CORONAVIRUS

# 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]*

Understanding the circumstances that lead to pandemics is important for their prevention. We analyzed 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 most probably 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 to 8 December), and 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 before 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 2019 (COVID-19) pandemic that caused more than 5 million confirmed deaths in the 2 years after 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 11 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 and T28144), whereas lineage A viruses have a "T/C" pattern (C8782T and 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 because their existence suggests within-human evolution of one lineage toward the other by way of a transitional form.

Questions about these lineages remain: If lineage B viruses are more distantly related to sarbecoviruses from *Rhinolophus* bats, then (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 that contain 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 that 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 (fig. S1 and supplementary text) (*11*). We confirmed that a C/C genome from South Korea sharing three such mutations had low sequencing depth at position 28144 (≤10×), a T/T genome sampled in Singapore had low coverage at both 8782 and 28144 (≤10×), 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 11 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 in the Diamond Princess cruise ship outbreak and subsequent COVID-19 waves in New York City

[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, MD 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] Department of Microbiology and Immunology, School of Medicine, Tulane University, New Orleans, LA 70112, USA. [16] Zalgen Labs, 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.)**

and San Diego (figs. S2 to S5 and supplementary text). Instances of convergent evolution are identifiable because SARS-CoV-2 phylogenies exist in "near-perfect" tree space, in which 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 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 nonrecombinant 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 (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 distinct reversions and 631 distinct 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 of 381, 6.04%), compared with substitutions at all other sites (631 of 29,134, 2.17%).

Most reversions were C-to-T mutations (19 of 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; for example, WA1) and C29095T (for example, 20SF012), as well as C24023T, C25000T, C4276T, and C22747T in mid-late January and February 2020. Hence, triple revertant genomes, such as 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. We continue to observe these reversion patterns throughout the pandemic, including in the emergence of

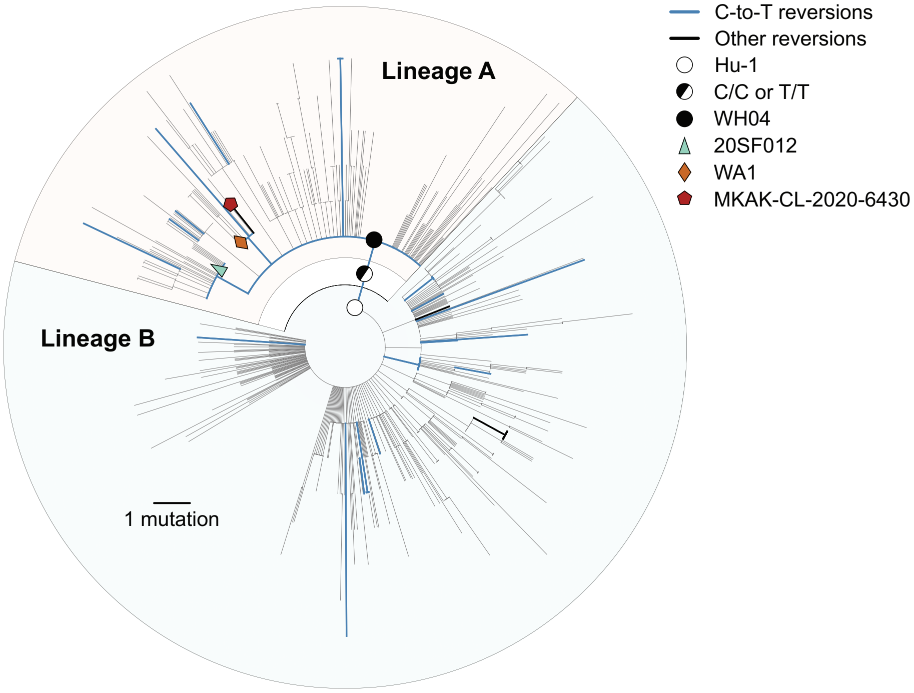

**Fig. 1. Maximum likelihood phylogeny of the early SARS-CoV-2 pandemic, showing nucleotide reversions and putative candidates for the ancestral haplotype at the MRCA.** Putative ancestral haplotypes are identified with colored shapes. Reversions from the Hu-1 reference genotype to the recCA are colored.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

**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 genome with 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 > 3.2; **BF > 10; ***BF > 100. 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+C29095T (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.

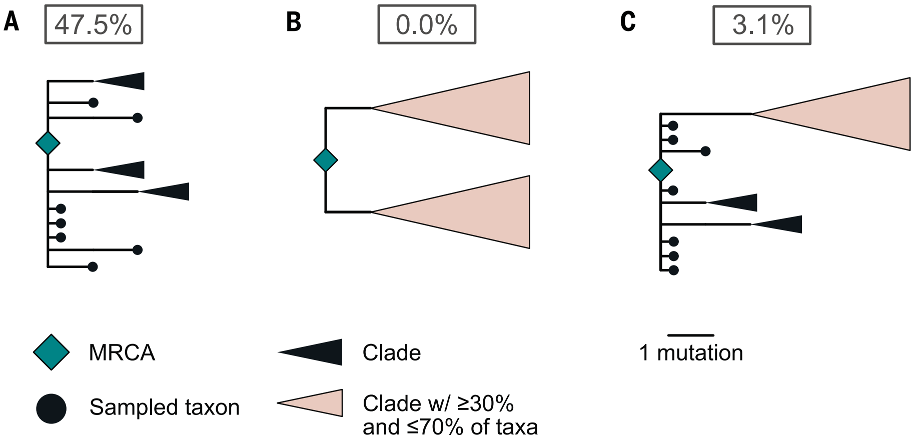

**Fig. 2. Probability of phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations.** (**A**) A large polytomy of at least 100 descendent lineages, which is consistent with the base of both lineages A and B. (**B**) Topology matching a C/C ancestral haplotype: two clades, each one mutation from the ancestor, both with polytomies of at least 100 descendent lineages. (**C**) Topology matching either a lineage A or lineage B ancestral haplotype: a basal polytomy with at least 100 descendent lineages, including a large clade separated by two mutations, also possessing a polytomy of at least 100 descendent lineages. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the respective boxes.

World Health Organization (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 nonreversible, 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) (*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 (BF = 9.4), 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 time of MRCA (tMRCA) for SARS-CoV-2 to be 11 December 2019 [95% highest posterior density (HPD) interval, 25 November to 12 December] by using unconstrained rooting. It has been suggested that a phylogenetic root in lineage A would produce an older tMRCA than would a lineage B rooting (*21*). Therefore, we developed an approach to assign a haplotype as the SARS-CoV-2 MRCA (A, B, C/C, A.1, or A+C29095T) and inferred the tMRCA (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 (such as 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 to 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, respectively, and each lineage is characterized by a large polytomy (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 (for example, 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), which is present in 47.5% of simulated epidemics (Fig. 2A).

By 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 0.0% 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 3.1% of our simulations (Fig. 2C and supplementary text).

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

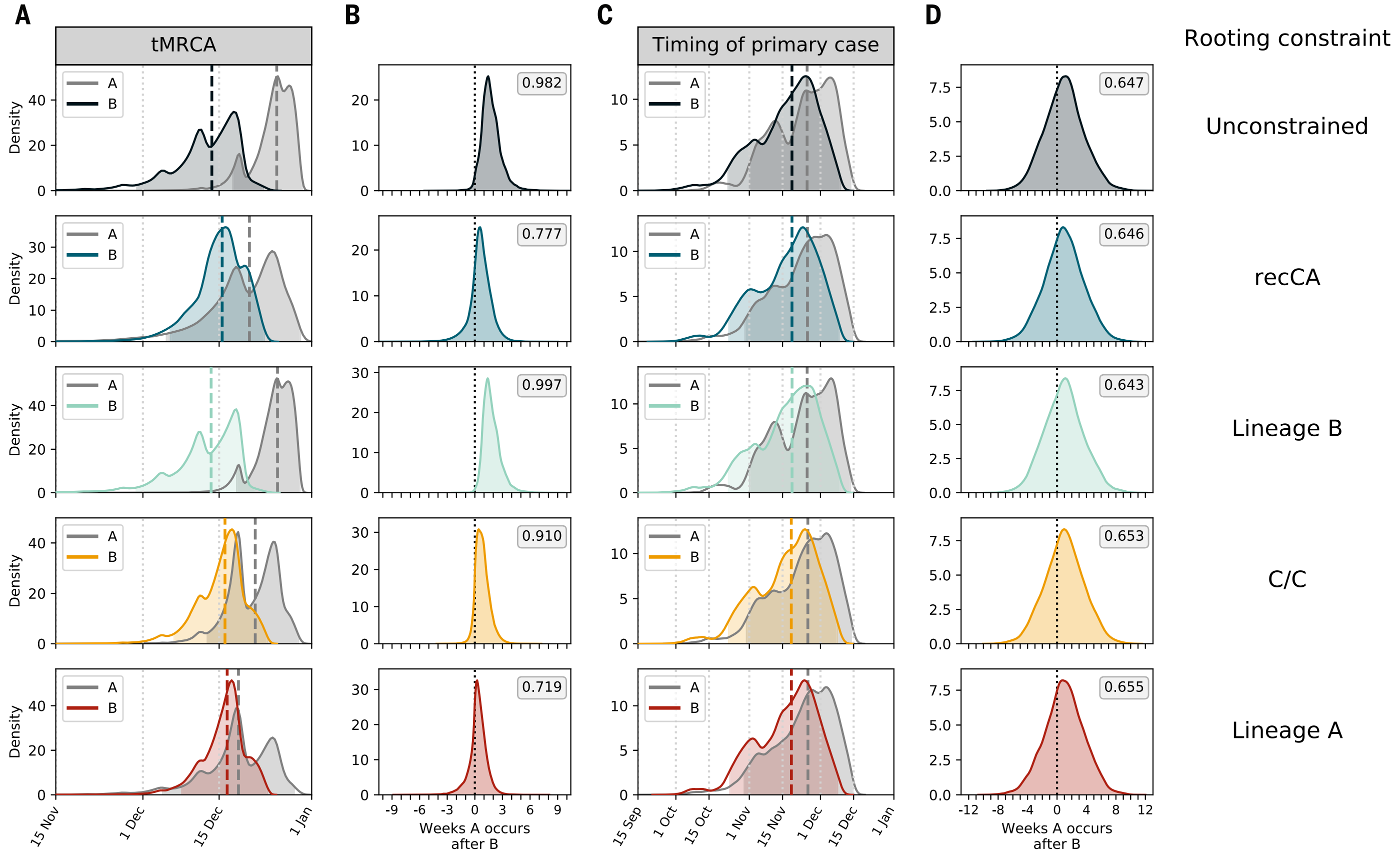

**Fig. 3. Comparison of the tMRCA and primary case dates for lineage A and lineage B in late 2019 across rooting strategies.** Each row represents a different rooting constraint in phylodynamic analysis, with lineage B, C/C, and lineage A representing a fixed ancestral haplotype. (**A**) The tMRCA for 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 indicates 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 gray box in each graph in (B) and (D).

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 inferred substantial support favoring separate introductions of lineages A and B (BF = 4.3 and BF = 4.2 by using the recCA and unconstrained rooting, respectively) [supplementary materials (SM), materials and 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 (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 that combines epidemic simulations and phylodynamic tMRCA inference (SM materials and 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 that this individual had a lineage B virus (supplementary text) because 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 for which 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 strategy and fixing the plausible ancestral haplotype to lineage A,

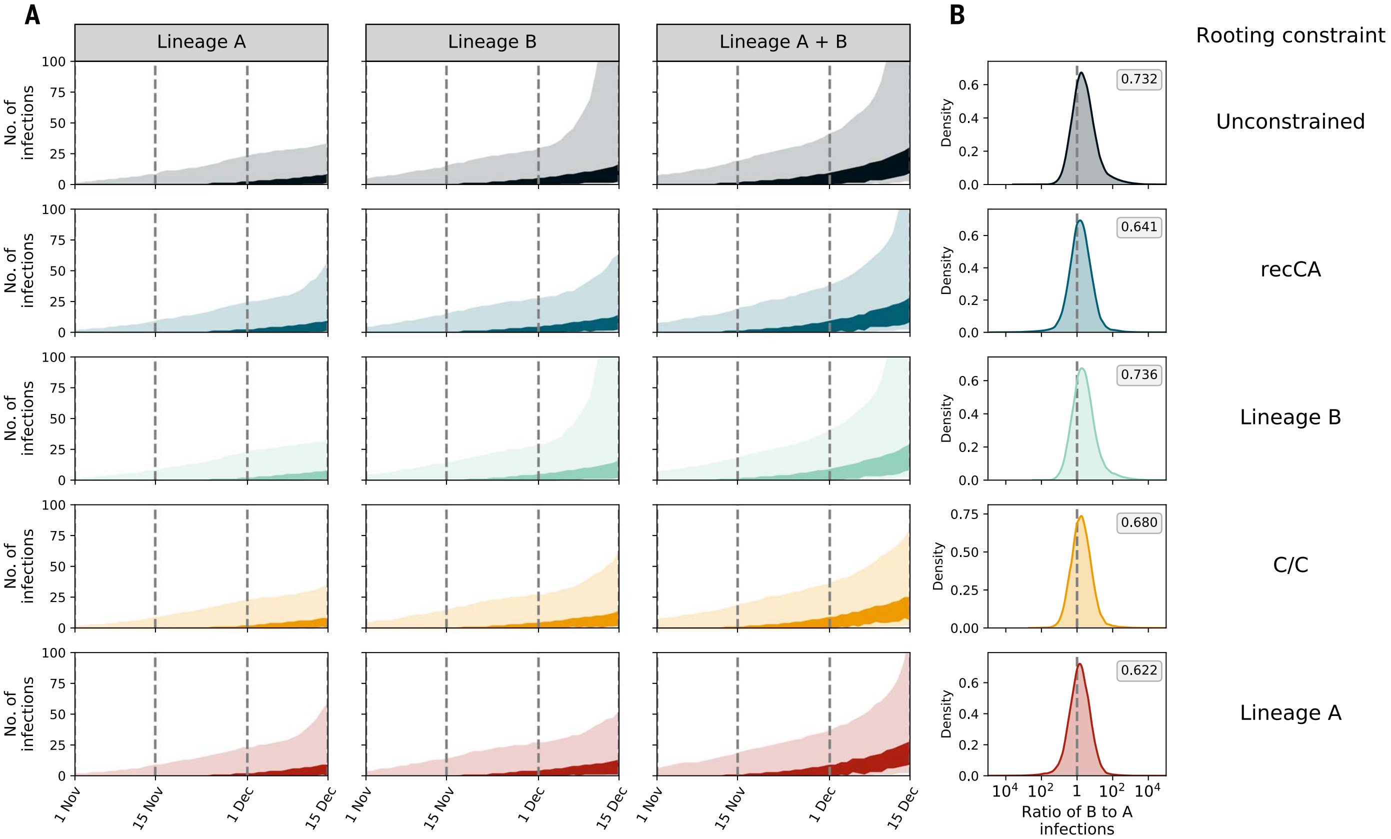

**Fig. 4. Dynamics of simulated SARS-CoV-2 epidemics resulting from separate introductions of lineages A and B in late 2019.** Each row represents a different rooting constraint in phylodynamic analysis, with lineage B, C/C, and lineage A representing a fixed ancestral haplotype. (**A**) Estimated number of infections. The header of each column indicates whether the number of infections is caused by lineage A, lineage B, or the two lineages combined. Darker and lighter shading indicates 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 gray box in each graph.

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 found that an earlier introduction of lineage B led 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 observed 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 to 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 that there were any COVID-19–related hospitalizations before December (*36*) because the simulated epidemics show a median of zero (95% HPD, 0 to 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 before December 2019 (table S13).

### *Additional introductions*

The extinction rate of our simulated epidemics (simulations that did not produce self-sustaining transmission chains) indicate that 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 to 23) introductions led to the establishment of both lineage A and B in humans.

### *Limitations*

Our analysis of the putative intermediate haplotypes suggests that 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. We lack direct evidence of a virus closely related to SARS-CoV-2 in nonhuman 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 before its discovery. Although we simulated epidemics across a range of plausible epidemiological dynamics, our models represent a time frame before the ascertainment of COVID-19 cases and sequencing of SARS-CoV-2 genomes and thus before 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. We resolved this paradox by showing that early SARS-CoV-2 genomic diversity and epidemiology are best explained by at least two separate zoonotic transmissions, in which lineage A and B progenitor viruses were both circulating in nonhuman mammals before 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 intraspecies 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 that established 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, Middle East respiratory syndrome coronavirus (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 that 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 (*55*) 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 (*56*), 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 by using MAFFT v7.453 (*57*) 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.* (*58*). 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 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 nonrecombinant regions (*13*) with IQ-TREE v2.0.7 (*59*), (ii) inferred the sequence of the ancestor of SARS-CoV-2 in each tree with TreeTime v0.8.1 (*60*), 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 (*61*) with the 787-taxon dataset to infer the ancestral haplotype and the tMRCA of SARS-CoV-2 (and the tMRCAs of lineages A and B), using a nonreversible 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*, *62*) using a scale-free network of 5 million individuals and a customized extension of the SAPHIRE model (*63*), producing coalescent trees on which we simulated mutations. We calculated the BF comparing the support of two introductions of SARS-CoV-2 with 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 the number of introductions.

We connected the phylodynamic inference and epidemic simulations by means of 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 (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, *Lancet Infect. Dis.* **20**, 533–534 (2020).
2. L.-L. Ren *et al.*, *Chin. Med. J.* **133**, 1015–1024 (2020).
3. H. Ritchie *et al.*, *Our World in Data* (2022); https://ourworldindata.org/covid-deaths.
4. A. Rambaut *et al.*, *Nat. Microbiol.* **5**, 1403–1407 (2020).
5. F. Wu *et al.*, *Nature* **579**, 265–269 (2020).
6. R. Lu *et al.*, *Lancet* **395**, 565–574 (2020).
7. S. Lytras *et al.*, *Genome Biol. Evol.* **14**, evac018 (2022).
8. M. Worobey, *Science* **374**, 1202–1204 (2021).
9. R. F. Garry, *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 *et al.*, *Virological* (2020); https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
11. M. Worobey *et al.*, *Science* **370**, 564–570 (2020).
12. J. O. Wertheim, M. Steel, M. J. Sanderson, *Syst. Biol.* **71**, 426–438 (2022).
13. S. Temmam *et al.*, *Nature* **604**, 330–336 (2022).
14. J. B. Pease, M. W. Hahn, *Evolution* **67**, 2376–2384 (2013).
15. J. Ratcliff, P. Simmonds, *Virology* **556**, 62–72 (2021).
16. P. Simmonds, *MSphere* **5**, e00408-20 (2020).
17. P. Simmonds, M. A. Ansari, *PLOS Pathog.* **17**, e1009596 (2021).
18. P. Forster, L. Forster, C. Renfrew, M. Forster, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 9241–9243 (2020).
19. J. D. Bloom, *Mol. Biol. Evol.* **38**, 5211–5224 (2021).
20. M. A. Caraballo-Ortiz *et al.*, *Bioinformatics* **38**, 2719–2726 (2022).
21. S. Kumar *et al.*, *Mol. Biol. Evol.* **38**, 3046–3059 (2021).
22. N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, S. Mirarab, *Bioinformatics* **35**, 1852–1861 (2019).
23. J. Pekar, M. Worobey, N. Moshiri, K. Scheffler, J. O. Wertheim, *Science* **372**, 412–417 (2021).
24. S. Hsiang *et al.*, *Nature* **584**, 262–267 (2020).
25. A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, D. Sledge, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 16732–16738 (2020).
26. S. Sanche *et al.*, *Emerg. Infect. Dis.* **26**, 1470–1477 (2020).
27. T. Bedford *et al.*, *Science* **370**, 571–575 (2020).
28. M. Zeller *et al.*, *Cell* **184**, 4939–4952.e15 (2021).
29. C. Alteri *et al.*, *Nat. Commun.* **12**, 434 (2021).
30. L. du Plessis, O. Pybus, *Virological* (2020); https://virological.org/t/further-musings-on-the-tmrca/340.
31. J. Giesecke, *Lancet* **384**, 2024 (2014).
32. Centers for Disease Control and Prevention (CDC), *MMWR Morb. Mortal. Wkly. Rep.* **52**, 986–987 (2003).
33. A. Marí Saéz *et al.*, *EMBO Mol. Med.* **7**, 17–23 (2015).
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 *et al.*, *Nature* **583**, 437–440 (2020).
36. E. O. Nsoesie, B. Rader, Y. L. Barnoon, L. Goodwin, J. Brownstein, *Dig. Acc. Scholar. Harv.* **2**, 019 (2020).
37. L. Chang *et al.*, *Protein Cell* 10.1093/procel/pwac013 (2022).
38. E. C. Holmes *et al.*, *Cell* **184**, 4848–4856 (2021).
39. M. Worobey *et al.*, *Science* **377**, 951–959 (2022).
40. X. Xiao, C. Newman, C. D. Buesching, D. W. Macdonald, Z.-M. Zhou, *Sci. Rep.* **11**, 11898 (2021).
41. C. M. Freuling *et al.*, *Emerg. Infect. Dis.* **26**, 2982–2985 (2020).
42. S. M. Porter, A. E. Hartwig, H. Bielefeldt-Ohmann, A. M. Bosco-Lauth, J. Root, bioRxiv 478082 [Preprint] (2022); https://doi.org/10.1101/2022.01.27.478082.
43. G. Gao *et al.*, *Research Square* [Preprint] (2022); https://doi.org/10.21203/rs.3.rs-1370392/v1.
44. L. du Plessis *et al.*, *Science* **371**, 708–712 (2021).
45. Chinese SARS Molecular Epidemiology Consortium, *Science* **303**, 1666–1669 (2004).
46. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, *eLife* **7**, e31257 (2018).
47. J. A. Lednicky *et al.*, *Nature* **600**, 133–137 (2021).
48. B. Kan *et al.*, *J. Virol.* **79**, 11892–11900 (2005).
49. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, *Nat. Med.* **26**, 450–452 (2020).
50. V. L. Hale *et al.*, *Nature* **602**, 481–486 (2022).
51. J. C. Chandler *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2114828118 (2021).
52. L. Lu *et al.*, *Nat. Commun.* **12**, 6802 (2021).
53. B. B. Oude Munnink *et al.*, *Science* **371**, 172–177 (2021).
54. S. V. Kuchipudi *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2121644119 (2022).
55. H.-L. Yen *et al.*, *Lancet* **399**, 1070–1078 (2022).
56. Y. Shu, J. McCauley, *Euro Surveill.* **22**, 30494 (2017).
57. K. Katoh, D. M. Standley, *Mol. Biol. Evol.* **30**, 772–780 (2013).
58. N. De Maio *et al.*, *Virological* (2020); https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480.
59. B. Q. Minh *et al.*, *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
60. P. Sagulenko, V. Puller, R. A. Neher, *Virus Evol.* **4**, vex042 (2018).
61. M. A. Suchard *et al.*, *Virus Evol.* **4**, vey016 (2018).
62. 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.
63. X. Hao *et al.*, *Nature* **584**, 420–424 (2020).
64. J. E. Pekar *et al.*, Code for: The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. Zenodo (2022); doi: 10.5281/zenodo.6585475.
65. J. E. Pekar *et al.*, Data for: The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. Zenodo (2022); doi: 10.5281/zenodo.6887186.

## ACKNOWLEDGMENTS

We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories, who generated and shared through GISAID the viral genomic sequences and metadata on which this research is based (data S1) (*57*). We are greatly appreciative toward L. Chen, D. Liu, and Y. Yan for providing insight into the putative intermediate genomes and clarification regarding the relative sequencing depth at positions 8782 and 28144, M. Eloit and S. Temmam for sharing their sarbecovirus dataset and recombination analysis results, and M. 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 75N93021C00015 (M.W.). J.E.P. acknowledges support from NIH (T15LM011271). N.M. acknowledges support from the National Science Foundation (NSF) (NSF-2028040). J.I.L. acknowledges support from NIH (5T32AI007244-38). J.O.W. acknowledges support from NIH (R01AI135992 and R01AI136056). R.F.G. is supported by NIH (R01AI132223, R01AI132244, U19AI142790, U54CA260581, U54HG007480, and OT2HL158260), the Coalition for Epidemic Preparedness Innovation, the Wellcome Trust Foundation, Gilead Sciences, and the European and Developing Countries Clinical Trials Partnership Programme. M.A.S. and A.R. acknowledge the support from the Wellcome Trust (Collaborators Award 206298/Z/17/Z–ARTIC network), the European Research Council (grant agreement 725422–ReservoirDOCS), and NIH (R01AI153044). K.G.A. is supported by NIH (U19AI135995, U01AI151812, and UL1TR002550). E.C.H. is funded by an Australian Research Council Laureate Fellowship (FL170100022). J.L., H.P., and M.-S.P. 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). T.I.V. 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 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: J.E.P., M.A.S., K.G.A., M.W., and J.O.W. Methodology: J.E.P., A.M., N.M., M.A.S., K.G.A., M.W., and J.O.W. Software: J.E.P., A.M., N.M., K.G., and M.A.S. Validation: J.E.P., A.M., K.I., K.G., and M.A.S. Formal analysis: J.E.P., A.M., E.P., K.I., J.L.H., K.G., and J.O.W. Investigation: J.E.P., A.M., E.P., K.I., J.L.H., K.G., and J.O.W. Resources: M.A.S., K.G.A., and J.O.W. Data curation: J.E.P., E.P., K.G., M.Z., J.C.W., S.H., J.L., H.P., M.-S.P., K.C.Z.Y., R.T.P.L., M.N.M.I., Y.M.N., and J.O.W. Writing – original draft preparation: J.E.P., M.W., and J.O.W. Writing – review and editing: All authors. Visualization: J.E.P., J.L.H., K.G., and L.M.M.S.; Supervision: M.A.S, K.G.A., M.W., and J.O.W.; Project administration: M.A.S., K.G.A., M.W., and J.O.W.; Funding acquisition: M.A.S., K.G.A., M.W., and J.O.W. **Competing interests:** J.O.W. has received funding from the CDC (ongoing) through contracts or agreements to his institution unrelated to this research. M.A.S. 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. R.F.G. is cofounder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. M.W., E.C.H., A.R., M.A.S., J.O.W., and K.G.A. 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 (*64*). The following data are available on Zenodo (*65*): recCA sequence, BEAST phylogenetic inference output, and simulation and rejection sampling output for the primary analysis. **License information:** This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abp8337
Materials and Methods
Supplementary Text
Figs. S1 to S31
Tables S1 to S15
References (*66*–*109*)
MDAR Reproducibility Checklist
Data S1 to S3

View/request a protocol for this paper from *Bio-protocol*.

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

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

Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich, Jennifer L. Havens, Karthik Gangavarapu, Lorena Mariana Malpica Serrano, Alexander Crits-Christoph, Nathaniel L. Matteson, Mark Zeller, Joshua I. Levy, Jade C. Wang, Scott Hughes, Jungmin Lee, Heedo Park, Man-Seong Park, Katherine Ching Zi Yan, Raymond Tzer Pin Lin, Mohd Noor Mat Isa, Yusuf Muhammad Noor, Tetyana I. Vasylyeva, Robert F. Garry, Edward C. Holmes, Andrew Rambaut, Marc A. Suchard, Kristian G. Andersen, Michael Worobey, and Joel O. Wertheim

*Science* **377** (6609), .  DOI: 10.1126/science.abp8337

**Pandemic epicenter**
As 2019 turned into 2020, a coronavirus spilled over from wild animals into people, sparking what has become one of the best documented pandemics to afflict humans. However, the origins of the pandemic in December 2019 are controversial. Worobey *et al*. amassed the variety of evidence from the City of Wuhan, China, where the first human infections were reported. These reports confirm that most of the earliest human cases centered around the Huanan Seafood Wholesale Market. Within the market, the data statistically located the earliest human cases to one section where vendors of live wild animals congregated and where virus-positive environmental samples concentrated. In a related report, Pekar *et al*. found that genomic diversity before February 2020 comprised two distinct viral lineages, A and B, which were the result of at least two separate cross-species transmission events into humans (see the Perspective by Jiang and Wang). The precise events surrounding virus spillover will always be clouded, but all of the circumstantial evidence so far points to more than one zoonotic event occurring in Huanan market in Wuhan, China, likely during November–December 2019. —CA

**View the article online**
https://www.science.org/doi/10.1126/science.abp8337
**Permissions**
https://www.science.org/help/reprints-and-permissions

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