407 Medline
68. A. Rambaut, *figtree* (Github, 2018); https://github.com/rambaut/figtree/releases.
69. H. Li, Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). doi:10.1093/bioinformatics/bty191 Medline
70. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline
71. N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F. Michael, L. L. Coffey, N. J. Loman, K. G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019). doi:10.1186/s13059-018-1618-7 Medline
72. *gofasta* (Github, 2022); https://github.com/virus-evolution/gofasta.
73. G. Dudas, *baltic: baltic - backronymed adaptable lightweight tree import code for molecular phylogeny manipulation, analysis and visualisation* (Github, 2021); https://github.com/evogytis/baltic.
74. S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, S. D. W. Frost, GARD: A genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096–3098 (2006). doi:10.1093/bioinformatics/btl474 Medline
75. D. P. Martin, B. Murrell, M. Golden, A. Khoosal, B. Muhire, RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015). doi:10.1093/ve/vev003 Medline
76. H. M. Lam, O. Ratmann, M. F. Boni, Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Mol. Biol. Evol.* **35**, 247–251 (2018). doi:10.1093/molbev/msx263 Medline
77. M. F. Boni, P. Lemey, X. Jiang, T. T.-Y. Lam, B. W. Perry, T. A. Castoe, A. Rambaut, D. L. Robertson, Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020). doi:10.1038/s41564-020-0771-4 Medline
78. 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
79. 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
80. F. Li, Y.-Y. Li, M.-J. Liu, L.-Q. Fang, N. E. Dean, G. W. K. Wong, X.-B. Yang, I. Longini, M. E. Halloran, H.-J. Wang, P.-L. Liu, Y.-H. Pang, Y.-Q. Yan, S. Liu, W. Xia, X.-X. Lu, Q. Liu, Y. Yang, S.-Q. Xu, Household transmission of SARS-CoV-2 and risk factors for susceptibility and infectivity in Wuhan: A retrospective observational study. *Lancet Infect. Dis.* **21**, 617–628 (2021). doi:10.1016/S1473-3099(20)30981-6 Medline
81. *EpiNow2: Estimate Realtime Case Counts and Time-varying Epidemiological Parameters* (Github, 2020); https://github.com/epiforecasts/EpiNow2.
82. N. Moshiri, NiemaGraphGen: A memory-efficient global-scale contact network simulation toolkit. *GIGAbyte* 10.46471/gigabyte.37 (2022).
83. A. L. Barabasi, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). doi:10.1126/science.286.5439.509 Medline
84. S. Eubank, H. Guclu, V. S. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184 (2004). doi:10.1038/nature02541 Medline
85. 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
86. F. D. Sahneh, A. Vajdi, H. Shakeri, F. Fan, C. Scoglio, GEMFsim: A stochastic simulator for the generalized epidemic modeling framework. *J. Comput. Sci.* **22**, 36–44 (2017). doi:10.1016/j.jocs.2017.08.014
87. X. Yang, Y. Yu, J. Xu, H. Shu, J. Xia, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, Y. Wang, S. Pan, X. Zou, S. Yuan, Y. Shang, Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: A single-centered, retrospective, observational study. *Lancet Respir. Med.* **8**, 475–481 (2020). doi:10.1016/S2213-2600(20)30079-5 Medline
88. F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu, Y. Zhang, H. Chen, B. Cao, Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **395**, 1054–1062 (2020). doi:10.1016/S0140-6736(20)30566-3 Medline
89. J. Yang, X. Chen, X. Deng, Z. Chen, H. Gong, H. Yan, Q. Wu, H. Shi, S. Lai, M. Ajelli, C. Viboud, P. H. Yu, Disease burden and clinical severity of the first pandemic wave of COVID-19 in Wuhan, China. *Nat. Commun.* **11**, 5411 (2020). doi:10.1038/s41467-020-19238-2 Medline
90. N. Moshiri, TreeSwift: A massively scalable Python tree package. *SoftwareX* **11**, 100436 (2020). doi:10.1016/j.softx.2020.100436
91. J. Ma, First Chinese coronavirus cases may have been infected in October 2019, says new research. *South China Morning Post* (2021); https://www.scmp.com/news/china/science/article/3126499/first-chinese-covid-19-cases-may-have-been-infected-october-2019.
92. K. Andersen, Clock and TMRCA based on 27 genomes. *Virological* (2020); https://virological.org/t/clock-and-tmrca-based-on-27-genomes/347/6.
93. L. Pipes, H. Wang, J. P. Huelsenbeck, R. Nielsen, Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. *Mol. Biol. Evol.* **38**, 1537–1543 (2021). doi:10.1093/molbev/msaa316 Medline
94. T. Murata, A. Sakurai, M. Suzuki, S. Komoto, T. Ide, T. Ishihara, Y. Doi, Shedding of Viable Virus in Asymptomatic SARS-CoV-2 Carriers. *MSphere* **6**, e00019-21 (2021). doi:10.1128/mSphere.00019-21 Medline
95. T. Sekizuka, K. Itokawa, T. Kageyama, S. Saito, I. Takayama, H. Asanuma, N. Nao, R. Tanaka, M. Hashino, T. Takahashi, H. Kamiya, T. Yamagishi, K. Kakimoto, M. Suzuki, H. Hasegawa, T. Wakita, M. Kuroda, Haplotype networks of SARS-CoV-2 infections in the *Diamond Princess* cruise ship outbreak. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 20198–20201 (2020). doi:10.1073/pnas.2006824117 Medline
96. Y. Turakhia, B. Thornlow, A. S. Hinrichs, N. De Maio, L. Gozashti, R. Lanfear, D. Haussler, R. Corbett-Detig, Ultrafast Sample placement on Existing tRees (USHER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021). doi:10.1038/s41588-021-00862-7 Medline
97. 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
98. M. Ghafari, L. du Plessis, J. Raghwani, S. Bhatt, B. Xu, O. G. Pybus, A. Katzourakis, Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. *Mol. Biol. Evol.* **39**, msac009 (2022). doi:10.1093/molbev/msac009 Medline
99. S. Duchêne, E. C. Holmes, S. Y. W. Ho, Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. Biol. Sci.* **281**, 20140732 (2014). doi:10.1098/rspb.2014.0732 Medline

100. J. Dushoff, S. W. Park, Speed and strength of an epidemic intervention. *Proc. Biol. Sci.* **288**, 20201556 (2021). doi:10.1098/rspb.2020.1556 Medline
101. J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, G. M. Leung, Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat. Med.* **26**, 506–510 (2020). doi:10.1038/s41591-020-0822-7 Medline
102. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020). doi:10.1016/S0140-6736(20)30183-5 Medline
103. R. Ke, E. Romero-Severson, S. Sanche, N. Hengartner, Estimating the reproductive number R_0 of SARS-CoV-2 in the United States and eight European countries and implications for vaccination. *J. Theor. Biol.* **517**, 110621 (2021). doi:10.1016/j.jtbi.2021.110621 Medline
104. L. Pellis, F. Scarabel, H. B. Stage, C. E. Overton, L. H. K. Chappell, E. Fearon, E. Bennett, K. A. Lythgoe, T. A. House, I. Hall; University of Manchester COVID-19 Modelling Group, Challenges in control of COVID-19: Short doubling time and long delay to effect of interventions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **376**, 20200264 (2021). doi:10.1098/rstb.2020.0264 Medline
105. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, Z. Feng, Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi:10.1056/NEJMoa2001316 Medline
106. M. Chinazzi, J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore Y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini Jr., A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020). doi:10.1126/science.aba9757 Medline
107. R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**, 489–493 (2020). doi:10.1126/science.abb3221 Medline
108. N. Moshiri, CoaTran: Coalescent tree simulation along a transmission network. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.11.10.377499.
109. K. M. Braun, G. K. Moreno, C. Wagner, M. A. Accola, W. M. Rehrauer, D. A. Baker, K. Koelle, D. H. O'Connor, T. Bedford, T. C. Friedrich, L. H. Moncla, Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLOS Pathog.* **17**, e1009849 (2021). doi:10.1371/journal.ppat.1009849 Medline
110. J. Ma, Coronavirus: China's first confirmed Covid-19 case traced back to November 17. *South China Morning Post* (2020); https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back.

ACKNOWLEDGMENTS

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared via GISAID the viral genomic sequences and metadata on which this research is based (data S1) (*57*). We are greatly appreciative toward Lu Chen, Di Liu, and Yi Yan for providing insight into the putative intermediate genomes and clarification regarding the relative sequencing depth at positions 8782 and 28144, Marc Eloit and Sarah Temmam for sharing their sarbecovirus dataset and recombination analysis results, and Matthew Kuehnert for general feedback. Figure S30 was created with Biorender.com. **Funding:** This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services, under Contract No. 75N93021C00015 (MW). JEP acknowledges support from the NIH (T15LM011271). NM acknowledges support from the National Science Foundation (NSF) (NSF-2028040). JIL acknowledges support from the NIH (5T32AI007244-38). JOW acknowledges support from the NIH (R01AI135992 and R01AI136056). RFG is supported by the NIH (R01AI132223, R01AI132244, U19AI142790, U54CA260581, U54HG007480, OT2HL158260), the Coalition for Epidemic Preparedness Innovation, the Wellcome Trust Foundation, Gilead Sciences, and the European and Developing Countries Clinical Trials Partnership Programme. MAS and AR acknowledge the support of the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network), the European Research Council (grant agreement no. 725422 – ReservoirDOCS) and the NIH (R01AI153044). KGA is supported by the NIH (U19AI135995, U01AI151812, and UL1TR002550). ECH is funded by an Australian Research Council Laureate Fellowship (FL170100022). JL, HP, and MSP acknowledge support from the National Research Foundation of Korea, funded by the Ministry of Science and Information and Communication Technologies, Republic of Korea (NRF-2017M3A9E4061995 and NRF-2019R1A2C2084206). TIV acknowledges support from the Branco Weiss Fellowship. We thank AMD for the donation of critical hardware and support resources from its HPC Fund that made this work possible. This work was supported (in part) by the Epidemiology and Laboratory Capacity (ELC) for Infectious Diseases Cooperative Agreement (Grant Number: ELC DETECT (6NU50CK000517-01-07) funded by the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or the Department of Health and Human Services. **Author contributions:** Conceptualization: JEP, MAS, KGA, MW, JOW; Methodology: JEP, AM, NM, MAS, KGA, MW, JOW; Software: JEP, AM, NM, KG, MAS; Validation: JEP, AM, KI, KG, MAS; Formal analysis: JEP, AM, EP, KI, JLH, KG, JOW; Investigation: JEP, AM, EP, KI, JLH, KG, JOW; Resources: MAS, KGA, JOW; Data Curation: JEP, EP, KG, MZ, JCW, SH, JL, HP, MP, KCZY, RTPL, MNMI, YMN, JOW; Writing - original draft preparation: JEP, MW, JOW; Writing - review and editing: All Authors; Visualization: JEP, JLH, KG, LMMS; Supervision: MAS, KGA, MW, JOW; Project administration: MAS, KGA, MW, JOW; Funding acquisition: MAS, KGA, MW, JOW. **Competing interests:** JOW has received funding from the CDC (ongoing) via contracts or agreements to his institution unrelated to this research. MAS receives contracts and grants from the US Food and Drug Administration, the US Department of Veterans Affairs and Janssen Research and Development unrelated to this research. RFG is co-founder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. MW, ECH, AR, MAS, JOW, and KGA have received consulting fees and/or provided compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic. **Data and materials availability:** Genome accessions are available in data S1 and S2, and raw data for two genomes were deposited to NCBI SRA (PRJNA806767 and PRJNA802993). Code is available on Zenodo (*65*). The following data are available on Data Dryad (*66*): recCA sequence, BEAST phylogenetic inference output, and simulation and rejection sampling output for the primary analysis.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abp8337
Materials and Methods
Supplementary Text
Figs. S1 to S31
Tables S1 to S15
References (*67*–*110*)
MDAR Reproducibility Checklist
Data S1 to S3

Submitted 3 March 2022; accepted 18 July 2022
Published online 26 July 2022
10.1126/science.abp8337

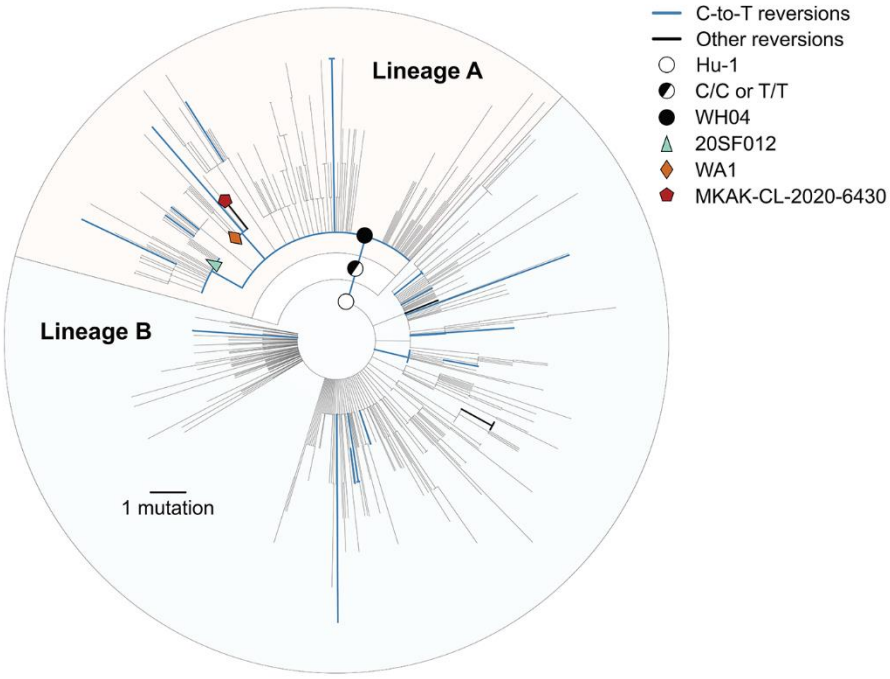


Fig. 1. Maximum likelihood phylogeny of the early SARS-CoV-2 pandemic, showing nucleotide reversions and putative candidates for the ancestral haplotype at the most common recent ancestor (MRCA). Putative ancestral haplotypes are identified with colored shapes. Reversions from the Hu-1 reference genotype to the recCA are colored. Blue represents C-to-T reversions and black indicates all other reversions. The tree is rooted on Hu-1 to show reversion dynamics to the recCA.

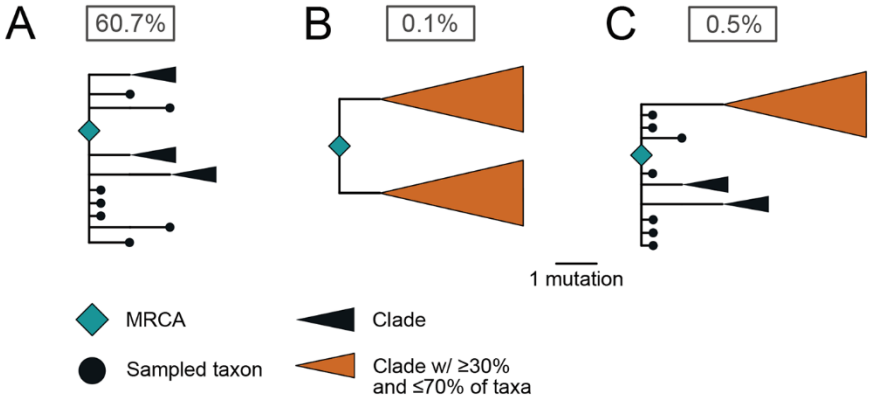


Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations. (**A**) A large polytomy of at least 20 descendent lineages, consistent with the base of both lineages A and B. (**B**) Topology matching a C/C ancestral haplotype: two clades each one mutation from the ancestor, both with polytomies of at least 20 descendent lineages. (**C**) Topology matching either a lineage A or lineage B ancestral haplotype: a basal polytomy with at least 20 descendent lineages including a large clade separated by two mutations, also possessing a polytomy of at least 20 descendent lineages. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the box.

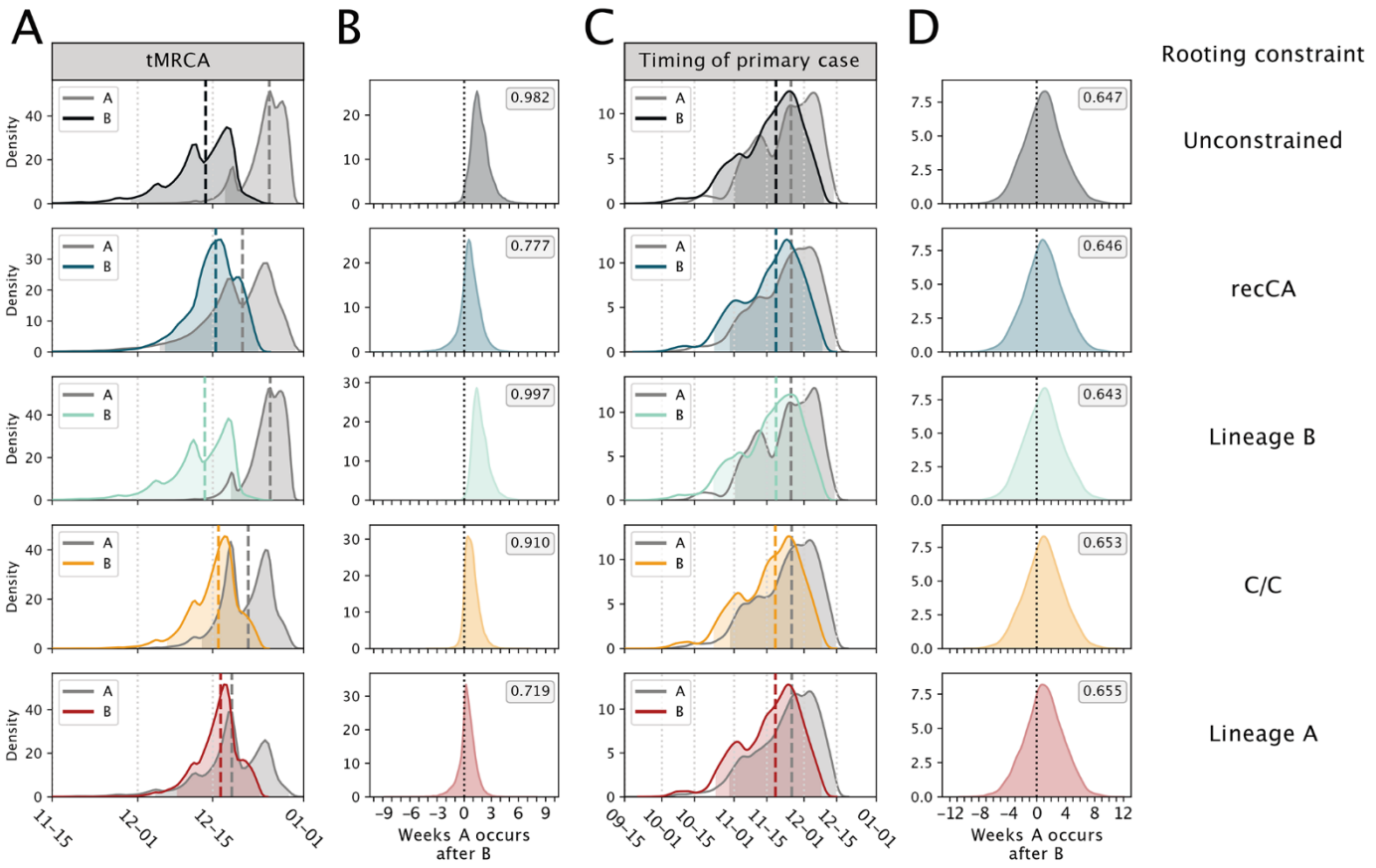


Fig. 3. Comparison of the tMRCA and primary case dates for lineage A and lineage B across rooting strategies. Each row represents a different rooting constraint in phylodynamic analysis, with lineage B, C/C and lineage A representing a fixed ancestral haplotype. (**A**) The tMRCA for lineages A and B. (**B**) The number of weeks the tMRCA of lineage A occurs after the tMRCA of lineage B. (**C**) The timing of the primary case for lineages A and B. (**D**) The number of weeks the time of the primary case of lineage A occurs after the time of the primary case of lineage B. Long dashed lines indicate the median and shading represents the 95% HPD for each distribution. Short dashed lines indicate 0 weeks difference between lineages A and B. Posterior probability that lineage A originated after lineage B is reported in the grey box.

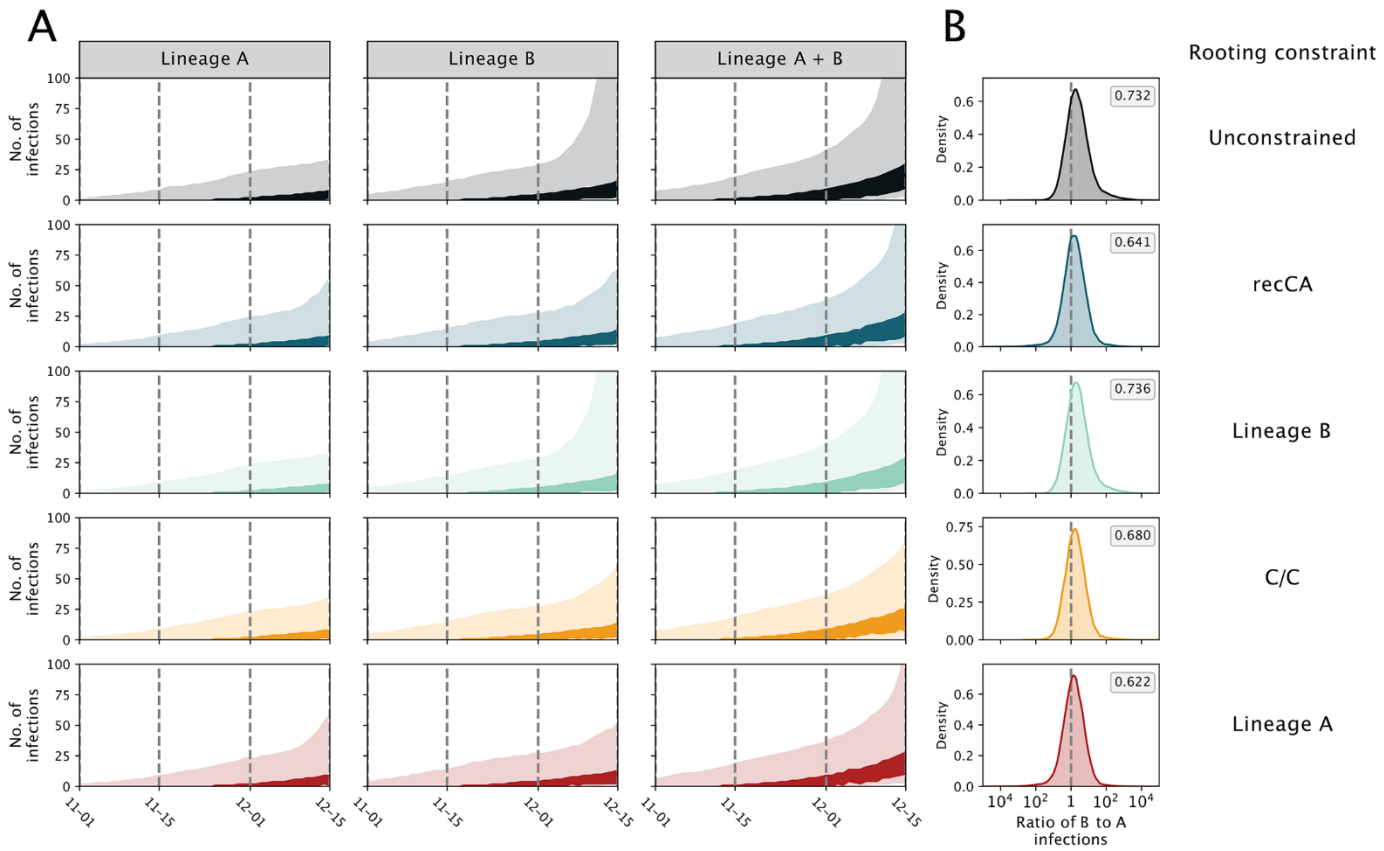


Fig. 4. Dynamics of simulated SARS-CoV-2 epidemics resulting from separate introductions of lineages A and B. Each row represents a different rooting constraint in phylodynamic analysis, with lineage B, C/C and lineage A representing a fixed ancestral haplotype. (**A**) Estimated number of infections. The header of each column indicates whether the number of infections are caused by lineage A, lineage B, or the two lineages combined. Darker and lighter shading represent the 50% and 95% HPD, respectively. (**B**) The log ratio of lineage B to lineage A infections on 15 December 2019. Posterior probability of having more lineage B infections than lineage A reported in the grey box.

Table 1. Posterior probabilities of inferred ancestral haplotype at the MRCA of SARS-CoV-2. Positions 8782 and 28144 are indicated in parentheses. Representative genome is that with its sequence matching the haplotype. "No market" excludes 15 market-associated genomes (13 lineage B genomes associated with the Huanan market plus one lineage A and one lineage B genome not associated with the Huanan market). *BF > 10. **BF > 100. ***BF > 1000; BFs are in favor of hypothesis rejection.

Haplotype	Mutations from Hu-1 reference	Representative genome	Phylodynamic analysis		
			Unconstrained (%)	No market (%)	recCA (%)
B (C/T)	N/A	Hu-1	80.85†	62.96†	8.18
A (T/C)	C8782T+T28144C	WH04	1.68*	5.73*	77.28†
C/C	T28144C	N/A	10.32	23.02	10.49
T/T	C8782T	N/A	0.92*	1.68*	3.71*
A+C29025T (T/C)	C8782T+T28144C+C29095T	20SF012	<0.01***	<0.01***	0.20**
A.1 (T/C)	C8782T+T28144C+C18060T	WA1	<0.01***	<0.01***	0.04***

†Haplotype with greatest posterior probability; reference for BF.

The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2

Jonathan E. PekarAndrew MageeEdyth ParkerNiema MoshiriKatherine IzhikevichJennifer L. HavensKarthik GangavarapuLorena Mariana Malpica SerranoAlexander Crits-ChristophNathaniel L. MattesonMark ZellerJoshua I. LevyJade C. WangScott HughesJungmin LeeHeedo ParkMan-Seong ParkKatherine Zi Yan ChingRaymond Tzer Pin LinMohd Noor Mat IsaYusuf Muhammad NoorTetyana I. VasylyevaRobert F. GarryEdward C. HolmesAndrew RambautMarc A. SuchardKristian G. AndersenMichael WorobeyJoel O. Wertheim

Science, **Ahead of Print** • DOI: 10.1126/science.abp8337

The Washington Post *Democracy Dies in Darkness*

Prominent scientist who said lab-leak theory of covid-19 origin should be probed now believes evidence points to Wuhan market

Controversy continues amid sketchy data and lack of transparency from Chinese authorities

By Joel Achenbach

Yesterday at 2:00 p.m. EST

The location of early coronavirus infections in late 2019 in Wuhan, China, suggests the virus probably spread to humans from a market where wild and domestically farmed animals were sold and butchered, according to a peer-reviewed article published Thursday in the journal Science that is the latest salvo in the debate over how the pandemic began.

The article, by University of Arizona evolutionary virologist Michael Worobey — a specialist in the origins of viral epidemics — does not purport to answer all questions about the pandemic's origins, nor is it likely to quell speculation that the virus might have emerged somehow from risky laboratory research.

Worobey has been open to the theory of a lab leak. He was one of the 18 scientists who wrote a much-publicized letter to Science in May calling for an investigation of all possible sources of the virus, including a laboratory accident. But he now contends that the geographic pattern of early cases strongly supports the hypothesis that the virus came from an infected animal at the Huanan Seafood Market — an argument that will probably revive the broader debate about the virus's origins.

Worobey notes that more than half of the earliest documented illnesses from the virus were among people with a direct connection to the market, and he argues this was not merely the result of the early focus on the market as a potential source of the outbreak. He concludes that the first patient known to fall ill with the virus was a female seafood vendor at the market who became symptomatic on Dec. 11, 2019.

That contradicts a report earlier this year from investigators for the World Health Organization and China, who concluded that the first patient was a 41-year-old accountant with no connection to the market who became sick on Dec. 8. But Worobey said the accountant's medical records reveal he visited the dentist that day to deal with retained baby teeth that needed to be pulled, but did not show symptoms from the coronavirus until Dec. 16, and was hospitalized six days after that.

The stealthy nature of the virus, which can spread asymptomatically, makes it highly likely that the pathogen began to spread many weeks before any of the cases that were identified. But Worobey said the locations and occupations of the first known patients point to a market origin, with the virus radiating outward into the city of 11 million.

"It becomes almost impossible to explain that pattern if that epidemic didn't start there," Worobey said in an interview.

Geography has been central to theories about the origin of the virus. Wuhan is home to the Wuhan Institute of Virology, where researchers study and conduct experiments upon coronaviruses that circulate abundantly in bats in

central and southern China. The institute has been a focus of those who argue that an accidental leak from one of its research labs is the most likely explanation for the spillover of the virus into humans.

The Huanan Seafood Market is many miles, and across the Yangtze River, from the virology institute. Few of the early documented cases were anywhere near the laboratory. A second laboratory studying coronaviruses at the Wuhan CDC, which oversaw the city's coronavirus response, relocated in late 2019 to a spot close to the market.

Worobey's article immediately drew skeptical responses from two prominent scientists who, like Worobey, have been deeply engaged in the debate over the most likely scenario for the start of the pandemic.

"It is based on fragmentary information and to a large degree, hearsay," David A. Relman, a professor of microbiology at Stanford University, said in an email after reading an embargoed copy. "In general, there is no way of verifying much of what he describes, and then concludes."

Jesse Bloom, a computational biologist at the Fred Hutchinson Cancer Research Center in Seattle, said the quality of the data from China on early coronavirus infections is too poor to support any conclusion.

"I don't feel like anything can be concluded with high or even really modest confidence about the exact origin of SARS-CoV-2 in Wuhan, simply because the underlying data are so limited," Bloom said. He contends that genetic evidence from early virus samples points to the market as a superspreader event, but not as the location of the first set of infections.

Bloom has been among those sounding alarms about what he feels is overly risky research conducted at the Wuhan Institute of Virology. That research has generated tremendous controversy, with some Republican lawmakers and conservative media figures focusing on funding for some of the experiments, funneled via a nonprofit group, EcoHealth Alliance, from the National Institute of Allergy and Infectious Diseases, which is led by President Biden's chief pandemic medical adviser, Anthony S. Fauci.

Worobey's paper drew strong praise from those favoring the natural zoonosis theory.

"Mike's piece shows beyond a shadow of a doubt that in fact the Huanan market was the epicenter of the outbreak," said Robert F. Garry Jr., a virologist at Tulane University and one of the most vocal proponents of the zoonosis hypothesis.

Benjamin Neuman, a virologist at Texas A&M University who was one of the coronavirus experts to give SARS-CoV-2 its name in early 2020, called the report "detailed and compelling, in a way that the most detailed conspiracy timelines have not been. … When the evidence is laid out like this, the association with the market is strong long before anyone realized it — right from the start."

Worobey and critics Relman and Bloom have one thing in common: They signed the letter to the journal Science in May that called for continued investigation into the virus's origins, including the possibility of a lab leak.

Soon, public opinion polls showed more people favored the lab-leak theory than the market origin. And Biden ordered his intelligence agencies to look into the matter and report back within 90 days.

In the months since he signed the Science letter, Worobey has become more convinced that the pandemic began as a spillover in the market, where animals known to be capable of harboring the virus — such as raccoon dogs — were sold.

The Science letter was influential in taking conjecture that had once been derided as a conspiracy theory and propelling it into the mainstream of virus-origin debates, even making it, as Worobey puts it, "the leading contender" in the public mind for the origin of the pandemic.

"The pendulum has swung way too far to the other side," he said.

It has been known since the start of the pandemic that the Huanan market was linked to many early cases, and the

It has been known since the start of the pandemic that the Huanan market was linked to many early cases, and the first news reports invariably cited it as the likely source of viral spillover. But the joint report from the WHO and China this year presented a murkier picture, noting that some cases in December 2019 had no link to the market: "No firm conclusion therefore about the role of the Huanan market in the origin of the outbreak, or how the infection was introduced into the market, can currently be drawn."

The market was quickly closed, the animals culled before any were screened for SARS-CoV-2, and everything cleaned and sanitized soon after the outbreak began. Still, a subsequent investigation showed that traces of the virus were found on surfaces in the market, including drains, particularly in the area where vendors sold animals.

Worobey acknowledged that the clustering of infections could be misleading, saying the early focus on the market might have skewed data because epidemiologists might have looked for market-linked infections and missed infections occurring in areas getting less attention — a common tendency in research known as "ascertainment bias." But he concluded that the timeline and geography of early cases rule out such an error.

Chinese officials have said the Huanan market was not the source of the pandemic. China's government has pushed the idea that the coronavirus could have been brought to China from overseas, including from Fort Detrick in Maryland and through frozen food imports.

Worobey does not contend that he has proved definitively how the pandemic began. And his article is not a research study presenting all-new data, but rather is labeled a "Perspective" piece. Such articles typically aggregate and interpret information that for the most part has already been in the public domain.

Although the lab-leak idea was at first derided by many scientists and in the mainstream media as a conspiracy theory — one embraced by President Donald Trump and his allies as part of their rhetorical attacks on China and the "China virus" — the failure to find an animal host of the immediate precursor to SARS-CoV-2 has kept all hypotheses on the table.

The 90-day investigation conducted by U.S. intelligence agencies at the behest of Biden was inconclusive. Most agencies favored the natural zoonosis theory. One favored the lab leak. The only firm conclusion was that the virus was not a bioweapon.

Worobey said he was open to the possibility of a lab leak, simply because of the proximity of the Wuhan Institute of Virology to the first outbreak. But he examined the geography question more closely. If the virus came out of the lab, why did the first cases cluster in and around the market many miles away? And that market, he notes, had sold animals that were implicated in the first SARS epidemic of 2002-2003.

"It becomes almost absurd, in my mind, to imagine that this virus started at the Wuhan Institute of Virology, and almost immediately that person went to one of the few places that sold raccoon dogs and other animals that were implicated in SARS-1," he said.

His paper does not mention the Wuhan CDC laboratory. Chinese officials have insisted that SARS-CoV-2 was never in one of the country's laboratories, nor has it been found through tests in wild or domesticated animals.

Proponents of the lab-leak theory point to the lack of transparency of Chinese officials and the removal of experimental data from a database at the Wuhan Institute of Virology several months before the pandemic. Worobey's market-origin theory suggests an alternative scenario, one in which authorities were not eager to find proof that the spillover happened in a market with live animals that may have been illegally captured and sold.

Worobey also suggests that Chinese officials may have been embarrassed that the country's system for identifying and rapidly responding to novel pneumonia-like illnesses — a system put in place after the original SARS epidemic — was slow to detect the outbreak of illnesses caused by the novel coronavirus.

Eva Dou contributed to this report.

Cite as: M. Worobey, *Science* 10.1126/science.abm4454 (2021).

Dissecting the early COVID-19 cases in Wuhan

By **Michael Worobey**

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. Email: worobey@arizona.edu

Elucidating the origin of the pandemic requires understanding of the Wuhan outbreak

Some key questions lie at the heart of investigations into the origin of the COVID-19 pandemic, including what is known about the earliest COVID-19 cases in Wuhan, China, and what can be learned from them? Despite assertions to the contrary (*1*), it is now clear that live mammals susceptible to coronaviruses, including raccoon dogs (*Nyctereutes procyonoides*), were sold at Huanan Market and three other live-animal markets in Wuhan before the pandemic (*2*, *3*). Severe acute respiratory syndrome–related coronaviruses (SARSr-CoVs) were found in raccoon dogs during the SARS outbreak, which was facilitated by animal-to-human contact in live-animal markets in China. However, because of the early public health focus on Huanan Market, it remains unclear whether the apparent preponderance of hospitalized COVID-19 cases associated with this market was truly reflective of the initial outbreak. Answering these questions requires resolving several crucial events that took place in December 2019 and early January 2020.

On 30 December 2019, the Wuhan Municipal Health Commission (WHC) issued two emergency notices for internal circulation to local hospitals alerting them to patients with unexplained pneumonia—several of whom worked at Huanan Market—and laying out a treatment and response plan (see fig. S1). The first official public report was WHC's announcement the next day that they had carried out case searches and retrospective investigations related to Huanan Market and found 27 patients. Forty-one of the first known patients formed the basis of an influential study that reported that 66%—i.e., not all early cases—had a link to Huanan Market (*4*). They had been transferred between 29 December and 2 January from other hospitals to Jinyintan Hospital, Wuhan's premier infectious disease center. Notably, individuals were enrolled according to clinical presentation, not epidemiologic information, such as connections to Huanan Market (*4*).

China's Viral Pneumonia of Unknown Etiology (VPUE) mechanism was set up in the wake of SARS to be an early warning reporting system for detecting unknown viral diseases and is overseen by the China Center for Disease Control and Prevention (CCDC) (*5*). PUE cases are supposed to be rapidly reported by clinicians to the national notifiable disease reporting system through an internet-based platform. Evidently, that did not happen in Wuhan in December. The system appears to have been in active use only from 3 January. Although it favored cases having a connection to Huanan Market (*6*–*8*), the VPUE mechanism could not have improperly inflated the proportion of Huanan Market–linked cases in December (*1*). Moreover, reporting began only after the 41 patients were transferred from other hospitals to Jinyintan Hospital. Nevertheless, it is possible that a disproportionate number of cases linked to Huanan Market were transferred to Jinyintan Hospital because of public health officials' early focus there.

There is, however, a way to step back to a period before any such bias could have crept in, by considering what happened in the hospitals that first pieced together that a new viral outbreak was underway. Although not mentioned by name in scientific publications (*9*), media reports reveal that Hubei Provincial Hospital of Integrated Chinese and Western Medicine (HPHICWM) was the first hospital to alert district, municipal, and provincial public health authorities about the mysterious pneumonia cases (see fig. S1). Zhang Jixian, director of respiratory and critical care medicine, noticed on 27 December that an elderly couple had large "ground glass" opacities in computed tomography (CT) images of their lungs, distinct from those she had seen in other cases of viral pneumonia. Zhang insisted that the couple's son, who was not a patient and had no symptoms, undergo a CT scan, and the same unusual lesions were observed. The husband and wife evidently are "cluster 1" in the World Health Organization (WHO)–China report (*1*): They are the earliest known case cluster and the only cluster admitted by 26 December. They had no known connection to Huanan Market.

Another patient with similar CT imaging, a worker at Huanan Market, was admitted on 27 December. Zhang, concerned about a new, probably infectious viral disease, reported the four cases to hospital officials, who alerted the Jianghan District CDC that same day. Over 28 and 29 December, three more patients, all of whom worked at Huanan Market, were admitted and recognized to have the same unknown respiratory disease. A vice president of HPHICWM, Xia Wenguang, brought together 10 experts from the hospital, including Zhang, for an emergency meeting on 29 December, and they concluded that the situation was extraordinary. Upon learning of similar patients, also linked to Huanan Market, at Tongji and Union (Xiehe) Hospitals, Xia alerted

the Wuhan and Hubei CDCs on 29 December.

A notably similar situation unfolded at Wuhan Central Hospital. On 18 December, Ai Fen, director of the emergency department, encountered her first unexplained pneumonia patient, a 65-year-old man who had become ill on either 13 or 15 December. Unbeknownst to Ai at the time, the patient was a deliveryman at Huanan Market. A CT scan revealed infection in both lungs, and he did not respond to antibiotics or anti-influenza drugs. On 24 December, a bronchoalveolar lavage specimen collected from him was sent to Vision Medicals, a metagenomics sequencing company. They identified a new SARSr-CoV on 26 December and relayed the finding by telephone to the hospital on 27 December. By 28 December, Wuhan Central Hospital had identified seven cases, of which four turned out to be linked to Huanan Market. Notably, these seven cases, like those at HPHICWM, were ascertained before epidemiologic investigations concerning Huanan Market commenced on 29 December.

At Zhongnan Hospital in the Wuchang District of Wuhan, 15 km away from Huanan Market and on the opposite bank of the Yangtze River, Vice President Yuan Yufeng asked units on 31 December to search for unexplained pneumonia cases, and the Respiratory Medicine Department reported two. The first lived in Wuchang District but worked at Huanan Market (in Jianghan District). The second did not work at Huanan Market but had friends who did and who had visited his home. On 3 January, three more cases were identified—a family cluster unlinked to Huanan Market. Clearly, hospitals in the first weeks of the outbreak were identifying cases both with and without a known connection to Huanan Market. And Wuhan hospitals were not swamped with unexplained pneumonia cases at the end of December—that would come later.

Thus, 10 of these hospitals' 19 earliest COVID-19 cases were linked to Huanan Market (~53%), comparable both to Jinyintan's 66% (of 41 cases) (*4*) and to the WHO-China report's 33% of 168 retrospectively identified cases across December 2019 (*1*). Regarding cases at the Wuhan Central Hospital and HPHICWM, patients with a history of exposure at Huanan Market could not have been "cherry picked" before anyone had identified the market as an epidemiologic risk factor. Hence, there was a genuine preponderance of early COVID-19 cases associated with Huanan Market.

How can this knowledge inform our understanding of the pandemic? If Huanan Market was the source, why were only one- to two-thirds of early cases linked to the market? Perhaps a better question is why would one expect all cases ascertained weeks into the outbreak to be confined to one market? Given the high transmissibility of SARS-CoV-2 and the high rate of asymptomatic spread, many symptomatic cases would inevitably soon lack a direct link to the location of the pandemic's origin. And some cases counted as "unlinked" may have been only one or two transmissions away, as exemplified by the second patient identified at Zhongnan Hospital. That so many of the >100 COVID-19 cases from December (*1*) with no identified epidemiologic link to Huanan Market nonetheless lived in its direct vicinity is notable (see the figure) and provides compelling evidence that community transmission started at the market.

Additionally, the earliest known cases should not necessarily be expected to be the first infected or linked to Huanan Market: They probably postdated the outbreak's index case by a considerable period (*10*) because only ~7% of SARS-CoV-2 infections lead to hospitalization (*11*); most fly under the radar. Similarly, it is entirely expected that early, ascertained cases from a seafood market would be workers who were not necessarily directly associated with wildlife sales because the outbreak spread from human to human. The index case was most likely one of the ~93% who never required hospitalization and indeed could have been any of hundreds of workers who had even brief contact with infected live mammals.

Crucially, however, the now famous "earliest" COVID-19 case (*1*), a 41-year-old male accountant, who lived 30 km south of Huanan Market and had no connection to it—illness onset reported as 8 December—appears to have become ill with COVID-19 considerably later (*12*). When interviewed, he reported that his COVID-19 symptoms started with a fever on 16 December; the 8 December illness was a dental problem related to baby teeth retained into adulthood (*12*). This is corroborated by hospital records and a scientific paper that reports his COVID-19 onset date as 16 December and date of hospitalization as 22 December (*13*). This indicates that he was infected through community transmission after the virus had begun spreading from Huanan Market. He believed that he may have been infected in a hospital (presumably during his dental emergency) or on the subway during his commute; he had also traveled north of Huanan Market shortly before his symptoms began (*12*). His symptom onset came after multiple cases in workers at Huanan Market, making a female seafood vendor there the earliest known case, with illness onset 11 December (*12*). Notably, she reported knowledge of several possible COVID-19 cases in clinics and hospitals that were near Huanan Market from 11 December, and Huanan Market patients were hospitalized at Union Hospital as early as 10 December (see fig. S1).

Although a widely cited report (*7*) credits the VPUE mechanism with uncovering the pandemic, it was HPHICWM that identified both the outbreak and the Huanan Market connection and passed on these fully formed discoveries to district, municipal, and provincial public health officials by 29 December (*9*). National officials reportedly did not learn about the outbreak until CCDC Director George Gao encountered online group chats about the WHC emergency notices on the evening of 30 December. Concerned that so many cases had

not been reported to the VPUE system, he quickly notified the National Health Commission (*14*) (see fig. S1).

Therefore, the preponderance of early cases connected to Huanan Market could not have been an artifact of ascertainment bias introduced by case definitions in the VPUE system. Although mechanisms like China's VPUE system are potentially invaluable, they will fail without both widespread buy-in from health care providers and rapid data sharing from local to central authorities. Key problems with the VPUE system were known before the pandemic, including that most clinicians in China had little awareness of the VPUE system and were not reporting cases to it—for example, 0 of 335 PUE cases in one study from 2019 (*5*). China should be commended, however, for having such a system, which is lacking in most countries. The focus now should be on fixing the problems that COVID-19 has exposed and blanketing the globe with a highly functional PUE early warning system.

Samples from the earliest COVID-19 patients in Wuhan have been sequenced, and two distinct SARS-CoV-2 lineages, A and B, have been identified. Given that the elderly couple at HPHICWM was the WHO report's cluster 1, it follows that the husband, illness onset 26 December (*1*), must be the source of the earliest lineage A sequence, Wuhan/IME-WH01/2019 (GenBank accession number MT291826) (see fig. S1), which he most likely got from his wife, who became ill 15 December. This raises the possibility that the Yangchahu market that they visited may have been a site of a separate animal spillover. The recent discovery that there may be no true lineage A or B intermediates in humans (*15*) also raises the possibility of separate spillovers of both lineages. However, the earliest known lineage A genomes have close geographical connections to Huanan Market: one from a patient (age and gender not reported) who stayed in a hotel near Huanan Market in the days before illness onset in December (*13*) and the other from the 62-year-old husband in cluster 1 who visited Yangchahu Market, just a few blocks north of Huanan Market (*1*), and lived just to the south (see the figure). Therefore, if lineage A had a separate animal origin from lineage B, both most likely occurred at Huanan Market, and the association with Yangchahu Market, which does not appear to have sold live mammals, is likely due to community transmission starting in the neighborhoods surrounding Huanan Market.

With SARS, live-animal markets continued to sell infected animals for many months, allowing zoonotic spillover to be established as the origin and revealing multiple independent jumps from animals into humans (*3*). Unfortunately, no live mammal collected at Huanan Market or any other live-animal market in Wuhan has been screened for SARS-CoV-2–related viruses (*1*), and Huanan Market was closed and disinfected on 1 January 2020. Nevertheless, that most early symptomatic cases were linked to Huanan Market—specifically to the western section (*1*) where raccoon dogs were caged (*2*)—provides strong evidence of a live-animal market origin of the pandemic.

This would explain the extraordinary preponderance of early COVID-19 cases at one of the handful of sites in Wuhan—population 11 million—that sell some of the same animals that brought us SARS. Although it may never be possible to recover related viruses from animals if they were not sampled at the time of emergence, conclusive evidence of a Huanan Market origin from infected wildlife may nonetheless be obtainable through analysis of spatial patterns of early cases and from additional genomic data, including SARS-CoV-2–positive samples from Huanan Market, as well as through integration of additional epidemiologic data. Preventing future pandemics depends on this effort.

REFERENCES AND NOTES

1. WHO, WHO-convened global study of origins of SARS-CoV-2: China Part (2021); https://bit.ly/3wjUXze.
2. X. Xiao, C. Newman, C. D. Buesching, D. W. Macdonald, Z.-M. Zhou, Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. *Sci. Rep.* **11**, 11898 (2021). doi:10.1038/s41598-021-91470-2 Medline
3. E. C. Holmes, S. A. Goldstein, A. L. Rasmussen, D. L. Robertson, A. Crits-Christoph, J. O. Wertheim, S. J. Anthony, W. S. Barclay, M. F. Boni, P. C. Doherty, J. Farrar, J. L. Geoghegan, X. Jiang, J. L. Leibowitz, S. J. D. Neil, T. Skern, S. R. Weiss, M. Worobey, K. G. Andersen, R. F. Garry, A. Rambaut, The origins of SARS-CoV-2: A critical review. *Cell* **184**, 4848–4856 (2021). doi:10.1016/j.cell.2021.08.017 Medline
4. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020). doi:10.1016/S0140-6736(20)30183-5 Medline
5. N. Xiang, Y. Song, Y. Wang, J. Wu, A. J. Millman, C. M. Greene, Z. Ding, J. Sun, W. Yang, G. Guo, R. Wang, P. Guo, Z. Ren, L. Gong, P. Xu, S. Zhou, D. Lin, D. Ni, Z. Feng, Q. Li, Lessons from an active surveillance pilot to assess the pneumonia of unknown etiology surveillance system in China, 2016: The need to increase clinician participation in the detection and reporting of emerging respiratory infectious diseases. *BMC Infect. Dis.* **19**, 770 (2019). doi:10.1186/s12879-019-4345-0 Medline
6. T. K. Tsang, P. Wu, Y. Lin, E. H. Y. Lau, G. M. Leung, B. J. Cowling, Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: A modelling study. *Lancet Public Health* **5**, e289–e296 (2020). doi:10.1016/S2468-2667(20)30089-X Medline
7. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, Z. Feng, Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi:10.1056/NEJMoa2001316 Medline
8. D. L. Yang, "Wuhan officials tried to cover up covid-19—and sent it careening outward," *Washington Post*, 10 March 2020.
9. The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team, *China CDC Weekly* **2**, 79 (2020).
10. J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, J. O. Wertheim, Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**, 412–417 (2021). doi:10.1126/science.abf8003 Medline
11. S. Mahajan, C. Caraballo, S.-X. Li, Y. Dong, L. Chen, S. K. Huston, R. Srinivasan, C. A. Redlich, A. I. Ko, J. S. Faust, H. P. Forman, H. M. Krumholz, SARS-CoV-2 infection hospitalization rate and infection fatality rate among the non-congregate population in Connecticut. *Am. J. Med.* **134**, 812–816.e2 (2021).

doi:10.1016/j.amjmed.2021.01.020 Medline
12. L. Bao, "Looking for the first infected person in the South China Seafood Market," video, *The Paper*, 25 March 2020; https://bit.ly/2YikwEa.
13. R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020). doi:10.1016/S0140-6736(20)30251-8 Medline
14. D. L. Yang, "China's early warning system didn't work on covid-19. Here's the story," *Washington Post*, 24 February 2020.
15. J. Pekar, E. Parker, J. L. Havens, M. A. Suchard, K. G. Andersen, N. Moshiri, M. Worobey, A. Rambaut, J. O. Wertheim, Virological 754 (2021); https://virological.org/t/evidence-against-the-veracity-of-sars-cov-2-genomes-intermediate-between-lineages-a-and-b/754.
Acknowledgments: Thanks to four anonymous reviewers and to A. Crits-Christoph, E. Holmes, D. Robertson, J. Wertheim, J. Pekar, K. Andersen, S. Goldstein, A. Rambaut, H. Mourant, D. Yang, L. Wang, S. Chen, C. Di, and Q. Jiang for assistance and discussions. The author is supported by the David and Lucile Packard Foundation and the Bill and Melinda Gates Foundation.

ACKNOWLEDGMENTS

Thanks to four anonymous reviewers and to A. Crits-Christoph, E. Holmes, D. Robertson, J. Wertheim, J. Pekar, K. Andersen, S. Goldstein, A. Rambaut, H. Mourant, D. Yang, L. Wang, S. Chen, C. Di, and Q. Jiang for assistance and discussions. The author is supported by the David and Lucile Packard Foundation and the Bill and Melinda Gates Foundation.

SUPPLEMENTARY MATERIALS

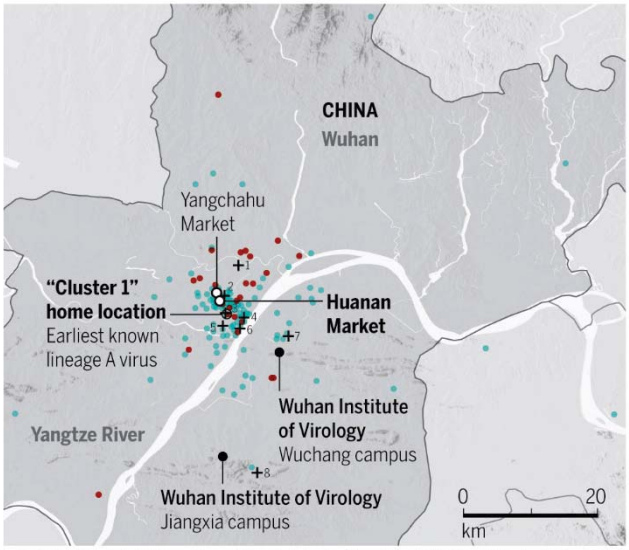
science.org/doi/10.1126/science.abm4454

Published online 18 November 2021
10.1126/science.abm4454

COVID-19 cases in Wuhan in December 2019

The map shows that most of the earliest cases of COVID-19 were in close proximity to Huanan Market, even if they were not directly connected with the market through working there or visiting. This suggests that transmission in the community around the market was occurring in December 2019. The map is based on a subset of data from 174 COVID-19 cases in and around Wuhan (*1*).

● Home address of cases with epidemiological link to Huanan Market
● No identified link to Huanan Market ○ Market + Hospital



1, Jinyintan Hospital; 2, Wuhan Central Hospital, Houhu Branch (no. 2); 3, Hubei Provincial Hospital of Integrated Chinese and Western Medicine; 4, Wuhan Central Hospital, Nanjing Road Branch; 5, Tongji Hospital; 6, Union Hospital; 7, Zhongnan Hospital; 8, Wuhan Jiangxia First People's Hospital

GRAPHIC: K. FRANKLIN/*SCIENCE*

The New York Times https://www.nytimes.com/2021/06/23/science/coronavirus-sequences.html

Scientist Finds Early Virus Sequences That Had Been Mysteriously Deleted

By rooting through files stored on Google Cloud, a researcher says he recovered 13 early coronavirus sequences that had disappeared from a database last year.

By Carl Zimmer

June 23, 2021

About a year ago, genetic sequences from more than 200 virus samples from early cases of Covid-19 in Wuhan disappeared from an online scientific database.

Now, by rooting through files stored on Google Cloud, a researcher in Seattle reports that he has recovered 13 of those original sequences — intriguing new information for discerning when and how the virus may have spilled over from a bat or another animal into humans.

The new analysis, released on Tuesday, bolsters earlier suggestions that a variety of coronaviruses may have been circulating in Wuhan before the initial outbreaks linked to animal and seafood markets in December 2019.

As the Biden administration investigates the contested origins of the virus, known as SARS-CoV-2, the study neither strengthens nor discounts the hypothesis that the pathogen leaked out of a famous Wuhan lab. But it does raise questions about why original sequences were deleted, and suggests that there may be more revelations to recover from the far corners of the internet.

“This is a great piece of sleuth work for sure, and it significantly advances efforts to understand the origin of SARS-CoV-2,” said Michael Worobey, an evolutionary biologist at the University of Arizona who was not involved in the study.

Jesse Bloom, a virologist at the Fred Hutchinson Cancer Research Center who wrote the new report, called the deletion of these sequences suspicious. It “seems likely that the sequences were deleted to obscure their existence,” he wrote in the paper, which has not yet been peer-reviewed or published in a scientific journal.

Dr. Bloom and Dr. Worobey belong to an outspoken group of scientists who have called for more research into how the pandemic began. In a letter published in May, they complained that there wasn’t enough information to determine whether it was more likely that a lab leak spread the coronavirus, or that it leapt to humans from contact with an infected animal outside of a lab.

The genetic sequences of viral samples hold crucial clues about how SARS-CoV-2 shifted to our species from another animal, most likely a bat. Most precious of all are sequences from early in the pandemic, because they take scientists closer to the original spillover event.

As Dr. Bloom was reviewing what genetic data had been published by various research groups, he came across a March 2020 study with a spreadsheet that included information on 241 genetic sequences collected by scientists at Wuhan University. The spreadsheet indicated that the scientists had uploaded the sequences to an online database called the Sequence Read Archive, managed by the U.S. government’s National Library of Medicine.

But when Dr. Bloom looked for the Wuhan sequences in the database earlier this month, his only result was “no item found.”

Puzzled, he went back to the spreadsheet for any further clues. It indicated that the 241 sequences had been collected by a scientist named Aisi Fu at Renmin Hospital in Wuhan. Searching medical literature, Dr. Bloom eventually found another study posted online in March 2020 by Dr. Fu and colleagues, describing a new experimental test for SARS-CoV-2. The Chinese scientists published it in a scientific journal three months later.

In that study, the scientists wrote that they had looked at 45 samples from nasal swabs taken “from outpatients with suspected Covid-19 early in the epidemic.” They then searched for a portion of SARS-CoV-2’s genetic material in the swabs. The researchers did not publish the actual sequences of the genes they fished out of the samples. Instead, they only published some mutations in the viruses.

But a number of clues indicated to Dr. Bloom that the samples were the source of the 241 missing sequences. The papers included no explanation as to why the sequences had been uploaded to the Sequence Read Archive, only to disappear later.

Perusing the archive, Dr. Bloom figured out that many of the sequences were stored as files on Google Cloud. Each sequence was contained in a file in the cloud, and the names of the files all shared the same basic format, he reported.

Dr. Bloom swapped in the code for a missing sequence from Wuhan. Suddenly, he had the sequence. All told, he managed to recover 13 sequences from the cloud this way.

With this new data, Dr. Bloom looked back once more at the early stages of the pandemic. He combined the 13 sequences with other published sequences of early coronaviruses, hoping to make progress on building the family tree of SARS-CoV-2.

Working out all the steps by which SARS-CoV-2 evolved from a bat virus has been a challenge because scientists still have a limited number of samples to study. Some of the earliest samples come from the Huanan Seafood Wholesale Market in Wuhan, where an outbreak occurred in December 2019.

But those market viruses actually have three extra mutations that are missing from SARS-CoV-2 samples collected weeks later. In other words, those later viruses look more like coronaviruses found in bats, supporting the idea that there was some early lineage of the virus that did not pass through the seafood market.

Dr. Bloom found that the deleted sequences he recovered from the cloud also lack those extra mutations. “They’re three steps more similar to the bat coronaviruses than the viruses from the Huanan fish market,” Dr. Bloom said.

The Wuhan Huanan Wholesale Seafood Market in January 2020. Dake Kang/Associated Press

This suggests, he said, that by the time SARS-CoV-2 reached the market, it had been circulating for awhile in Wuhan or beyond. The market viruses, he argued, aren't representative of full diversity of coronaviruses already loose in late 2019.

"Maybe our picture of what was present early in Wuhan from what has been sequenced might be somewhat biased," he said.

In his report, Dr. Bloom acknowledged that this conclusion would have to be confirmed with a deeper analysis of the virus sequences. Dr. Worobey said that he and his colleagues are working on a large-scale study of SARS-CoV-2 genes to better understand its origin and that they'll now add Dr. Bloom's 13 recovered sequences.

"These additional data will play a big role in that effort," Dr. Worobey said.

It's not clear why this valuable information went missing in the first place. Scientists can request that files be deleted by sending an email to the managers of the Sequence Read Archive. The National Library of Medicine, which manages the archive, said that the 13 sequences were removed last summer.

"These SARS-CoV-2 sequences were submitted for posting in SRA in March 2020 and subsequently requested to be withdrawn by the submitting investigator in June 2020," said Renate Myles, a spokeswoman for the National Institutes of Health.

She said that the investigator, whom she did not name, told the archive managers that the sequences were being updated and would be added to a different database. But Dr. Bloom has searched every database he knows of, and has yet to find them. "Obviously I can't rule out that the sequences are on some other database or web page somewhere, but I have not been able to find them any of the obvious places I've looked," he said.

Three of the co-authors of the 2020 testing study that produced the 13 sequences did not immediately respond to emails inquiring about Dr. Bloom's finding. That study did not give contact information for another co-author, Dr. Fu, who was also named on the spreadsheet from the other study.

Some scientists are skeptical that there is anything sinister behind the removal of the sequences. "I don't really understand how this points to a cover-up," said Stephen Goldstein, a virologist at the University of Utah.

Dr. Goldstein noted that the testing paper listed the individual mutations the Wuhan researchers found in their tests. Although the full sequences are no longer in the archive, the key information has been public for over a year, he said. It was just tucked away in a format that is hard for researchers to find.

"We all missed this relatively obscure paper," Dr. Goldstein said.

"You can't really say why they were removed," Dr. Bloom acknowledged in an interview. "You can say that the practical consequence of removing them was that people didn't notice they existed." He also noted that the Chinese government ordered the destruction of a number of early samples of the virus and barred the publication of papers on the coronavirus without its approval.

For his part, Dr. Worobey still wants answers. "I hope we hear from the authors who generated, but then deleted, these crucial sequences so we can understand more about their motivation for doing so," he said. "It certainly is strange at face value and really demands an explanation."

Regardless of what happened to these 13 sequences, Dr. Bloom now wonders what other clues might be discovered online. In order to reconstruct the origin of Covid-19, all those clues potentially matter.

"Ideally, we need to try to find as many other early sequences as possible," he said. "And I think this study suggests that we should look everywhere."

Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic

Jesse D. Bloom
Fred Hutchinson Cancer Research Center
Howard Hughes Medical Institute
Seattle, WA, USA

ABSTRACT The origin and early spread of SARS-CoV-2 remains shrouded in mystery. Here I identify a data set containing SARS-CoV-2 sequences from early in the Wuhan epidemic that has been deleted from the NIH's Sequence Read Archive. I recover the deleted files from the Google Cloud, and reconstruct partial sequences of 13 early epidemic viruses. Phylogenetic analysis of these sequences in the context of carefully annotated existing data suggests that the Huanan Seafood Market sequences that are the focus of the joint WHO-China report are not fully representative of the viruses in Wuhan early in the epidemic. Instead, the progenitor of known SARS-CoV-2 sequences likely contained three mutations relative to the market viruses that made it more similar to SARS-CoV-2's bat coronavirus relatives.

Understanding the spread of SARS-CoV-2 in Wuhan is crucial to tracing the origins of the virus, including identifying events that led to infection of patient zero. The first reports outside of China at the end of December 2019 emphasized the role of the Huanan Seafood Market (ProMED 2019), which was initially suggested as a site of zoonosis. However, this theory became increasingly tenuous as it was learned that many early cases had no connection to the market (Cohen 2020; Huang *et al.* 2020; Chen *et al.* 2020). Eventually, Chinese CDC Director Gao Fu dismissed the theory, stating "At first, we assumed the seafood market might have the virus, but now the market is more like a victim. The novel coronavirus had existed long before" (Global Times 2020).

Indeed, there were reports of cases that far preceded the outbreak at the Huanan Seafood Market. The *Lancet* described a confirmed case having no association with the market whose symptoms began on December 1, 2019 (Huang *et al.* 2020). The *South China Morning Post* described nine cases from November 2019 including details on patient age and sex, noting that none were confirmed to be "patient zero" (Ma 2020). Professor Yu Chuanhua of Wuhan University told the *Health Times* that records he reviewed showed two cases in mid-November, and one suspected case on September 29 (Health Times 2020). At about the same time as Professor Chuanhua's interview, the Chinese CDC issued an order forbidding sharing of information about the COVID-19 epidemic without approval (China CDC 2020), and shortly thereafter Professor Chuanhua re-contacted the *Health Times* to say the November cases could not be confirmed (Health Times 2020). Then China's State Council issued a much broader order requiring central approval of all publications related to COVID-19 to ensure they were coordinated "like moves in a game of chess" (Kang *et al.* 2020a). In 2021, the joint WHO-China report dismissed all reported cases prior to December 8 as not COVID-19, and revived the theory that the virus might have originated at the Huanan Seafood Market (WHO 2021).

In other outbreaks where direct identification of early cases has been stymied, it has increasingly become possible to use genomic epidemiology to infer the timing and dynamics of spread from analysis of viral sequences. For instance, analysis of SARS-CoV-2 sequences has enabled reconstruction of the initial spread of SARS-CoV-2 in North America and Europe (Bedford *et al.* 2020; Worobey *et al.* 2020; Deng *et al.* 2020; Fauver *et al.* 2020).

But in the case of Wuhan, genomic epidemiology has also proven frustratingly inconclusive. Some of the problem is simply limited data: despite the fact that Wuhan has advanced virology labs, there is only patchy sampling of SARS-CoV-2 sequences from the first months of the city's explosive outbreak. Other than a set of multiply sequenced samples collected in late December of 2019 from a dozen patients connected to the Huanan Seafood Market (WHO 2021), just a handful of Wuhan sequences are available from before late January of 2020 (see analysis in this study below). This paucity of sequences could be due in part to an order that unauthorized Chinese labs destroy all coronavirus samples from early in the outbreak, reportedly for "laboratory biological safety" reasons (Pingui 2020).

However, the Wuhan sequences that are available have also confounded phylogenetic analyses designed to infer the "progenitor" of SARS-CoV-2, which is the sequence from which all other currently known sequences are descended (Kumar *et al.* 2021). Although there is debate about exactly how SARS-CoV-2 entered the human population, it is universally accepted that the virus's deep ancestors are bat coronaviruses (Lytras *et al.* 2021). But the earliest known SARS-CoV-2 sequences, which are mostly derived from the Huanan Seafood Market, are notably more different from these bat coronaviruses than other sequences collected at later dates outside Wuhan. As a result, there is a direct conflict between the two major principles used to infer an outbreak's progenitor: namely that it should be among the earliest sequences, and that it should be most closely related to deeper ancestors (Pipes *et al.* 2021).

Here I take a step towards resolving these questions by identifying and recovering a deleted data set of partial SARS-CoV-2 sequences from outpatient samples collected early in the Wuhan epidemic. Analysis of these new sequences in conjunction with careful annotation of existing ones suggests that the early Wuhan

Manuscript compiled: Friday 18th June, 2021
Corresponding author: jbloom@fredhutch.org

Figure 1 Accessions from deep sequencing project PRJNA612766 have been removed from the SRA. Shown is the result of searching for "SRR11313485" in the SRA search toolbar. This result has been digitally archived on the Wayback Machine at https://web.archive.org/web/20210502131630/https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR11313485.

samples that have been the focus of most studies including the joint WHO-China report (WHO 2021) are not fully representative of the viruses actually present in Wuhan at that time. These insights help reconcile phylogenetic discrepancies, and suggest two plausible progenitor sequences, one of which is identical to that inferred by Kumar *et al.* (2021). Furthermore, the approach taken here hints it may be possible to advance understanding of SARS-CoV-2's origins or early spread even without further on-the-ground studies, such as by more deeply probing data archived by the NIH and other entities.

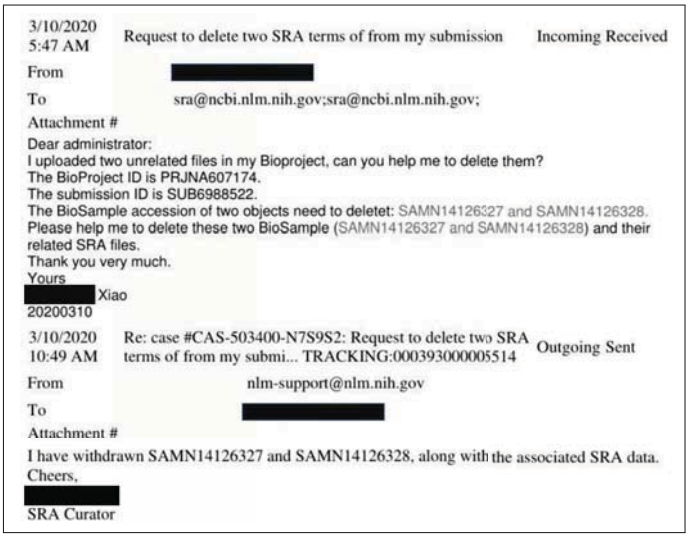
Results

Identification of a SARS-CoV-2 deep sequencing data set that has been removed from the Sequence Read Archive

During the course of my research, I read a paper by Farkas *et al.* (2020) that analyzed SARS-CoV-2 deep sequencing data from the Sequence Read Archive (SRA), which is a repository maintained by the NIH's National Center for Biotechnology Information. The first supplementary table of Farkas *et al.* (2020) lists all SARS-CoV-2 deep sequencing data available from the SRA as of March 30, 2020.

The majority of entries in this table refer to a project (BioProject PRJNA612766) by Wuhan University that is described as nanopore sequencing of SARS-CoV-2 amplicons. The table indicates this project represents 241 of the 282 SARS-CoV-2 sequencing run accessions in the SRA as of March, 30, 2020. Because I had never encountered any other mention of this project, I performed a Google search for "PRJNA612766," and found no search hits other than the supplementary table itself. Searching for "PRJNA612766" in the NCBI's SRA search box returned a message of "No items found." I then searched for individual sequencing run accessions from the project in the NCBI's SRA search box. These searches returned messages indicating that the sequencing runs had been removed (Figure 1).

The SRA is designed as a permanent archive of deep sequencing data. The SRA documentation states that after a sequencing run is uploaded, "neither its files can be replaced nor filenames can be changed," and that data can only be deleted by e-mailing SRA staff (SRA 2021). An example of this process from another study is in Figure 2, which shows an e-mail by the lead author of a paper on pangolin coronaviruses (Xiao *et al.* 2020) requesting deletion of two sequencing runs. Subsequent to March 30, 2020, a similar e-mail request must have been made to fully delete SARS-CoV-2 deep sequencing project PRJNA612766.


3/10/2020 5:47 AM Request to delete two SRA terms of from my submission Incoming Received
From [redacted]
To sra@ncbi.nlm.nih.gov;sra@ncbi.nlm.nih.gov;
Attachment #
Dear administrator:
I uploaded two unrelated files in my Bioproject, can you help me to delete them?
The BioProject ID is PRJNA607174.
The submission ID is SUB6988522.
The BioSample accession of two objects need to deletet: SAMN14126327 and SAMN14126328.
Please help me to delete these two BioSample (SAMN14126327 and SAMN14126328) and their related SRA files.
Thank you very much.
Yours
[redacted] Xiao
20200310

3/10/2020 10:49 AM Re: case #CAS-503400-N7S9S2: Request to delete two SRA terms of from my submi... TRACKING:000393000005514 Outgoing Sent
From nlm-support@nlm.nih.gov
To [redacted]
Attachment #
I have withdrawn SAMN14126327 and SAMN14126328, along with the associated SRA data.
Cheers,
[redacted]
SRA Curator

Figure 2 Example of the process to delete SRA data. The image shows e-mails between the lead author of the pangolin coronavirus paper Xiao *et al.* (2020) and SRA staff excerpted from USRTK (2020).

The deleted data set contains sequencing of viral samples collected early in the Wuhan epidemic

The metadata in the first supplementary table of Farkas *et al.* (2020) indicates that the samples in deleted project PRNJA612766 were collected by Aisu Fu and Renmin Hospital of Wuhan University. Google searching for these terms revealed the samples were related to a study posted as a pre-print on *medRxiv* in early March of 2020 (Wang *et al.* 2020a), and subsequently published in the journal *Small* in June of 2020 (Wang *et al.* 2020b).

The study describes an approach to diagnose infection with SARS-CoV-2 and other respiratory viruses by nanopore sequencing. This approach involved reverse-transcription of total RNA from swab samples, followed by PCR with specific primers to generate amplicons covering portions of the viral genome. These amplicons were then sequenced on an Oxford Nanopore GridION, and infection was diagnosed if the sequencing yielded sufficient reads aligning to the viral genome. Importantly, the study notes that this approach yields information about the sequence of the virus as well enabling diagnosis of infection.

The pre-print (Wang *et al.* 2020a) says the approach was applied to "45 nasopharyngeal swab samples from outpatients with suspected COVID-19 early in the epidemic." The digital object identifier (DOI) for the pre-print indicates that it was processed by *medRxiv* on March 4, 2020, which is one day after China's State Council ordered that all papers related to COVID-19 must be centrally approved (Kang *et al.* 2020a). The final published manuscript (Wang *et al.* 2020b) from June of 2020 updated the description from "early in the epidemic" to "early in the epidemic (January 2020)." Both the pre-print and published manuscript say that 34 of the 45 early epidemic samples were positive in the sequencing-based diagnostic approach. In addition, both state that the approach was later applied to 16 additional samples collected on February 11–12, 2020, from SARS-CoV-2 patients hospitalized at Renmin Hospital of Wuhan University.

There is complete concordance between the accessions for project PRJNA612766 in the supplementary table of Farkas *et al.* (2020) and the samples described by Wang *et al.* (2020a). There are 89 accessions corresponding to the 45 early epidemic sam-

ples, with these samples named like wells in a 96-well plate (A1, A2, etc). The number of accessions is approximately twice the number of early epidemic samples because each sample has data for two sequencing runtimes except one sample (B5) with just one runtime. There are 31 accessions corresponding to the 16 samples collected in February from Renmin Hospital patients, with these samples named R01, R02, etc. Again, all but one sample (R04) have data for two sequencing runtimes. In addition, there are 7 accessions corresponding to positive and negative controls, 2 accessions corresponding to other respiratory virus samples, and 112 samples corresponding to plasmids used for benchmarking of the approach. Together, these samples and controls account for all 241 accessions listed for PRJNA612766 in the supplementary table of Farkas *et al.* (2020).

Neither the pre-print (Wang *et al.* 2020a) nor published manuscript (Wang *et al.* 2020b) contain any correction or note that indicates a scientific reason for deleting the study's sequencing data from the SRA. I e-mailed both corresponding authors of Wang *et al.* (2020a) to ask why they had deleted the deep sequencing data and to request details on the collection dates of the early outpatient samples, but received no reply.

Recovery of deleted sequencing data from the Google Cloud

As indicated in Figure 1, none of the deleted sequencing runs could be accessed through the SRA's web interface. In addition, none of the runs could be accessed using the command-line tools of the SRA Toolkit. For instance, running `fastq-dump SRR11313485` or `vdb-dump SRR11313485` returned the message "err: query unauthorized while resolving query within virtual file system module - failed to resolve accession 'SRR11313485'".

However, the SRA has begun storing all data on the Google and Amazon clouds. While inspecting the SRA's web interface for other sequencing accessions, I noticed that SRA files are often available from links to the cloud such as `https://storage.googleapis.com/nih-sequence-read-archive/run/<ACCESSION>/<ACCESSION>`.

Based on the hypothesis that deletion of sequencing runs by the SRA might not remove files stored on the cloud, I interpolated the cloud URLs for the deleted accessions and tested if they still yielded the SRA files. This strategy was successful; for instance, as of June 3, 2021, going to https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313485/SRR11313485 downloads the SRA file for accession SRR11313485. I have archived this file on the Wayback Machine at https://web.archive.org/web/20210502130820/https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313485/SRR11313485.

I automated this strategy to download the SRA files for 97 of the 99 sequencing runs corresponding to the 34 SARS-CoV-2 positive early epidemic samples and the 16 hospital samples from February (files for SRR11313490 and SRR11313499 were not accessible via the cloud). I used the SRA Toolkit to get the object timestamp (`vdb-dump --obj_timestamp`) and time (`vdb-dump --info`) for each SRA file. For all files, the object timestamp is February 15, 2020, and the time is March 16, 2020. Although the SRA Toolkit does not clearly document these two properties, my *guess* is that the object timestamp may refer to when the SRA file was created from a FASTQ file uploaded to the SRA, and the time may refer to when the accession was made public.

The data are sufficient to determine the viral sequence from the start of spike through the end of ORF10 for some samples

Wang *et al.* (2020a) sequenced PCR amplicons covering nucleotide sites 21,563 to 29,674 of the SARS-CoV-2 genome, which spans from the start of the spike gene to the end of ORF10. They also sequenced a short amplicon generated by nested PCR that covered a fragment of ORF1ab spanning sites ∼15,080 to 15,550. In this paper, I only analyze the region from spike through ORF10 because this is a much longer contiguous sequence and the amplicons were generated by conventional rather than nested PCR. I slightly trimmed the region of interest to 21,570 to 29,550 because many samples had poor coverage at the termini.

I aligned the recovered deep sequencing data to the SARS-CoV-2 genome using `minimap2` (Li 2018), combining accessions

sample	fraction sites called (21570-29550)	patient group	substitutions relative to proCoV2
A4	0.9827	early outpatient	none
C1	0.9966	early outpatient	G22081A (A=924, C=4, G=9), C28144T (C=6, T=1185), T29483G (C=1, G=45, T=1)
C2	0.9962	early outpatient	C29095T (C=1, G=1, T=751)
C9	0.9536	early outpatient	C28144T (C=3, T=823), G28514T (G=1, T=36)
D9	0.9585	early outpatient	C28144T (C=4, T=1653)
D12	0.9970	early outpatient	C28144T (C=8, T=2400)
E1	0.9759	early outpatient	C28144T (T=125)
E5	0.9758	early outpatient	C24034T (A=5, C=3, T=74), T26729C (C=12), G28077C (C=142, G=4)
E11	0.9877	early outpatient	C25460T (C=2, T=246), C28144T (C=1, T=412)
F11	0.9594	early outpatient	T25304A (A=9, T=1), C28144T (C=6, G=1, T=1328)
G1	0.9959	early outpatient	none
G11	0.9677	early outpatient	none
H9	0.9941	early outpatient	C28144T (C=2, T=1254)
R11	0.9987	hospital patient (Feb)	C21707T (T=401), C28144T (A=1, C=18, T=4265)

Table 1 Samples for which the SARS-CoV-2 sequence could be called at ≥95% of sites between 21,570 and 29,550, and the substitutions in this region relative to the putative SARS-CoV-2 progenitor proCoV2 inferred by Kumar *et al.* (2021). Numbers in parentheses after each substitution give the deep sequencing reads with each nucleotide identity.

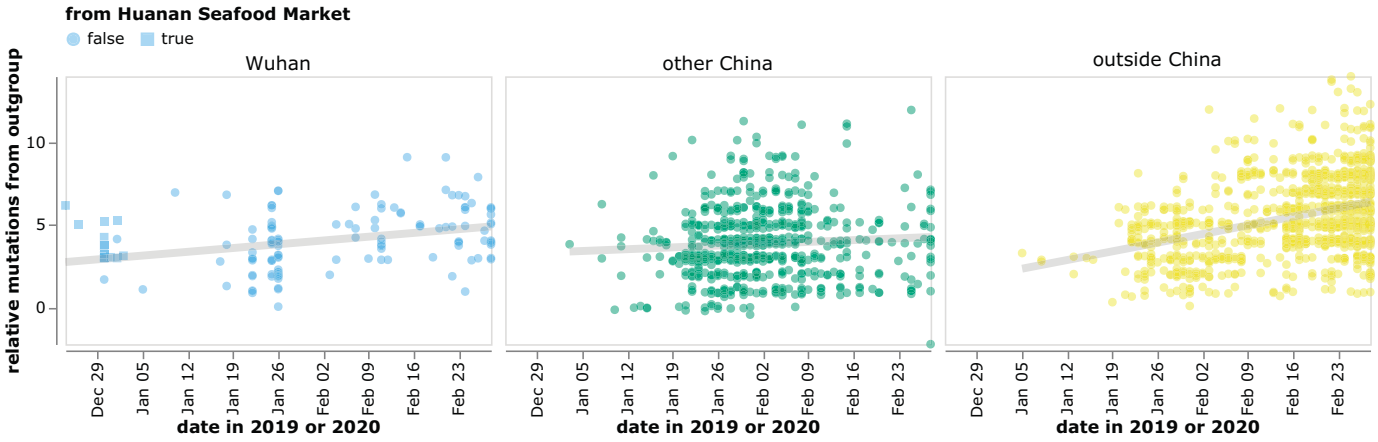


Figure 3 The reported collection dates of SARS-CoV-2 sequences in GISAID versus their relative mutational distances from the RaTG13 bat coronavirus outgroup. Mutational distances are relative to the putative progenitor proCoV2 inferred by Kumar *et al.* (2021). The plot shows sequences in GISAID collected no later than February 28, 2020. Sequences that the joint WHO-China report (WHO 2021) describes as being associated with the Wuhan Seafood Market are plotted with squares. Points are slightly jittered on the y-axis. Go to https://jbloom.github.io/SARS-CoV-2_PRJNA612766/deltadist.html for an interactive version of this plot that enables toggling of the outgroup to RpYN06 and RmYN02, mouseovers to see details for each point including strain name and mutations relative to proCoV2, and adjustment of the y-axis jittering. Static versions of the plot with RpYN06 and RmYN02 outgroups are in Figure S3.

for the same sample. Figure S1 shows the sequencing coverage for the 34 virus-positive early epidemic samples and the 16 hospitalized patient samples over the region of interest; a comparable plot for the whole genome is in Figure S2.

I called the consensus viral sequence for each sample at each site with coverage $\geq$3 and >80% of the reads concurring on the nucleotide identity. With these criteria, 13 of the early outpatient samples and 1 of the February hospitalized patient samples had sufficient coverage to call the consensus sequence at >95% of the sites in the region of interest (Table 1), and for the remainder of this paper I focus on these high-coverage samples. Table 1 also shows the mutations in each sample relative to proCoV2, which is a putative progenitor of SARS-CoV-2 inferred by Kumar *et al.* (2021) that differs from the widely using Wuhan-Hu-1 reference sequence by three mutations (C8782T, C18060T, and T28144C). Although requiring coverage of only $\geq$3 is relatively lenient, Table 1 shows that all sites with mutations have coverage $\geq$10. In addition, the mutations I called from the raw sequence data in Table 1 concord with those mentioned in Wang *et al.* (2020b).

I also determined the consensus sequence of the plasmid control used by Wang *et al.* (2020a) from the recovered sequencing data, and found that it had mutations C28144T and G28085T relative to proCoV2, which means that in the region of interest this control matches Wuhan-Hu-1 with the addition of G28085T. Since none of the viral samples in Table 1 contain G28085T and the samples that prove most relevant below also lack C28144T (which is a frequent natural mutation among early Wuhan sequences), plasmid contamination did not afflict the viral samples in the deleted sequencing project.

Analysis of existing SARS-CoV-2 sequences emphasizes the perplexing discordance between collection date and distance to bat coronavirus relatives

To contextualize the viral sequences recovered from the deleted project, I first analyze early SARS-CoV-2 sequences already available in the GISAID database (Shu and McCauley 2017). The analyses described in this section are not entirely novel, but synthesize observations from multiple prior studies (Kumar *et al.* 2021; Pekar *et al.* 2021; Rambaut *et al.* 2020; Forster *et al.* 2020; Pipes *et al.* 2021) to provide key background.

Known human SARS-CoV-2 sequences are consistent with expansion from a single progenitor sequence (Kumar *et al.* 2021; Pekar *et al.* 2021; Rambaut *et al.* 2020; Forster *et al.* 2020; Pipes *et al.* 2021). However, attempts to infer this progenitor have been confounded by a perplexing fact: the earliest reported sequences from Wuhan are *not* the sequences most similar to SARS-CoV-2's bat coronavirus relatives (Pipes *et al.* 2021). This fact is perplexing because although the proximal origin of SARS-CoV-2 remains unclear (i.e., zoonosis versus lab accident), all reasonable explanations agree that at a deeper level the SARS-CoV-2 genome is derived from bat coronaviruses (Lytras *et al.* 2021). One would therefore expect the first reported SARS-CoV-2 sequences to be the most similar to these bat coronavirus relatives—but this is not the case.

This conundrum is illustrated in Figure 3, which plots the collection date of SARS-CoV-2 sequences in GISAID versus the relative number of mutational differences from RaTG13 (Zhou *et al.* 2020b), which is the bat coronavirus with the highest full-genome sequence identity to SARS-CoV-2. The earliest SARS-CoV-2 sequences were collected in Wuhan in December, but these sequences are more distant from RaTG13 than sequences collected in January from other locations in China or even other countries (Figure 3). The discrepancy is especially pronounced for sequences from patients who had visited the Huanan Seafood Market (WHO 2021). All sequences associated with this market differ from RaTG13 by at least three more mutations than sequences subsequently collected at various other locations (Figure 3)—a fact that is difficult to reconcile with the idea that the market was the original location of spread of a bat coronavirus into humans. Importantly, all these observations also hold true if SARS-CoV-2 is compared to other related bat coronaviruses (Lytras *et al.* 2021) such as RpYN06 (Zhou *et al.* 2021) or RmYN02 (Zhou *et al.* 2020a) rather than RaTG13 (Figure S3).

This conundrum can be visualized in a phylogenetic con-

Figure 4 Phylogenetic trees of SARS-CoV-2 sequences in GISAID with multiple observations among viruses collected before February, 2020. The trees are identical except they are rooted to make the progenitor each of the three sequences with highest identity to the RaTG13 bat coronavirus outgroup. Nodes are shown as pie charts with areas proportional to the number of observations of that sequence, and colored by where the viruses were collected. The mutations on each branch are labeled, with mutations towards the nucleotide identity in the outgroup in purple. The labels at the top of each tree give the first known virus identical to each putative progenitor, as well as mutations in that progenitor relative to proCoV2 (Kumar *et al.* 2021) and Wuhan-Hu-1. The monophyletic group containing C28144T is collapsed into a node labeled "clade B" in concordance with the naming scheme of Rambaut *et al.* (2020); this clade contains Wuhan-Hu-1. Figure S4 shows identical results are obtained if the outgroup is RpYN06 or RmYN02.

text by rooting a tree of early SARS-CoV-2 sequences so that the progenitor sequence is closest to the bat coronavirus outgroup. If we limit the analysis to sequences with at least two observations among strains collected no later than January 2020, there are three ways to root the tree in this fashion since there are three different sequences equally close to the outgroup (Figure 4, Figure S4). Importantly, none of these rootings place any Huanan Seafood Market viruses (or other Wuhan viruses from December 2019) in the progenitor node—and only one of the rootings has any virus from Wuhan in the progenitor node (in the leftmost tree in Figure 4, the progenitor node contains Wuhan/0126-C13/2020, which was reportedly collected on January 26, 2020). Therefore, inferences about the progenitor of SARS-CoV-2 based on comparison to related bat viruses are inconsistent with other evidence suggesting the progenitor is an early virus from Wuhan (Pipes *et al.* 2021).

Several plausible explanations have been proposed for the discordance of phylogenetic rooting with evidence that Wuhan was the origin of the pandemic. Rambaut *et al.* (2020) suggest that viruses from the clade labeled "B" in Figure 4 may just "happen" to have been sequenced first, but that other SARS-CoV-2 sequences are really more ancestral as implied by phylogenetic rooting. Pipes *et al.* (2021) discuss the conundrum in detail, and suggest that phylogenetic rooting could be incorrect due to technical reasons such as high divergence of the outgroup or unusual mutational processes not captured in substitution models. Kumar *et al.* (2021) agree that phylogenetic rooting is problematic, and circumvent this problem by using an alternative algorithm to infer a progenitor for SARS-CoV-2 that they name proCoV2. Notably, proCoV2 turns out to be identical to one of the putative progenitors yielded by my approach in Figure 4 of simply placing the root at the nodes closest to the outgroup. However, neither the sophisticated algorithm of Kumar *et al.* (2021) nor my more simplistic approach explain *why* the progenitor should be so different from the earliest sequences reported from Wuhan.

Before moving to the next section, I will also briefly address two less plausible explanations for the discordance between phylogenetic rooting and epidemiological data that have gained traction in discussion of SARS-CoV-2's origins. The first explanation, which has circulated on social media, suggests that the RaTG13 sequence might be faked in a way that confounds phylogenetic inference of SARS-CoV-2's progenitor. But although there are un-

usual aspects of RaTG13's primary sequencing data (Singla *et al.* 2020; Rahalkar and Bahulikar 2020), the conundrum about inferring the progenitor holds for other outgroups such as RpYN06, RmYN02, and more distant bat coronaviruses reported before emergence of SARS-CoV-2 such as ZC45 (Tang *et al.* 2020). The second explanation, which was proposed in a blog post by Garry (2021) and amplified by a popular podcast (Racaniello *et al.* 2021), is that there were multiple zoonoses from distinct markets, with the Huanan Seafood Market being the source of viruses in clade B, and some other market being the source of viruses that lack the T8782C and C28144T mutations. However, inspection of Figure 4 shows that clade B is connected to viruses lacking T8782C and C28144T by single mutational steps via other human isolates, so this explanation requires not only positing two markets with two progenitors differing by just two mutations, but also the exceedingly improbable evolution of one of these progenitors towards the other after it had jumped to humans.

Sequences recovered from the deleted project and better annotation of Wuhan-derived viruses help reconcile inferences about SARS-CoV-2's progenitor

To examine if the sequences recovered from the deleted data set help resolve the conundrum described in the previous section, I repeated the analyses including those sequences. In the process, I noted another salient fact: four GISAID sequences collected in Guangdong that fall in a putative progenitor node are from two different clusters of patients who traveled to Wuhan in late December of 2019 and developed symptoms before or on the day that they returned to Guangdong, where their viruses were ultimately sequenced (Chan *et al.* 2020; Kang *et al.* 2020b). Since these patients were clearly infected in Wuhan even though they were sequenced in Guangdong, I annotated them separately from both the other Wuhan and other China sequences.

Repeating the analysis of the previous section with these changes shows that several sequences from the deleted project and all sequences from patients infected in Wuhan but sequenced in Guangdong are more similar to the bat coronavirus outgroup than sequences from the Huanan Seafood Market (Figure 5). This fact suggests that the market sequences, which are the primary focus of the genomic epidemiology in the joint WHO-China report (WHO 2021), are not representative of the viruses that were circulating in Wuhan in late December of 2019 and early January of 2020.

Furthermore, it is immediately apparent that the discrepancy between outgroup rooting and the evidence that Wuhan was the origin of SARS-CoV-2 is alleviated by adding the deleted sequences and annotating Wuhan infections sequenced in Guangdong. The rooting of the middle tree in Figure 6 is now highly plausible, as half its progenitor node is derived from early Wuhan infections, which is more than any other equivalently large node. The first known sequence identical to this putative progenitor (Guangdong/HKU-SZ-002/2020) is from a patient who developed symptoms on January 4 while visiting Wuhan (Chan *et al.* 2020). This putative progenitor has three mutations towards the bat coronavirus outgroup relative to Wuhan-Hu-1 (C8782T, T28144C, and C29095T), and two mutations relative to proCoV2 (T18060C away from the outgroup and C29095T towards the outgroup). The leftmost tree in Figure 6, which has a progenitor identical to proCoV2 (Kumar *et al.* 2021) also looks plausible, with some weight from Wuhan sequences. However, analysis of this rooting is limited by the fact that the defining C18060T mutation is in a region not covered in the deleted sequences. The rightmost tree in Figure 6 looks less plausible, as it has almost no weight from Wuhan and the first sequence identical to its progenitor was not collected until January 24.

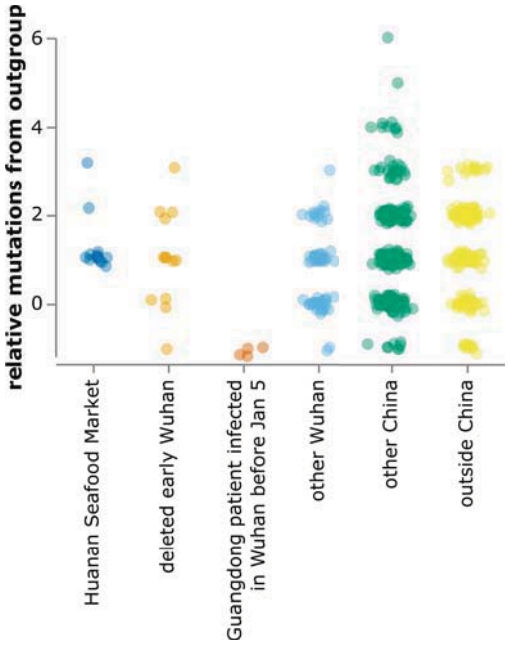


Figure 5 Relative mutational distance from RaTG13 bat coronavirus outgroup calculated *only* over the region of the SARS-CoV-2 genome covered by sequences from the deleted data set (21,570–29,550). The plot shows sequences in GISAID collected before February of 2020, as well as the 13 early Wuhan epidemic sequences in Table 1. Mutational distance is calculated relative to proCoV2, and points are jittered on the y-axis. Go to https://jbloom.github.io/SARS-CoV-2_PRJNA612766/deltadist_jitter.html for an interactive version of this plot that enables toggling the outgroup to RpYN06 or RmYN02, mouseovers to see details for each point, and adjustment of jittering.

We can also qualitatively examine the three progenitor placements in Figure 6 using the principle employed by Worobey *et al.* (2020) to help evaluate scenarios for emergence of SARS-CoV-2 in Europe and North America: namely that during a growing outbreak, a progenitor is likely to give rise to multiple branching lineages. This principle is especially likely to hold for the scenarios in Figure 6, since there are multiple individuals infected with each putative progenitor sequence, implying multiple opportunities to transmit descendants with new mutations. Using this qualitative principle, the middle scenario in Figure 6 seems most plausible, the leftmost (proCoV2) scenario also seems plausible, and the rightmost scenario seems less plausible. I acknowledge these arguments are purely qualitative and lack the formal statistical analysis of Worobey *et al.* (2020)—but as discussed below, there may be wisdom in qualitative reasoning when there are valid concerns about the nature of the underlying data.

Discussion

I have identified and recovered a deleted set of partial SARS-CoV-2 sequences from the early Wuhan epidemic. Analysis of these sequences leads to several conclusions. First, the Huanan Seafood Market sequences that were the focus of the joint WHO-China report (WHO 2021) are not representative of all SARS-CoV-2 in Wuhan early in the epidemic. The deleted data as well

Figure 6 Phylogenetic trees like those in Figure 4 with the addition of the early Wuhan epidemic sequences from the deleted data set, and Guangdong patients infected in Wuhan prior to January 5 annotated separately. Because the deleted sequences are partial, they cannot all be placed unambiguously on the tree. Therefore, they are added to each compatible node proportional to the number of sequences already in that node. The deleted sequences with C28144T (clade B) or C29095T (putative progenitor in middle tree) can be placed relatively unambiguously as defining mutations occur in the sequenced region, but those that lack either of these mutations are compatible with a large number of nodes including the proCoV2 putative progenitor. Figure S4 demonstrates that the results are identical if RpYN06 or RmYN02 is instead used as the outgroup.

as existing sequences from Wuhan-infected patients hospitalized in Guangdong show early Wuhan sequences often carried the T29095C mutation and were less likely to carry T8782C / C28144T than sequences in the joint WHO-China report (WHO 2021). Second, given current data, there are two plausible identities for the progenitor of all known SARS-CoV-2. One is proCoV2 described by Kumar *et al.* (2021), and the other is a sequence that carries three mutations (C8782T, T28144C, and C29095T) relative to Wuhan-Hu-1. Crucially, both putative progenitors are three mutations closer to SARS-CoV-2's bat coronavirus relatives than sequences from the Huanan Seafood Market. Note also that the progenitor of all known SARS-CoV-2 sequences could still be downstream of the sequence that infected patient zero depending on the transmission dynamics of the first infections.

The fact that such an informative data set was deleted has implications beyond those gleaned directly from the recovered sequences. Samples from early outpatients in Wuhan are a gold mine for anyone seeking to understand spread of the virus. Even my analysis of the partial sequences is revealing, and it clearly would have been more scientifically informative to fully sequence the samples rather than surreptitiously delete the partial sequences. There is no plausible scientific reason for the deletion: the sequences are perfectly concordant with the samples described in Wang *et al.* (2020a,b), there are no corrections to the paper, the paper states human subjects approval was obtained, and the sequencing shows no evidence of plasmid or sample-to-sample contamination. It therefore seems likely the sequences were deleted to obscure their existence. Particularly in light of the directive that labs destroy early samples (Pingui 2020) and multiple orders requiring approval of publications on COVID-19 (China CDC 2020; Kang *et al.* 2020a), this suggests a less than wholehearted effort to trace early spread of the epidemic.

Another important implication is that genomic epidemiology studies of early SARS-CoV-2 need to pay as much attention to the provenance and annotation of the underlying sequences as technical considerations. There has been substantial scientific effort expended on topics such as phylogenetic rooting (Pipes *et al.* 2021; Morel *et al.* 2021), novel algorithms (Kumar *et al.* 2021), and correction of sequencing errors (Turakhia *et al.* 2020). Future studies should devote equal effort to going beyond the annotations in GISAID to carefully trace the location of patient

infection and sample sequencing. The potential importance of such work is revealed by the observation that many of the sequences closest to SARS-CoV-2's bat coronavirus relatives are from early patients who were infected in Wuhan, but then sequenced in and attributed to Guangdong.

There are several caveats to this study. Most obviously, the sequences recovered from the deleted data set are partial and lack full metadata. Therefore, it is impossible to unambiguously place them phylogenetically, or determine exactly when they were collected. However, little can be done to mitigate this caveat beyond my failed attempt to contact the corresponding authors of Wang *et al.* (2020a). It is also important to note that my phylogenetic analyses use relatively simple methods to draw qualitative conclusions without formal statistical testing. Further application of more advanced methods would be a welcome advance. However, qualitative and visual analyses do have advantages when the key questions relate more to the underlying data than the sophistication of the inferences. Finally, both plausible putative progenitors require that an early mutation to SARS-CoV-2 was a reversion towards the bat coronavirus outgroups (either C18060T or C29095T) on a branch that subsequently gave rise to multiple distinct descendants. Such a scenario can only be avoided by invoking recombination very early in the pandemic, which is not entirely implausible for a coronavirus (Boni *et al.* 2020). However, because the outgroups have ∼4% nucleotide divergence from SARS-CoV-2, a mutation towards the outgroup is also entirely possible. Of course, future identification of additional early sequences could fully resolve these questions.

More broadly, the approach taken here suggests it may be possible to learn more about the origin or early spread of SARS-CoV-2 even without an international investigation. Minimally, it should be immediately possible for the NIH to determine the date and purported reason for deletion of the data set analyzed here, since the only way sequences can be deleted from the SRA is by an e-mail request to SRA staff (SRA 2021). In addition, I suggest it could be worthwhile to review e-mail records to identify other SRA deletions, which are already known to include SRR11119760 and SRR11119761 (USRTK 2020). Importantly, SRA deletions do not imply any malfeasance: there are legitimate reasons for removing sequencing runs, and the SRA houses >13-million runs making it infeasible for its staff to validate the rationale for all requests. However, the current study suggests that at least in one case, the trusting structures of science have been abused to obscure sequences relevant to the early spread of SARS-CoV-2 in Wuhan. A careful re-evaluation of other archived forms of scientific communication, reporting, and data could shed additional light on the early emergence of the virus.

Methods

Code and data availability

The computer code and input data necessary to reproduce all analyses described in this paper are available on GitHub at https://github.com/jbloom/SARS-CoV-2_PRJNA612766. This GitHub repository includes a `Snakemake` (Mölder *et al.* 2021) pipeline that fully automates all steps in the analysis except for downloading of sequences from GISAID, which must be done manually as described in the GitHub repository's README in order to comply with GISAID data sharing terms.

The deleted SRA files recovered from the Google Cloud are all available at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/tree/main/results/sra_downloads. I have suffixed the file extension `.sra` to all these files. The consensus sequences recovered from these deleted SRA files are linked to in the relevant Methods subsection below.

Archiving of key weblinks

I have digitally archived key weblinks in the Wayback Machine, including a subset of the SRA files from PRJNA612766 on the Google Cloud:

- The first supplementary table of Farkas *et al.* (2020) is archived at https://web.archive.org/web/20210502130356/https://dfzljdn9uc3pi.cloudfront.net/2020/9255/1/Supplementary_Table_1.xlsx.
- SRR11313485: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313485/SRR11313485
- SRR11313486: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313486/SRR11313486
- SRR11313274: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313274/SRR11313274
- SRR11313275: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313275/SRR11313275
- SRR11313285: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313285/SRR11313285
- SRR11313286: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313286/SRR11313286
- SRR11313448: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313448/SRR11313448
- SRR11313449: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313449/SRR11313449
- SRR11313427: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313427/SRR11313427
- SRR11313429: https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313429/SRR11313429

Recovery of SRA files from deleted project PRJNA612766

I parsed the first supplementary table of Farkas *et al.* (2020) to extract the accessions for sequencing runs for deleted SRA BioProject PRJNA612766. By cross-referencing the samples described in this table to Wang *et al.* (2020a,b), I identified the accessions corresponding to the 34 early outpatient samples who were positive, as well as the accessions corresponding to the 16 hospitalized patient samples from February. Samples had both 10 minute and 4 hour sequencing runtime accessions, which were combined in the subsequent analysis. I also identified the samples corresponding to the high-copy plasmid controls to enable analysis of the plasmid sequence to rule out contamination. The code used to parse the Excel table is available as a Jupyter notebook at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/tree/main/manual_analyses/PRJNA612766.

I recovered the SRA files from the Google Cloud by using `wget` to download files with from paths like `https://storage.googleapis.com/nih-sequence-read-archive/run/SRR11313485/SRR11313485`. Note that I cannot guarantee that these Google Cloud links will remain active, as my analyses of other deleted SRA runs (beyond the scope of this study) indicates that only sometimes are deleted SRA files still available via the Google Cloud. For this reason, key runs have been archived on the Wayback Machine as described above, and all downloaded SRA files relevant to this study are included in the GitHub repository. Note also that as described in this paper's main text, two SRA files could not be downloaded from the Google Cloud using the aforementioned method, and so are not part of this study.

Alignment of recovered reads and calling of consensus sequences

The downloaded SRA files were converted to FASTQ files using `fasterq-dump` from the SRA Toolkit. The FASTQ files were pre-processed with `fastp` (Chen *et al.* 2018) to trim reads and remove low-quality ones (the exact settings using in this pre-processing are specified in the `Snakemake` file in the GitHub repository).

The reads in these FASTQ files were then aligned to a SARS-CoV-2 reference genome using `minimap2` (Li 2018) with default settings. The reference genome used for the entirety of this study is proCoV2 (Kumar *et al.* 2021), which was generated by making the following three single-nucleotide changes to the Wuhan-Hu-1 reference (ASM985889v2) available on NCBI: C8782T, C18060T, and T28144C.

I processed the resulting alignments with `samtools` and `pysam` to determine the coverage at each site by aligned nucleotides with a quality score of at least 20. These coverage plots are in Figure S1 and Figure S2; the legends of these figures also link to interactive versions of the plots that enable zooming and mouseovers to get statistics for specific sites. I called the consensus sequence at a site if this coverage was ≥3 and >80% of the reads agreed on the identity. These consensus sequences over the entire SARS-CoV-2 genome are available at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/raw/main/results/consensus/consensus_seqs.csv; note that they are mostly `N` nucleotides

since the sequencing approach of Wang *et al.* (2020a) only covers part of the genome.

I only used the recovered consensus sequences in the downstream analyses if it was possible to call the consensus identity at ≥95% of the sites in the region of interest from site 21,570 to 29,550. These are the sequences listed in Table 1, and as described in that table, all mutation calls were at sites with coverage ≥10. These sequences in the region of interest (21,570 to 29,550) are available at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/results/recovered_seqs.fa.

Bat coronavirus outgroup sequences

For analyses that involved comparisons to SARS-CoV-2's bat coronavirus relatives (Lytras *et al.* 2021), the bat coronavirus sequences were manually downloaded from GISAID (Shu and McCauley 2017). The sequences used were RaTG13 (Zhou *et al.* 2020b), RmYN02 (Zhou *et al.* 2020a), and RpYN06 (Zhou *et al.* 2021)—although the multiple sequence alignment of these viruses to SARS-CoV-2 also contains PrC31 (Li *et al.* 2021), which was not used in the final analyses as it more diverged from SARS-CoV-2 than the other three bat coronaviruses at a whole-genome level. The GISAID accessions for these sequences are listed at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/comparator_genomes_gisaid/accessions.txt, and a table acknowledging the labs and authors is at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/comparator_genomes_gisaid/acknowledgments.csv. Sites in SARS-CoV-2 were mapped to their corresponding nucleotide identities in the bat coronavirus outgroups via a multiple sequence alignment of proCoV2 to the bat coronaviruses generated using `mafft` (Katoh and Standley 2013).

Curation and analysis of early SARS-CoV-2 sequences from GISAID

For the broader analyses of existing SARS-CoV-2 sequences, I downloaded all sequences from collected prior to March of 2020 from GISAID. The accessions of these sequences are at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/gisaid_sequences_through_Feb2020/accessions.txt, and a table acknowledging the labs and authors is at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/gisaid_sequences_through_Feb2020/acknowledgments.csv.

I then used `mafft` (Katoh and Standley 2013) to align these sequences to the proCoV2 reference described above, stripped any sites that were gapped relative to the reference, and filtered the sequences using the following criteria:

- I removed any sequences collected after February 28, 2020.
- I removed any sequences that had ≥4 mutations within any 10-nucleotide stretch, as such runs of mutations often indicate sequencing errors.
- I removed any sequence for which the alignment covered <90% of the proCoV2 sequence.
- I removed any sequence with ≥15 mutations relative to the reference.
- I removed any sequence with ≥5,000 ambiguous nucleotides.

I then annotated the sequences using some additional information. First, I annotated sequences based on the joint WHO-China report (WHO 2021) and also Zhu *et al.* (2020) to keep only one representative from multiply sequenced patients, and to indicate which sequences were from patients associated with the Huanan Seafood Market. My version of these annotations is at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/WHO_China_Report_Dec2019_cases.yaml. Next, I identified some sequences in the set that were clearly duplicates from the same patient, and removed these. The annotations used to remove these duplicates are at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/seqs_to_exclude.yaml. Finally, I used information from Chan *et al.* (2020) and Kang *et al.* (2020b) to identify patients who were infected in Wuhan before January 5 of 2020, but ultimately sequenced in Guangdong: these annotations are at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/data/Wuhan_exports.yaml.

I next removed any of the handful of mutations noted by Turakhia *et al.* (2020) to be lab artifacts that commonly afflict SARS-CoV-2 sequences. I also limited the analyses to the region of the genome that spans from the start of the first coding region (ORF1ab) to the end of the last (ORF10), because I noticed that some sequences had suspicious patterns (such as many mutations or runs of mutations) near the termini of the genome.

The plot in Figure 3 contains all of the GISAID sequences after this filtering. The plot in Figure 5 shows the filtered GISAID sequences collected before February of 2020 plus the 13 good coverage recovered partial early outpatient sequences (Table 1), considering only the region covered by the partial sequences (21,570 to 29,550).

Phylogenetic analyses

The phylogenetic trees were inferred using the GISAD sequences after the filtering and annotations described above, only considering sequences with ≥95% coverage over the region of interest that were collected before February of 2020. In addition, after generating this sequence set I removed any sequence variants with a combination of mutations that was not observed at least twice so the analysis only includes multiply observed sequence variants. A file indicating the unique sequences used for the phylogenetic analysis, their mutations relative to proCoV2, and other sequences in that cluster is at https://github.com/jbloom/SARS-CoV-2_PRJNA612766/blob/main/results/phylogenetics/all_alignment.csv.

I then used `IQ-Tree` (Minh *et al.* 2020) to infer a maximum-likelihood phylogenetic tree using a GTR nucleotide substitution model with empirical nucleotide frequencies, and collapsing zero-length branches to potentially allow a multifurcating tree. The inference yielded the tree topology and branch lengths shown in all figures in this study with phylogenetic trees. I then rendered the images of the tree using `ETE 3` (Huerta-Cepas *et al.* 2016), manually re-rooting the tree to place the first (progenitor) node at each of the three nodes that have the highest identity to the bat coronavirus outgroup. In these images, node sizes are proportional to the number of sequences in that node, and are colored in proportion to the location from which those sequences are derived. As indicated in the legend to Figure 4, the node containing the monophyletic set of sequences with C28144T is collapsed into a single node in the tree images.

For the trees in which I added the recovered sequences from the deleted data set (Figure 6), the actual trees are exactly the same as those inferred using the GISAID sequences above. The difference is that the sequences from the deleted data set are then added to each node with which they are compatible given their mutations in an amount proportional to the size of the node, the logic being that a sequence is more likely to fall into larger clusters.

Interactive versions of some figures

Interactive versions of some figures are available at https://jbloom.github.io/SARS-CoV-2_PRJNA612766/, and were created using `Altair` (VanderPlas *et al.* 2018)

Acknowledgments

I thank the citizens and scientists on Twitter who helped inspire this study and inform its background through their analyses and discussions of the early spread of SARS-CoV-2. I thank the scientists and labs who contributed sequences used in this study to the GISAID database; the names of these scientists and labs are listed in the tables linked to in the Methods. The scientific computing infrastructure used in this work was supported by the NIH Office of Research Infrastructure Programs under S10OD028685. The author is an Investigator of the Howard Hughes Medical Institute.

Competing interests

The author consults for Moderna on SARS-CoV-2 evolution and epidemiology, consults for Flagship Labs 77 on viral evolution and deep mutational scanning, and has the potential to receive a share of IP revenue as an inventor on a Fred Hutch licensed technology/patent (application WO2020006494) related to deep mutational scanning of viral proteins.

Literature Cited

Bedford, T., A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, *et al.*, 2020 Cryptic transmission of SARS-CoV-2 in washington state. Science **370**: 571–575.

Boni, M. F., P. Lemey, X. Jiang, T. T.-Y. Lam, B. W. Perry, *et al.*, 2020 Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature Microbiology **5**: 1408–1417.

Chan, J. F.-W., S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, *et al.*, 2020 A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet **395**: 514–523.

Chen, N., M. Zhou, X. Dong, J. Qu, F. Gong, *et al.*, 2020 Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet **395**: 507–513.

Chen, S., Y. Zhou, Y. Chen, and J. Gu, 2018 fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics **34**: i884–i890.

China CDC, 2020 Memo to the Offices of the Chinese Center for Disease Control and Prevention. https://www.documentcloud.org/documents/7340336-China-CDC-Sup-Regs.html.

Cohen, J., 2020 Wuhan seafood market may not be source of novel virus spreading globally. Science **10**: 10.1126/science.abb0611.

Deng, X., W. Gu, S. Federman, L. Du Plessis, O. G. Pybus, *et al.*, 2020 Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science **369**: 582–587.

Farkas, C., F. Fuentes-Villalobos, J. L. Garrido, J. Haigh, and M. I. Barría, 2020 Insights on early mutational events in SARS-CoV-2 virus reveal founder effects across geographical regions. PeerJ **8**: e9255.

Fauver, J. R., M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, *et al.*, 2020 Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell **181**: 990–996.

Forster, P., L. Forster, C. Renfrew, and M. Forster, 2020 Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences **117**: 9241–9243.

Garry, R. F., 2021 Early appearance of two distinct genomic lineages of SARS-CoV-2 in different wuhan wildlife markets suggests SARS-CoV-2 has a natural origin. https://virological.org/t/early-appearance-of-two-distinct-genomic-lineages-of-sars-cov-2-in-different-wuhan-wildlife-markets-suggests-sars-cov-2-has-a-natural-origin/691.

Global Times, 2020 Wuhan's huanan seafood market a victim of COVID-19: CDC director. https://www.globaltimes.cn/content/1189506.shtml, archived at https://web.archive.org/web/20200528062530/https://www.globaltimes.cn/content/1189506.shtml.

Health Times, 2020 Experts judge the source of the new coronavirus: December 8 last year may not be the earliest time of onset. https://www.guancha.cn/politics/2020_02_27_538822.shtml, archived and detailed at https://docs.google.com/document/d/e/2PACX-1vTQxG822DtqP7IZSjLj751Mrm8Ev7leksXfjBLsA9KJ0_tbGV6YJAAjuijPnwz_YmUQGY1PZUl5LcCI/pub.

Huang, C., Y. Wang, X. Li, L. Ren, J. Zhao, *et al.*, 2020 Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet **395**: 497–506.

Huerta-Cepas, J., F. Serra, and P. Bork, 2016 Ete 3: reconstruction, analysis, and visualization of phylogenomic data. Molecular biology and evolution **33**: 1635–1638.

Kang, D., M. Cheng, and S. McNeil, 2020a China clamps down in hidden hunt for coronavirus origins. https://apnews.com/article/united-nations-coronavirus-pandemic-china-only-on-ap-bats-24fbadc58cee3a40bca2ddf7a14d2955, The actual China State Council order described in the article is at https://www.documentcloud.org/documents/7340337-State-Research-regulations.html.

Kang, M., J. Wu, W. Ma, J. He, J. Lu, *et al.*, 2020b Evidence and characteristics of human-to-human transmission of sars-cov-2. medRxiv .

Katoh, K. and D. M. Standley, 2013 Mafft multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30**: 772–780.

Kumar, S., Q. Tao, S. Weaver, M. Sanderford, M. A. Caraballo-Ortiz, *et al.*, 2021 An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. Molecular Biology and Evolution msab118.

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**: 3094–3100.

Li, L., J. Wang, X. Ma, J. Li, X. Yang, *et al.*, 2021 A novel sars-cov-2 related virus with complex recombination isolated from bats in yunnan province, china. bioRxiv .

Lytras, S., J. Hughes, D. Martin, A. de Klerk, R. Lourens, *et al.*, 2021 Exploring the natural origins of SARS-CoV-2 in the light of recombination. bioRxiv .

Ma, J., 2020 Coronavirus: China's first confirmed Covid-19 case traced back to November 17. South China Morning Post, https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back, archived at https://web.archive.org/web/20200315011702/https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, *et al.*, 2020 IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution **37**: 1530–1534.

Mölder, F., K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, *et al.*, 2021 Sustainable data analysis with snakemake. F1000Research **10**: 33.

Morel, B., P. Barbera, L. Czech, B. Bettisworth, L. Hübner, *et al.*, 2021 Phylogenetic analysis of SARS-CoV-2 data is difficult. Molecular Biology and Evolution **38**: 1777–1791.

Pekar, J., M. Worobey, N. Moshiri, K. Scheffler, and J. O. Wertheim, 2021 Timing the SARS-CoV-2 index case in hubei province. Science **372**: 412–417.

Pingui, Z., 2020 China confirms unauthorised labs were told to destroy early coronavirus samples. South China Morning Post, https://www.scmp.com/news/china/society/article/3084635/china-confirms-unauthorised-labs-were-told-destroy-early, archived at https://web.archive.org/web/20210103124552/https://www.scmp.com/news/china/society/article/3084635/china-confirms-unauthorised-labs-were-told-destroy-early.

Pipes, L., H. Wang, J. P. Huelsenbeck, and R. Nielsen, 2021 Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. Molecular Biology and Evolution **38**: 1537–1543.

ProMED, 2019 Undiagnosed pneumonia—China (Hubei): Request for information. https://promedmail.org/promed-post/?id=6864153.

Racaniello, V., D. Despommier, A. Dove, R. Condit, and B. Barker, 2021 TWiV 762: SARS-CoV-2 origins with Robert Garry. https://www.microbe.tv/twiv/twiv-762/.

Rahalkar, M. and R. Bahulikar, 2020 The anomalous nature of the fecal swab data, receptor binding domain and other questions in RaTG13 genome. Preprints p. 2020080205.

Rambaut, A., E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, *et al.*, 2020 A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology **5**: 1403–1407.

Shu, Y. and J. McCauley, 2017 GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance **22**: 30494.

Singla, M., S. Ahmad, C. Gupta, and T. Sethi, 2020 De-novo assembly of RaTG13 genome reveals inconsistencies further

obscuring SARS-CoV-2 origins. Preprints p. 2020080595.

SRA, 2021 SRA data updates. https://www.ncbi.nlm.nih.gov/sra/docs/submitupdate/#how-do-i-withdraw-sra-data, Last accessed June 2, 2021; a copy of the page is digitally archived on the Wayback Machine at https://web.archive.org/web/20210505002105/https://www.ncbi.nlm.nih.gov/sra/docs/submitupdate/.

Tang, X., C. Wu, X. Li, Y. Song, X. Yao, *et al.*, 2020 On the origin and continuing evolution of SARS-CoV-2. National Science Review **7**: 1012–1023.

Turakhia, Y., N. De Maio, B. Thornlow, L. Gozashti, R. Lanfear, *et al.*, 2020 Stability of SARS-CoV-2 phylogenies. PLoS Genetics **16**: e1009175.

USRTK, 2020 Altered datasets raise more questions about reliability of key studies on coronavirus origins. https://usrtk.org/tag/pangolin-papers/, which links to the actual SRA e-mails at https://usrtk.org/wp-content/uploads/2020/12/NCBI-Emails.pdf.

VanderPlas, J., B. E. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, *et al.*, 2018 Altair: interactive statistical visualizations for Python. Journal of Open Source Software **3**: 1057.

Wang, M., A. Fu, B. Hu, Y. Tong, R. Liu, *et al.*, 2020a Nanopore target sequencing for accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. medRxiv **10.1101/2020.03.04.20029538**.

Wang, M., A. Fu, B. Hu, Y. Tong, R. Liu, *et al.*, 2020b Nanopore targeted sequencing for the accurate and comprehensive detection of SARS-CoV-2 and other respiratory viruses. Small **16**: 2002169.

WHO, 2021 WHO-convened global study of origins of SARS-CoV-2: China Part. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.

Worobey, M., J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, *et al.*, 2020 The emergence of SARS-CoV-2 in Europe and North America. Science **370**: 564–570.

Xiao, K., J. Zhai, Y. Feng, N. Zhou, X. Zhang, *et al.*, 2020 Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature **583**: 286–289.

Zhou, H., X. Chen, T. Hu, J. Li, H. Song, *et al.*, 2020a A novel bat coronavirus closely related to sars-cov-2 contains natural insertions at the s1/s2 cleavage site of the spike protein. Current Biology **30**: 2196–2203.

Zhou, H., J. Ji, X. Chen, Y. Bi, J. Li, *et al.*, 2021 Identification of novel bat coronaviruses sheds light on the evolutionary origins of sars-cov-2 and related viruses. Cell .

Zhou, P., X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, *et al.*, 2020b A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature **579**: 270–273.

Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, *et al.*, 2020 Brief report: A novel coronavirus from patients with pneumonia in china, 2019. The New England Journal of Medicine **382**: 727.