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

Journal
Science, 377(6609)

ISSN
0036-8075

Authors
Pekar, Jonathan E
Magee, Andrew
Parker, Edyth
et al.

Publication Date
2022-08-26

DOI
10.1126/science.abp8337

Peer reviewed

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

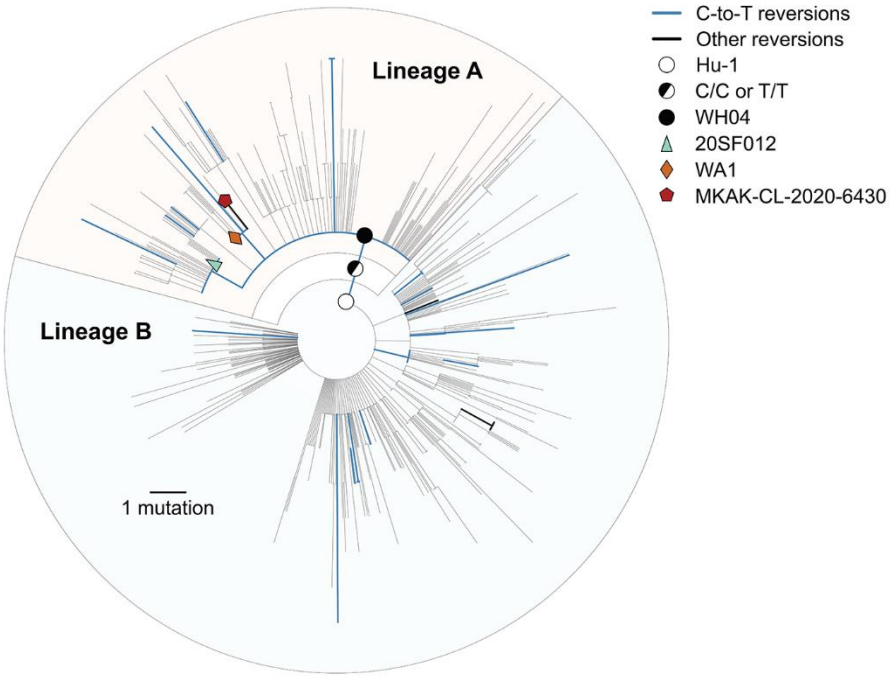

Fig. 1. Maximum likelihood phylogeny of the early SARS-CoV-2 pandemic, showing nucleotide reversions and putative candidates for the ancestral haplotype at the most common recent ancestor (MRCA). Putative ancestral haplotypes are identified with colored shapes. Reversions from the Hu-1 reference genotype 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.

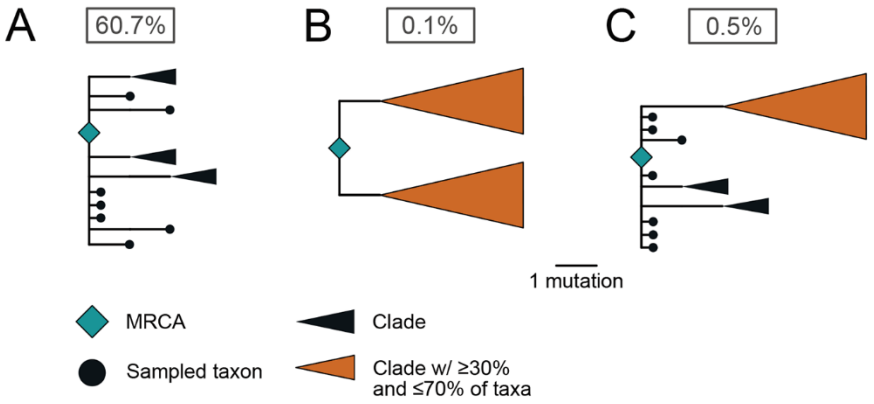

Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations. (**A**) A large polytomy of at least 20 descendent lineages, consistent with the base of both lineages A and B. (**B**) Topology matching a C/C ancestral haplotype: two clades each 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.

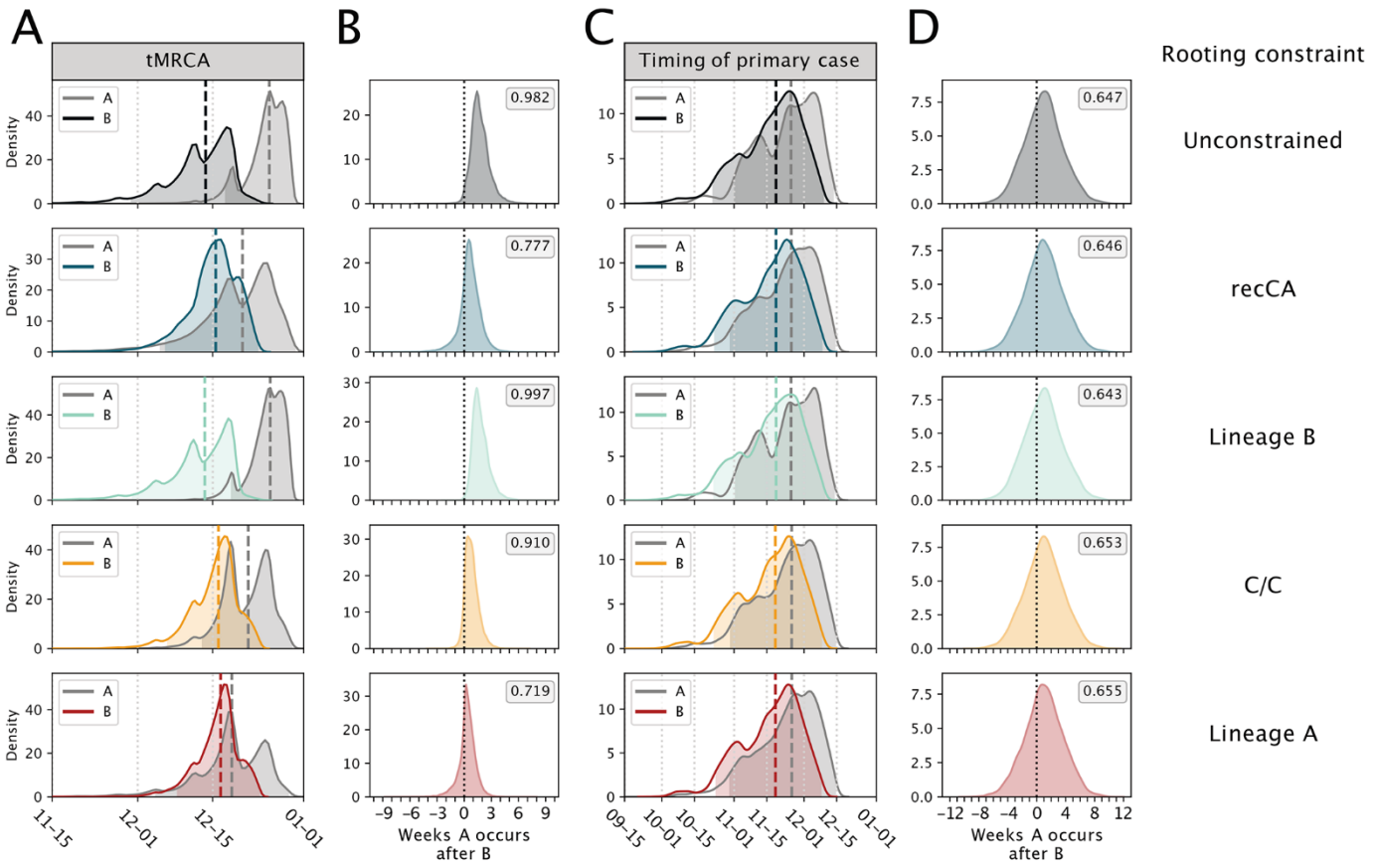

Fig. 3. Comparison of the tMRCA and primary case dates for lineage A and lineage B across rooting strategies. Each row represents a different rooting constraint in phylodynamic analysis, with lineage B, C/C and lineage A representing a fixed ancestral haplotype. (**A**) The tMRCA for lineages A and B. (**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.

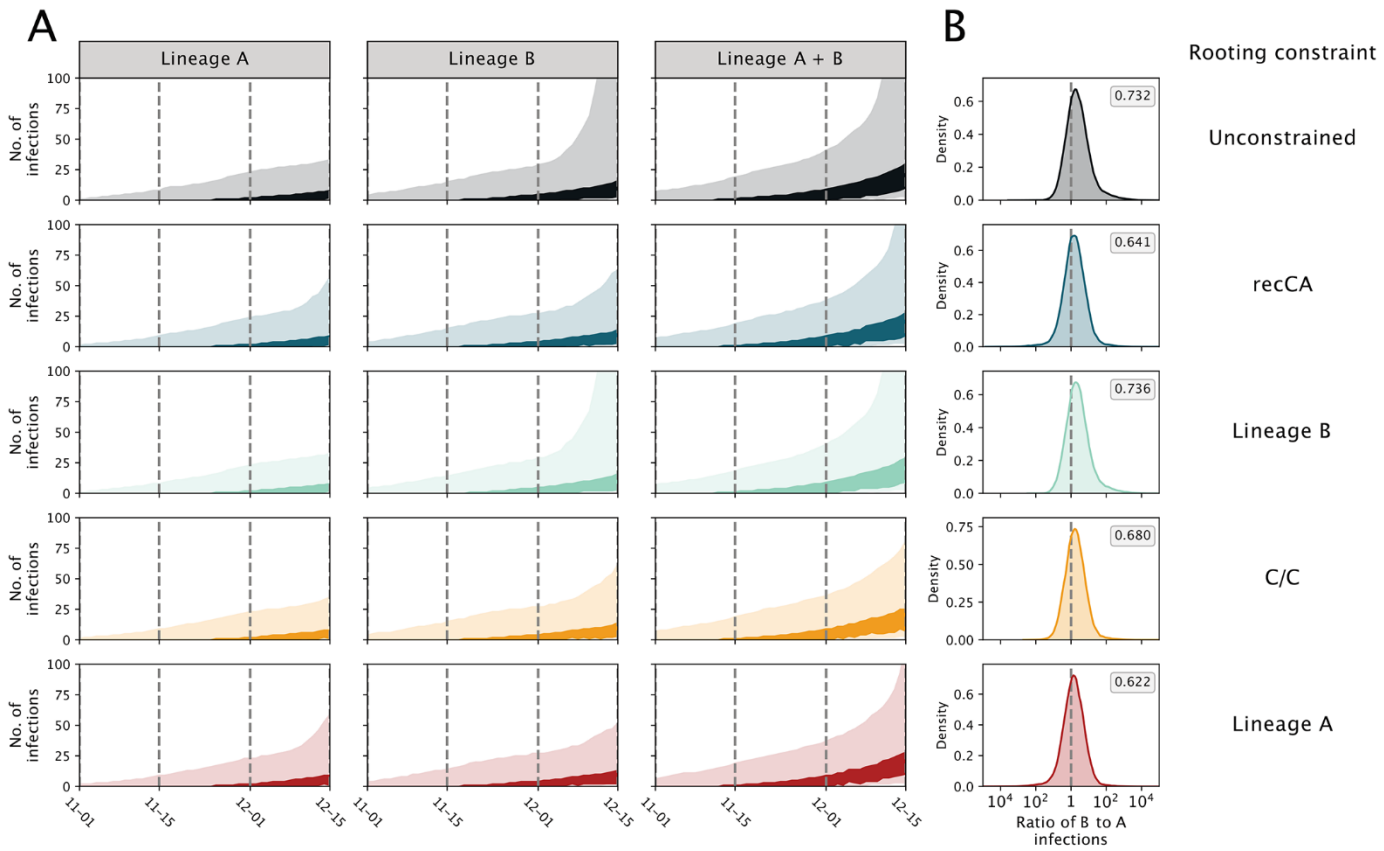

Fig. 4. Dynamics of simulated SARS-CoV-2 epidemics resulting from separate introductions of lineages A and B. Each row represents a different rooting constraint in phylodynamic analysis, with lineage B, C/C and lineage A representing a fixed ancestral haplotype. (**A**) Estimated number of infections. 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.