SARS-CoV-2 emergence very likely resulted from at least two zoonotic events

Authors: 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,*], and Joel O. Wertheim[14,*]

*Corresponding authors. Email: jepekar@ucsd.edu (JEP), msuchard@ucla.edu (MAS), andersen@scripps.edu (KGA), worobey@arizona.edu (MW); jwertheim@health.ucsd.edu (JOW)

Short Title: Repeated SARS-CoV-2 zoonoses

Affiliations:
[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, Germantown, MD 20876, USA.

[17]Global Virus Network (GVN), Baltimore, MD, 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

Abstract: Understanding the circumstances that lead to pandemics is critical to their prevention. Here, we analyze the pattern and origin of genomic diversity of SARS-CoV-2 early in the COVID-19 pandemic. We show that the SARS-CoV-2 genomic diversity prior to February 2020 comprised only two distinct viral lineages—denoted A and B—with no transitional haplotypes. Novel phylodynamic rooting methods, coupled with epidemic simulations, indicate that these two lineages were the result of at least two separate cross-species transmission events into humans. The first zoonotic transmission likely involved lineage B viruses and occurred in late-November/early-December 2019 and no earlier than the beginning of November 2019, while the introduction of lineage A likely occurred within weeks of the first event. These findings define the narrow window between when SARS-CoV-2 first jumped into humans and when the first cases of COVID-19 were reported. Hence, as with SARS-CoV-1 in 2002 and 2003, SARS-CoV-2 emergence likely resulted from multiple zoonotic events.

One sentence summary: SARS-CoV-2 genomic diversity early in the COVID-19 pandemic points to emergence via repeated zoonotic events.

Introduction

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

Lineage B has been the most common throughout the pandemic and includes all 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') (*4*), sampled on 24 and 26 December 2019, respectively. All published sequences from environmental samples taken at Huanan Market also fall in lineage B. The earliest lineage A viruses, Wuhan/IME-WH01/2019 and Wuhan/WH04/2020, were sampled on 30 December 2019 and 5 January 2020, respectively, and they differ from lineage B by two nucleotide substitutions: C8782T and T28144C (*5*). Notably, the nucleotides at these two positions in lineage A are identical to related viruses of *Rhinolophus* bats (*3*), which are presumed to represent the ultimate host reservoir (*6*). We designate lineage B viruses as having a "C/T" pattern at these key sites (C8782, T28144), and lineage A viruses as having a "T/C" pattern (C8782T, T28144C). The earliest lineage A genomes lack a direct epidemiological connection to the Huanan market, but, importantly, were sampled from individuals who lived or had recently stayed close to the market (*7*). Despite the availability of these data, three central questions remain: if lineage B viruses are more distantly related to sarbecoviruses from *Rhinolophus* bats, (i) why were they detected earlier than lineage A viruses, (ii) why were only lineage B viruses found in humans at the Huanan market, and (iii) why did lineage B predominate early in the pandemic?

Paramount to answering these questions is determining the ancestral haplotype: the genomic sequence characteristics of the most recent common ancestor (MRCA) at the root of the SARS-CoV-2 phylogeny. Although the root has often been inferred with Bayesian and maximum likelihood methods to fall on the branch leading to IPBCAMS-WH-01 (Lineage B) or other early genomes (*8*–*12*), reanalysis of sequence data from the earliest sampled viruses found that three previously reported mutations in IPBCAMS-WH-01 were spurious, and the genome was, in fact, identical to the Hu-1 reference genome (*13*). Other early genomes were also found to have spurious mutations (*13*), thereby decreasing the overall genetic diversity of early SARS-CoV-2 sequence data. This decreased diversity suggests that prior studies, including our own (*9*), may have incorrectly rooted the SARS-CoV-2 phylogeny. Alternatively, it has been suggested that multiple SARS-CoV-2 introductions from an intermediate host led to separate origins of lineages A and B and that the MRCA of SARS-CoV-2 existed in an animal reservoir, rather than in humans (*14*).

In this study, we analyze the evolutionary and epidemiological dynamics at the start of the COVID-19 pandemic. By combining genomic and epidemiological data, novel phylodynamic models, and epidemic simulations, we eliminate 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 no earlier than November 2019, with a lineage A virus jumping into humans after the introduction of a lineage B virus.

Sequencing artifacts inflate estimates of early SARS-CoV-2 genetic diversity

There are 787 near-full length genomes from lineages A and B sampled by 14 February 2020 available on GISAID (Data S1, S2). However, there are also 20 genomes of intermediate haplotypes from this time period containing either T28144C or C8782T but not both mutations. We refer to these haplotypes as C/C or T/T because they have the same nucleotide at these two key sites. These 'intermediate' genomes present a challenge to the hypothesis that lineages A and B were separate introductions, as their existence suggests there was within-human evolution of one lineage towards the other via a C/C or T/T transitional form. However, rather than representing transitional states, these apparent intermediates could also arise from issues, including contamination, or artifacts from sequencing or bioinformatics, leading to erroneous consensus genomes (*15*), as shown in our previous analysis on the emergence of SARS-CoV-2 in North America (*16*).

We identified multiple instances of C/C and T/T haplotypes 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.

Of the 16 C/C genomes in our data set, we found four that share nucleotide substitutions—other than T28144C—found in some lineage A genomes (Fig. 1A). If the C/C intermediates actually existed, 11 of the 19 additional unique mutations in the C/C genomes would need to be homoplasies: identical mutations arising separately in the C/C and lineage A genomes. For example, a C/C genome from Anhui province (EPI_ISL_1069206) shares A11430G with 6 lineage A genomes sampled across China, and a genome from Sichuan province (EPI_ISL_451325) shares C1342T and C18060T with three lineage A genomes. The authors of the latter example confirmed that low sequencing depth at position 8782 led to the erroneous calling of the reference genome nucleotide at this position in this genome (L. Chen Personal Communication). Furthermore, these authors confirmed that incorrect base calls, often due to low sequencing depth, led to erroneous assignment of 11 additional C/C genomes sampled in Wuhan and Sichuan province (four of which share substitutions with lineage B genomes, see below).

A similar pattern was observed in the five C/C genomes sharing substitutions found within lineage B (Fig. 1B), including a South Korean genome (EPI_ISL_413017) sharing G26640T, G26144T, and T26677C with another lineage B genome from South Korea. In this instance, we confirmed that low sequencing depth at position 28144 (<10x) resulted in this erroneous assignment. Critically, therefore, we are able to explain all C/C genomes as artifactual, with the exception of two genomes sampled in Beijing in late January and early February, whose additional mutations were not observed in early lineage A or B genomes and whose underlying data was not available.

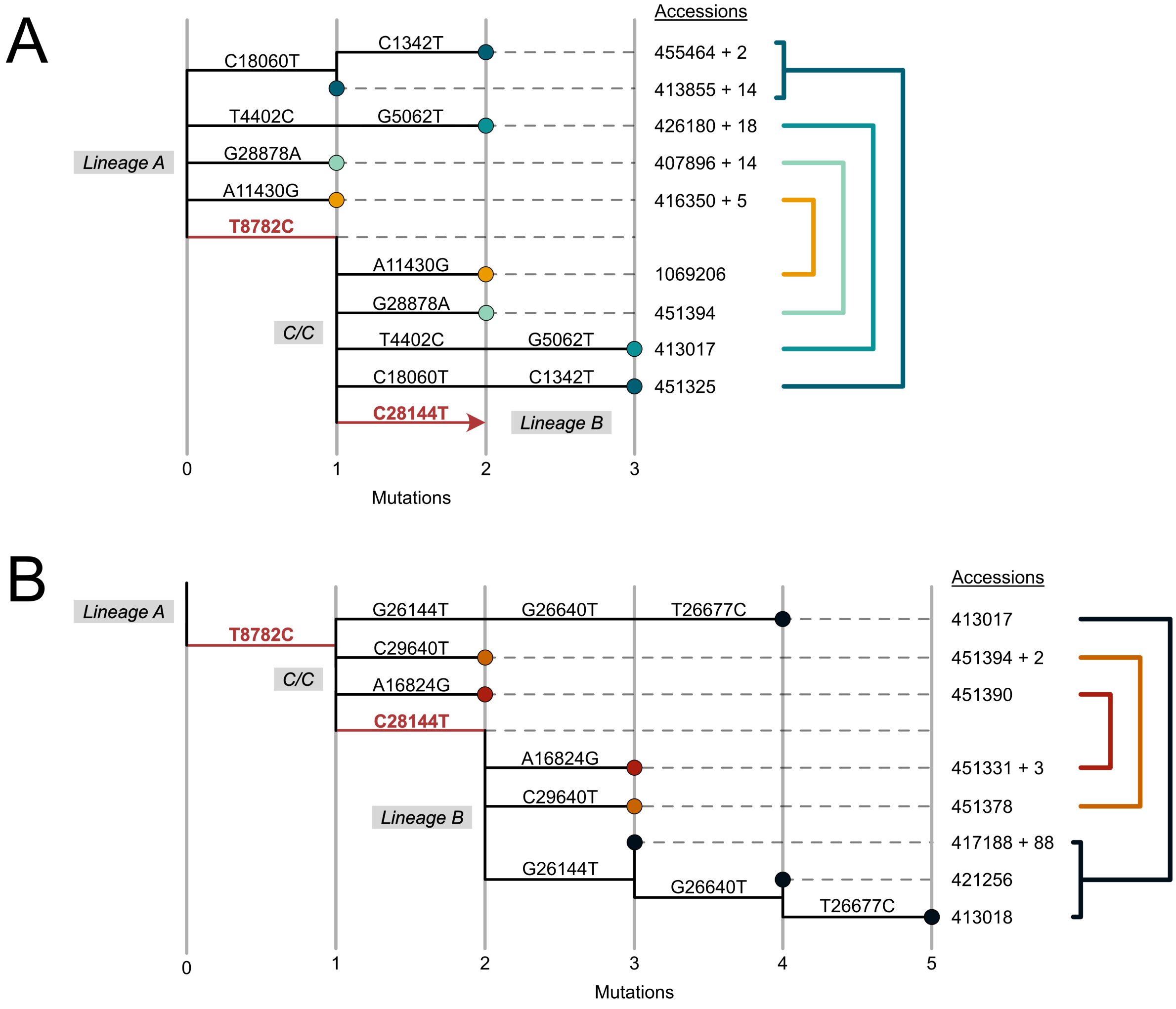

Figure 1. Phylogeny of SARS-CoV-2 intermediate C/C genomes and their shared mutations within lineages A and B. (**A**) Shared mutations across lineage A and C/C. (**B**) Shared mutations across lineage B and C/C. Mutations relative to the Hu-1 reference genome are shown above each branch. Lineage-defining mutations (8782 and 28144) are colored in red. Derived mutations not shared by both lineages are excluded. The taxon names are GISAID accession numbers, and the total number of additional matching homoplastic sequences are indicated. Sequences that share derived mutations are connected by the lines on the right, and brackets indicate that a group of sequences share the derived mutations that cannot be individually resolved.

Unlike the C/C genomes, none of the four T/T genomes shared additional mutations with lineage A or B genomes that would clarify their veracity. However, we confirmed that the T/T genome sampled in Singapore on 14 February 2020 (EPI_ISL_462306) had low coverage at both 8782 and 28144 (≤10x). Moreover, the 3 T/T genomes sampled in Wuhan on 26 January (EPI_ISL_493179, EPI_ISL_493182, EPI_ISL_493180) had low sequencing depth and indeterminate C/T nucleotide assignment at position 8782 (Table S1). These findings suggest all T/T genomes sampled by 14 February 2020 are similarly artifactual.

C/C and T/T genomes continue to be observed throughout the pandemic, but are either the result of random sequencing error or convergent evolution. Examples of convergent evolution at C8782T producing T/T can be seen aboard the Diamond Princess cruise ship outbreak and subsequent COVID-19 waves in New York City and San Diego (Fig. S1-S6, supplementary text).

These findings cast substantial 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.

Reconstructing the ancestral bat sarbecovirus

SARS-CoV-2 genomes sampled in 2019 at the Huanan market are more genetically distant to related bat sarbecoviruses than SARS-CoV-2 genomes identified by February 2020 (*3*, *8*, *13*, *17*, *18*), raising the question of whether the SARS-CoV-2 lineage with the greatest sequence similarity to the bat sarbecoviruses represents the ancestral haplotype that first circulated in humans. To answer this question, we reconstructed a hypothetical progenitor of SARS-CoV-2 across 15 non-recombinant regions from closely related non-human sarbecovirus genomes (*19*) (Fig. 2, S7). Specifically, we performed maximum likelihood ancestral state reconstruction to infer the sequence of the MRCA of SARS-CoV-2 and its most closely related bat virus in each non-recombinant region and then concatenated the inferred sequences to create the genome of the recombinant common ancestor (‘recCA’) (see Methods). The recCA is more informative than an outgroup sarbecovirus because it accounts for the closest relative across all recombinant segments and, as an internal node on the phylogeny, is more genetically similar to SARS-CoV-2 than any extant sarbecovirus.

The recCA differed from Hu-1 by just 381 substitutions, including C8782T and T28144C. Here, we define mutations away from the Hu-1 reference genome toward the recCA, such as C8782T and T28144C, as reversions. In this manner, lineage A (exemplified by Wuhan/WH04/2020) has two reversions, and hence has two fewer substitutions separating it from the recCA than lineage B. (Note that although these mutations are nominally ‘reversions’, if the true MRCA of SARS-CoV-2 were a lineage A virus, those lineage B to lineage A reversions would not actually have occurred). Additional reversions, C18060T and C29095T, have been separately identified in USA/WA1/2020 and Guangdong/20SF012/2020, respectively, and it has been argued that these haplotypes are the ancestral form of SARS-CoV-2 (*17*, *18*). We find that repeated substitutions at sites 8782, 18060, and 28144, are common among closely related sarbecoviruses (Fig. S11-13). In contrast, 29095 is strongly conserved among these sarbecoviruses but highly polymorphic in humans (Fig. S14). Absent from the recCA are two mutations, C2416T and C23929T, previously suggested to have been present in the immediate ancestor of SARS-CoV-2 (*18*); these mutations occur on the branch to the related bat sarbecovirus RaTG13 (Fig. S8-10) (*20*).

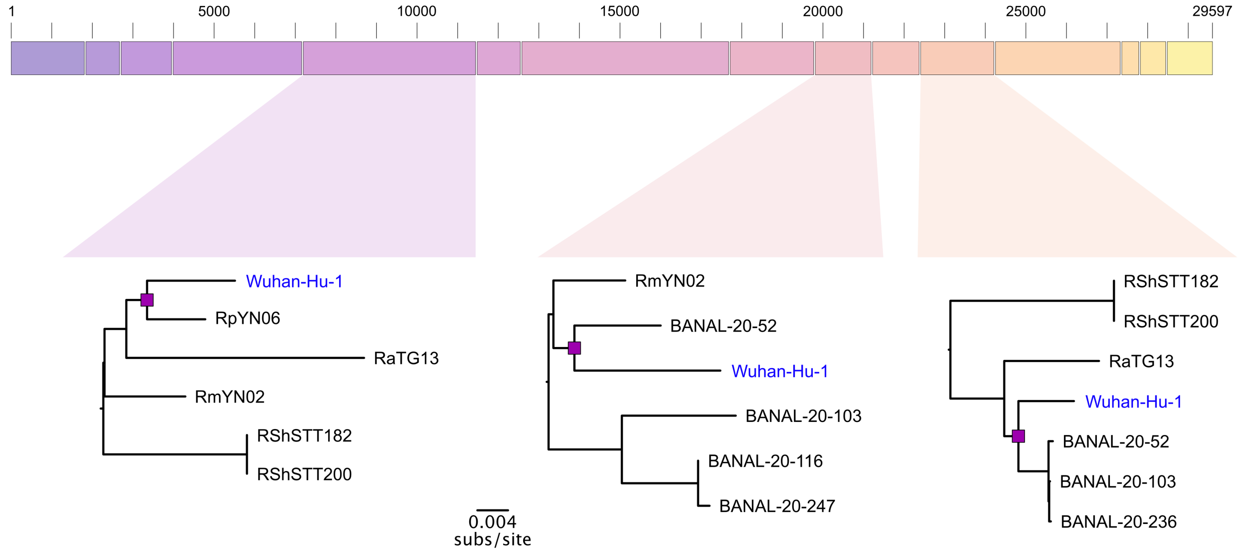

Figure 2. Reconstructing the recombinant common ancestor (recCA) of SARS-CoV-2 infecting a non-human animal. The figure identifies 15 non-recombinant regions of SARS-CoV-2-like sarbecovirus genomes. Subtrees from sarbecovirus maximum likelihood phylogenies of example regions 5, 9, and 11 show the genomes most closely related to SARS-CoV-2. Ancestral state reconstruction at the MRCA (purple square) of SARS-CoV-2 (Wuhan-Hu-1) and the most closely related sarbecovirus for each of the 15 fragments is concatenated to construct the recCA.

Frequent reversions across the SARS-CoV-2 phylogeny

We find that the ubiquity of reversion mutations within SARS-CoV-2 indicates that genetic similarity to the recCA 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. 3). Reversions were overrepresented among the 381 sites that separate the recCA from Hu-1 (23/381 = 6.04%), compared with substitutions at all other sites (631/29,134 = 2.17%).

Most of these reversions are C-to-T mutations (19/23 = 82.6%), matching the mutational bias of SARS-CoV-2 (*21–23*). 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, sampled from a Malaysian citizen traveling back from Wuhan on 4 February 2020, whose only four mutations from Hu-1 are all reversions (A.1+T6025C) (Fig. 3; Malaysia/MKAK-CL-2020-6430/2020). 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.

Frequent reversions, especially C-to-T reversions, are not restricted to early-2020, as we observe them throughout the pandemic, including in the emergence of WHO-named variants of concern/interest (Fig. S16, S17).

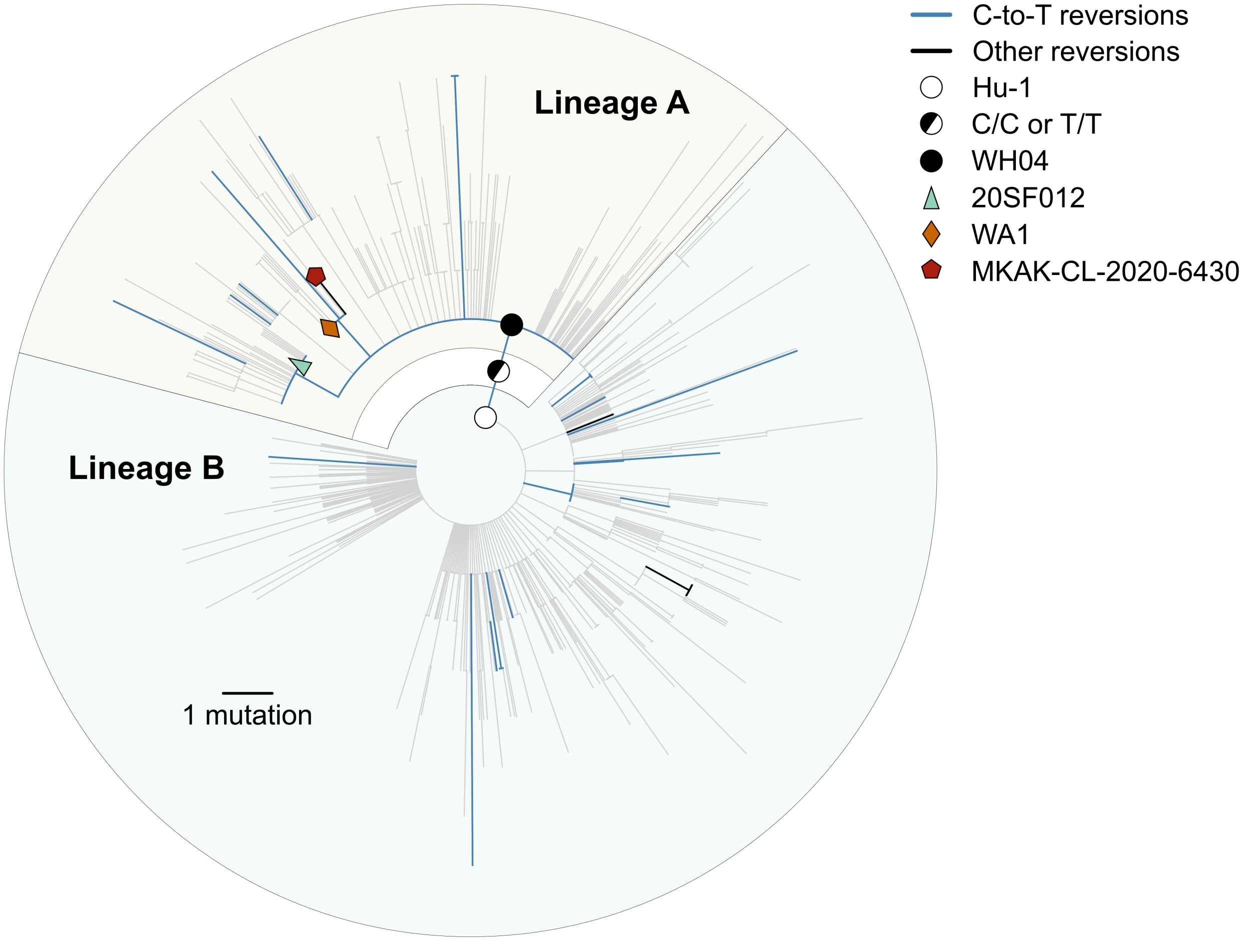

Figure 3. Maximum likelihood phylogeny of the early SARS-CoV-2 pandemic, showing nucleotide reversions and putative candidates for the ancestral haplotype at the most common recent ancestor (MRCA). Putative ancestral haplotypes are identified with colored shapes. Reversions from the Hu-1 reference genotype 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.

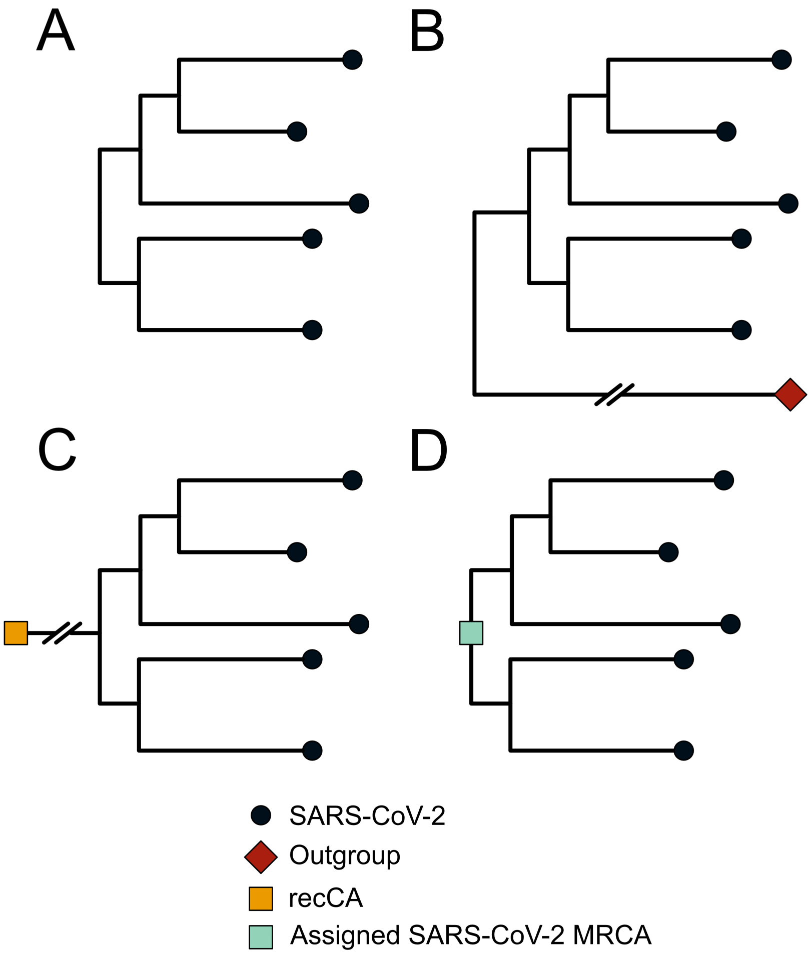

Figure 4. Schematic depicting the rooting strategies used in different phylodynamic models. (**A**) An unconstrained model with only SARS-CoV-2 where the root is inferred from the molecular clock calibrated using SARS-CoV-2 sampling dates. (**B**) An unconstrained model with SARS-CoV-2 and a sarbecovirus outgroup. (**C**) A constrained model where the ancestor of SARS-CoV-2 is constrained to be the recombinant common ancestor (recCA). (**D**) A constrained model with only SARS-CoV-2, but the MRCA forced to be a pre-specified haplotype.

Inferring the MRCA of SARS-CoV-2

To infer the ancestral SARS-CoV-2 haplotype, we developed and implemented a novel 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. S15, supplementary text). Using this model, referred to as the unconstrained model (Fig. 4A), we inferred the ancestral haplotype of the 787 lineages A and B genomes sampled by 14 February 2020; C/C and T/T genomes from this time period were excluded, because we determined they are the result of sequencing artifacts.

Our unconstrained model strongly favors a lineage B or C/C ancestral haplotype, weakly disfavors lineage A [Bayes factor (BF)>10], and strongly disfavors A.1 and A+C29095T (BF>1000) (Table 1). The T/T ancestral haplotype was disfavored (BF>10), likely because of the C-to-T bias present in SARS-CoV-2 (Fig. S15). However, we acknowledge that the earliest sampled lineage B genomes associated with the Huanan market could bias rooting using phylodynamic inference toward ancestral haplotypes matching these genomes. When we repeated this analysis excluding the 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), posterior support for a lineage B ancestral haplotype decreased from 80.85% to 62.96%; support for C/C and lineage A increased from 10.32% and 1.68% to 23.02% and 5.73%, respectively (Table 1). Extending this sensitivity analysis to exclude all genomes from Wuhan (59 lineage B, 37 lineage A), which represent the majority of genomes before mid-January 2020, increased support for a C/C and lineage A ancestral haplotype even further, to 32.02% and 11.03%, respectively. The overwhelming rejection of an A.1 or A+C29095T ancestral haplotype remained unchanged (BF>1000).

Table 1. Posterior probabilities of inferred ancestral haplotype at the MRCA of SARS-CoV-2.

Haplotype[1]	Mutations from Hu-1 reference	Representative genome[2]	Phylodynamic analysis			
			Unconstrained %	No market[3] %	No Wuhan[4] %	recCA %
B (C/T)	N/A	Hu-1	80.85†	62.96†	50.85†	8.18
A (T/C)	C8782T + T28144C	WH04	1.68*	5.73*	11.03	77.28†
C/C	T28144C	N/A	10.32	23.02	32.02	10.49
T/T	C8782T	N/A	0.92*	1.68*	2.02*	3.71*
A+C29025T (T/C)	C8782T + T28144C + C29095T	20SF012	<0.01***	<0.01***	<0.01***	0.20**
A.1 (T/C)	C8782T + T28144C + C18060T	WA1	<0.01***	<0.01***	<0.01***	0.04***

†Haplotype with greatest posterior probability, reference for Bayes Factor (BF)

* BF>10; ** BF>100; *** BF>1000; BFs are in favor of hypothesis rejection

[1] Positions 8782 and 28144 indicated in parentheses.

[2] Genome with sequence matching the haplotype.

[3] Excluding 15 market-associated genomes.

[4] Excluding all 96 genomes from Wuhan.

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 [*e.g.*, RaTG13, as in Fig. 4B and (*8*)], we developed a novel rooting method that assigns the recCA to be the progenitor of the inferred SARS-CoV-2 MRCA (Fig. 4C).

By including a recCA root, the posterior support for a lineage A ancestral haplotype increased to 77.28% and support for lineage B decreased to 8.18% (Table 1). Support for C/C was unchanged. Even though the recCA includes both C18060T and C29095T, present in A.1 and A+C29095T respectively, we can still confidently reject both of these haplotypes as the ancestral haplotype of SARS-CoV-2 when rooting with the recCA (BF>100). Hence, WA1-like and 20SF012-like haplotypes cannot plausibly represent the MRCA of SARS-CoV-2: the similarity of these genomes to the recCA is due to C-to-T reversions. Lastly, haplotypes not present in Table 1 were strongly disfavored as well (Data S3).

In sum, we infer only three plausible ancestral haplotypes: lineage A, lineage B, and C/C. Lineage A is two mutations closer to the recCA than lineage B, and C/C is one mutation closer (Fig. S11, S13). The repeated observation of C8782T convergent evolution and the C-to-T mutational bias suggest a C/C ancestral haplotype is plausible. Therefore, we maintain that lineage A and C/C are more likely than lineage B to be the haplotype at the MRCA of SARS-CoV-2.

Regardless of the ancestral haplotype, we find that the time of the most recent common ancestor (tMRCA) of SARS-CoV-2 is robust. Similar to our novel approach with recCA (Fig. 4C), we assigned different ancestral haplotypes to be the MRCA of SARS-CoV-2 and inferred the tMRCA (*i.e.*, A, B, C/C, A.1 or A+C29095T; as in Fig. 4D). Despite previous suggestions that a phylogenetic root in lineage A would produce older tMRCA estimates than a lineage B rooting (*18*), the tMRCA was consistent across all ancestral haplotypes. (Table S2; supplementary text).

Evidence against a single introduction of SARS-CoV-2

We next sought to determine whether the plausible ancestral haplotypes of SARS-CoV-2 (lineage A, lineage B, or C/C) are consistent with a single zoonotic introduction. We simulated SARS-CoV-2-like epidemics (*9*, *24*) with a doubling time of 2.65 days (95% range across simulations: 1.51-4.14) (*25*, *26*) to account for the rapid spread of SARS-CoV-2 before it was identified as the etiological agent of COVID-19 and prior to the *cordon sanitaire* and lockdown implemented in Wuhan on 23 January 2020 (*27*). We then simulated coalescent processes and viral genome evolution across these epidemics to determine if we could recapitulate the observed viral diversity emanating from a single origin event.

The C/C haplotype at the MRCA is inconsistent with a single origin event. The C/C rooting produces two clades, lineages A and B, each one mutation from the root with no transitional genomes (Fig. 5A). This topology, where there are only two clades of any size, each one mutation from the root, was present in 6.6% of phylogenies from our simulated epidemics. However, both lineages A and B are large clades, comprising 35.2% and 64.8% of the early SARS-CoV-2 genomes, respectively, and the smaller clade in these simulations was rarely this large. If we require our simulated clades to more realistically comprise at least 1% of the taxa, only 3.6% of the simulations match the C/C topology. If we require both clades to comprise ≥30% of the taxa—better reflecting empirical genomic diversity—only 0.3% of the simulations matches the C/C topology. These results indicate that a single introduction of C/C virus would not be expected to give rise to lineages A and B with no surviving ancestral C/C lineages.

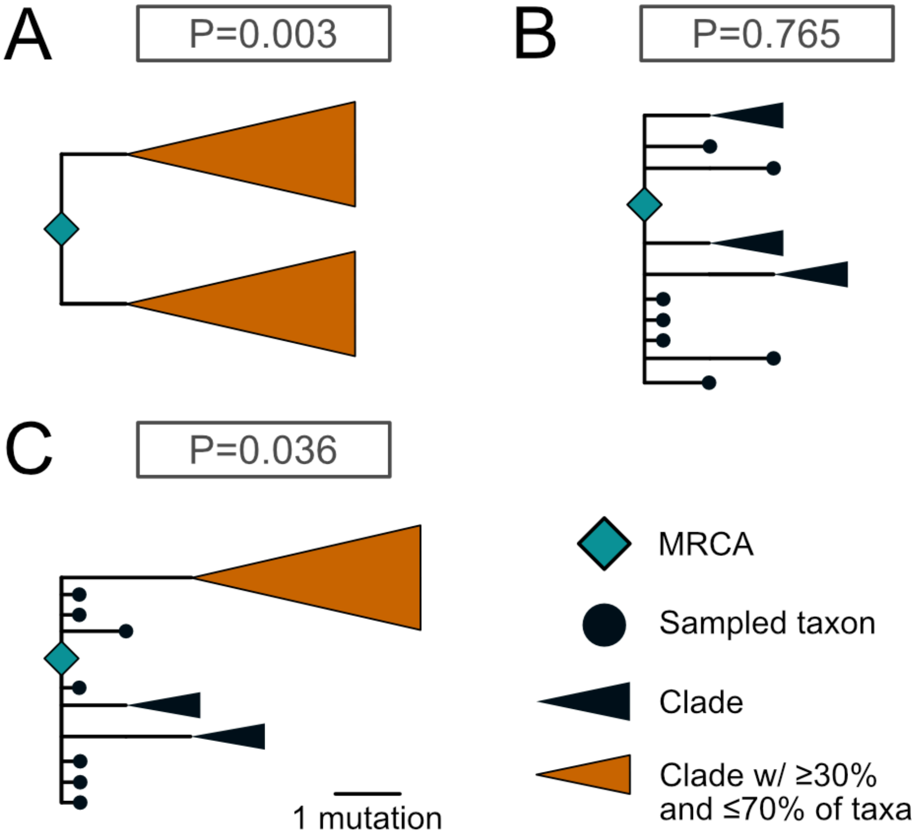

Figure 5. Probability of potential phylogenetic structures arising from a single introduction of SARS-CoV-2 in epidemic simulations. (**A**) Topology matching a C/C ancestral haplotype. (**B**) A large polytomy, consistent with the base of both lineages A and B. (**C**) Topology matching either a lineage A or lineage B ancestral haplotype. Basal taxa have short branch lengths for clarity. The probability of each phylogenetic structure after a single introduction is reported in the box.

The lineage A and lineage B ancestral haplotypes are also inconsistent with a single origin event. Both lineages A and B are characterized by large polytomies: many sampled lineages descending from a single node on the phylogenetic tree. There are 108 and 231 lineages (including basal taxa) descendent from the base of lineages A and B, respectively (Fig. 3). Large polytomies have also been observed for introductions of SARS-CoV-2 into other geographical regions (for example, see Fig. S18) (*16*, *28*, *29*). As expected, these basal polytomies are the most common topology observed in our simulations, present in 76.5% of the simulated epidemics (Fig. 5B).

Rooting on either the lineage A or B haplotype produces a large basal polytomy with the largest clade in the tree separated by two mutations from the root (Fig. 5C). Importantly, our simulations permit these two mutations to occur either within a single individual or during successive infected hosts (*30*). We see a large clade comprising a substantial fraction of the sampled taxa (*i.e.*, between 30% and 70%, reflecting either lineage A or B prevalence) in only 4.8% of the epidemic simulations. When we look for a large clade separated by at least two mutations from a basal polytomy of at least 10 branches—a conservative reflection of the 108- and 231-lineage polytomies characterizing lineages A and B, respectively—we observe this topology in only 3.64% of epidemic simulations (Fig. 5C).

Hence, if SARS-CoV-2 were introduced into humans in a single cross-species transmission event, there is only 0.27% probability of the empirical phylogeny if C/C were the ancestral haplotype and only a 3.64% probability if either lineage A or lineage B were the ancestral haplotype. Additionally, if lineage A were the singular ancestral haplotype, one would expect less genetic divergence from the ancestor in lineage B genomes relative to lineage A genomes than observed, because lineage A viruses would have been evolving at least as long in humans as lineage B viruses; instead, more divergence is seen within lineage B than within lineage A (Fig. S19, S20). Therefore, the phylogenetic structure of SARS-CoV-2 is best explained by separate introductions of lineages A and B, producing the two largest polytomies observed in the early pandemic at the base of each lineage.

The tMRCA of lineage B predates lineage A

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 in an animal reservoir). We inferred the median tMRCA of lineage B to be 13 December (95% HPD: 29 November to 23 December) and the median tMRCA of lineage A to be 25 December (95% HPD: 17 December to 30 December) (Fig. 6A), under the unconstrained model (Fig. 4A). Across the posterior sample of trees, the tMRCA of lineage B consistently occurs earlier than the tMRCA of lineage A (Fig. 6B). These results are robust when rooting with the recCA or constraining the ancestral haplotype to C/C or lineage A (Fig. 6A, 6B, Table S2).

We considered the possibility that the predominance of lineage B viruses in the beginning of the pandemic, particularly at the Huanan market, was biasing the earlier inference of the lineage B tMRCA. However, when we excluded all market-associated genomes, the median tMRCA of lineage B was 17 December (95% HPD: 29 November to 26 December), still earlier than the median tMRCA of lineage A: 25 December (95% HPD: 15 December to 4 January). Therefore, the predominance of lineage B at the Huanan market is not biasing its tMRCA to predate the tMRCA of lineage A.

Timing the primary cases of separate lineage A and B introductions

The primary case—the first human infected with SARS-CoV-2—could precede the tMRCA if basal lineages went extinct during cryptic transmission (*9*, *31*, *32*). Similarly, the primary case is very rarely the first identified case: the index case (*33*, *34*).

We estimated the timing of the lineage B and lineage A primary cases. Our previous analysis (*9*) inferring the timing of the SARS-CoV-2 primary case incorporated the date of index case ascertainment; however, uncertainty regarding the true index case persists (*7*, *13*, *35*) (supplementary text). To overcome this uncertainty, we extend our previously published approach, which combines the epidemic simulations and phylodynamics tMRCA inference (described in the above two sections), to condition the timing of the primary case on both the index case symptom onset date and earliest documented COVID-19 hospitalization date. We included hospitalization dates because they are less susceptible to recall bias than date of symptom onset.

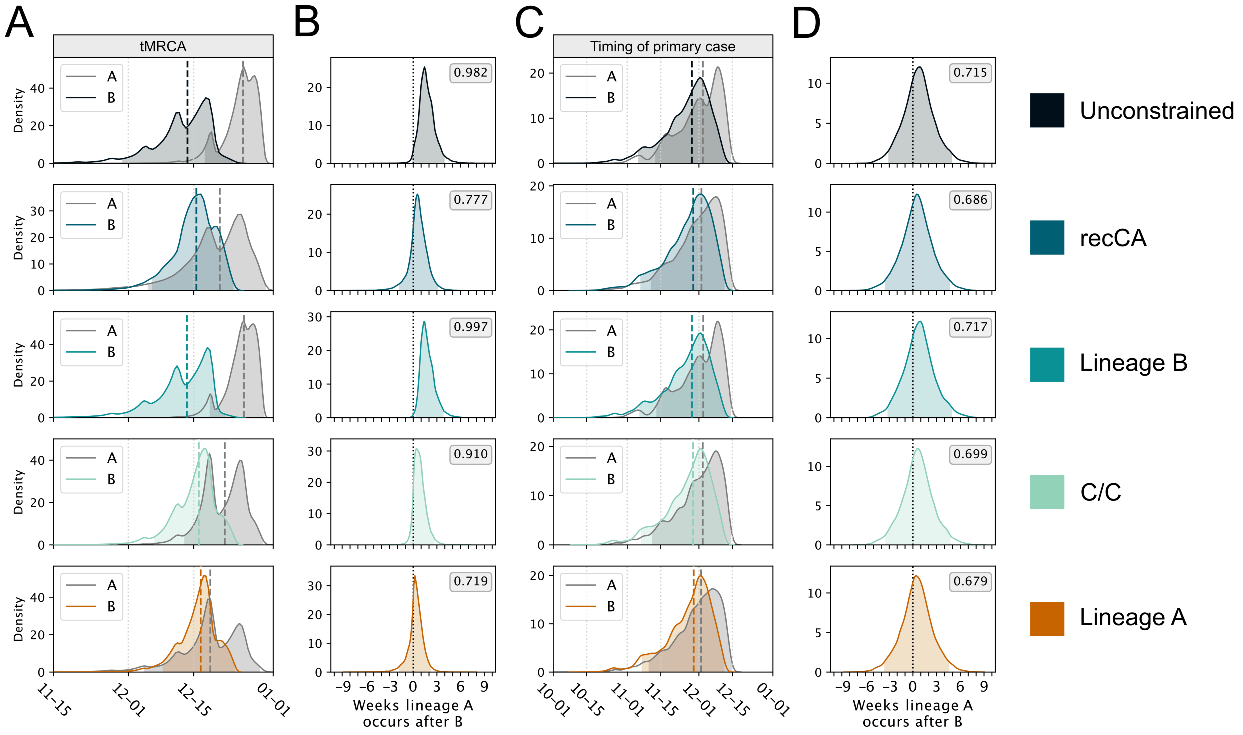

Figure 6. Comparison of the tMRCA and primary case for lineage A and lineage B across rooting strategies. (**A**) The tMRCA for lineages A and B. (**B**) The number of weeks the tMRCA of lineage A occurs after the tMRCA of lineage B. (**C**) The timing of the primary case for lineages A and B. (**D**) 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. The proportion of the posterior with lineage A occurring after lineage B is reported in the grey box. The key denotes the rooting strategy employed.

The earliest confirmed case of COVID-19—with symptom onset date of 10 December and hospitalization on 16 December—was a seafood vendor at the Huanan market; however, this case does not have an associated published genome (*7*). Nonetheless, we can reasonably assume this individual had a lineage B virus, as all human cases associated with the Huanan market and an environmental sample from the stall this vendor operated (EPI_ISL_408512), were lineage B. The earliest confirmed lineage A case, who does have an associated genome, had a symptom onset date of 26 December. However, the spouse of the first confirmed lineage A case was infected as well, becoming symptomatic on 15 December and hospitalized on 25 December (IME-WH01) (*13*). We can therefore reasonably assume the spouse was also infected with a lineage A virus and subsequently infected the earliest confirmed lineage A case (supplementary text). Conditioning on these dates, we inferred the infection date of the lineage B primary case to be 25 November (95% HPD: 4 November to 8 December) and the infection date of the primary case of lineage A to be 2 December (95% HPD: 12 November to 13 December), using the unconstrained model. These estimates were consistent when rooting with the recCA or fixing the plausible ancestral haplotype to lineage A, lineage B, or C/C (Fig. 6C, Table S3, S4).

It is probable that the lineage B primary case predated the lineage A primary case, by approximately 7 days in the unconstrained model (Fig. 6D; Table S5). We observed the lineage B primary case predating that of lineage A in 71.5% of the posterior sample for the unconstrained model. Similar trends are seen when we root on the recCA or force lineage A to be the ancestral haplotype (Fig. 6D; Table S5).

If we discount the possibility that the 10 December case was lineage B, we can instead rely on the earliest case with a confirmed lineage B genome sequence (Wuhan/IPBCAMS-WH-01/2020), who was symptomatic on 13 December and hospitalized on 18 December (*13*) (supplementary text). Conditioning on these dates, we inferred the timing of the lineage B primary case to be 28 November 2019 (95% HPD: 5 November to 11 December). If we instead condition on 8 December as the earliest lineage B case date, to match prior reports of an index case (supplementary text), the timing of the lineage B primary case similarly shifts to 25 November (95% HPD: 3 November to 7 December). These results are consistent across the various phylodynamic models used (Table S3, S4) and indicate that slightly earlier index case dates do not substantially affect the inferred timing of the lineage B primary case.

Our lineage A and B primary case inference is robust to foregoing index case dates and conditioning solely on the tMRCA and earliest hospitalization dates, as well as to assuming a slower growth rate at the start of the pandemic (Table S3, S4; supplementary text). Therefore, across multiple phylodynamics models, index case or earliest hospitalization dates, and epidemic doubling times, our results indicate that lineage B was introduced into humans no earlier than November 2019, and lineage A cross-species transmission likely occurred within days to weeks of the first event.

Epidemiological dynamics of multiple zoonotic origins

We then inferred the number of ascertained infections and hospitalizations arising from these separate introductions. We find that the earlier introduction of lineage B leads to a faster rise in lineage B-associated infections, which dominate the simulated epidemics, recapitulating the predominance of lineage B in China in early 2020 (*12*). We observe this pattern regardless of rooting strategy (unconstrained or recCA) or ancestral haplotype (B, A, or C/C) (Fig. 7, Table S6, S7). Similarly, simulated lineage B hospitalizations are more common than those from lineage A through January 2020 across all ancestral haplotypes, matching empirical data (Fig. S21). This inference is also robust to assuming a slower doubling time (Fig. S22, S23).

Counterintuitively, our modeling shows that separate lineage A and B introductions do not lead to substantially more infections or hospitalizations overall, early in the pandemic, compared with a single introduction. Due to the exponential nature of epidemic growth, cases emanating from a later introduction will be dwarfed by those from the first introduction (Fig. 7; supplementary text).

Additionally, if we assume a single introduction during our modeling (Fig. S24), we observe a similar number of infections and hospitalizations as multiple introductions (Fig. 7, S21; supplementary text).

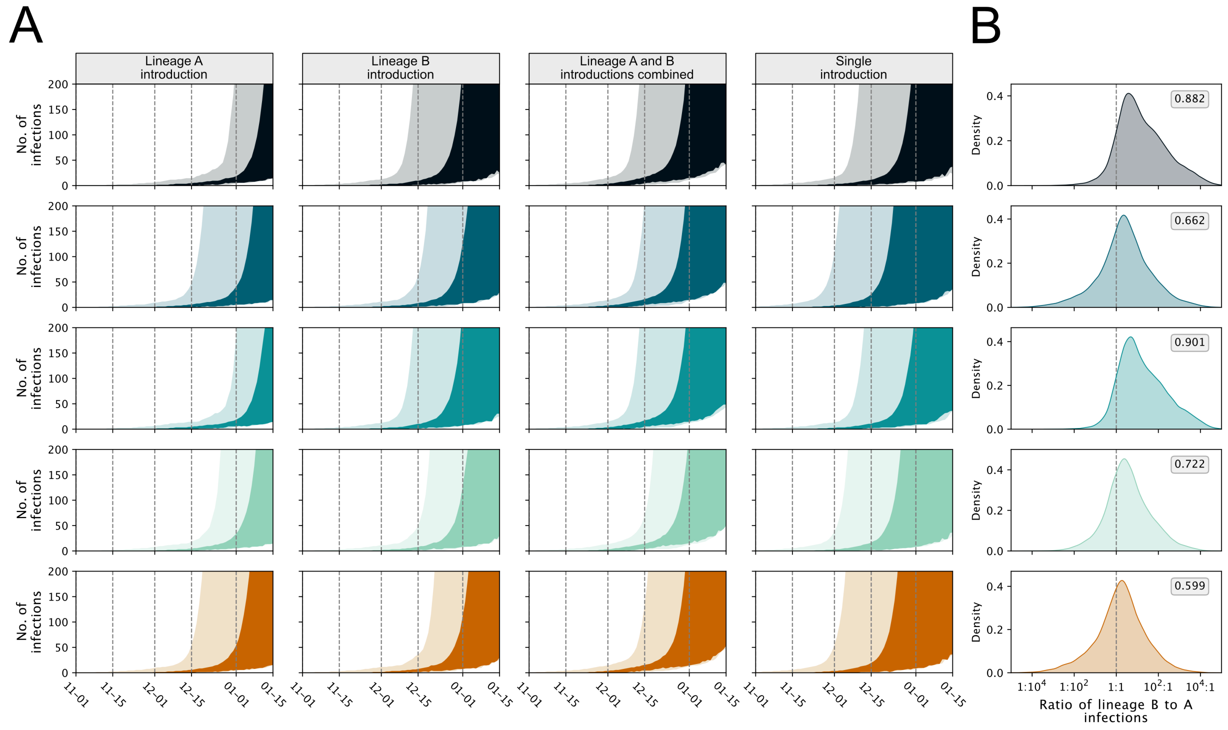

Figure 7. Dynamics of SARS-CoV-2 infections resulting from separate introductions of lineages A and B. (**A**) Estimated number of infections. The header of each column indicates whether the infections are caused by lineage A or B in a multi-introduction scenario, the two lineages together in multi-introduction scenario, or by a single introduction. (**B**) The log ratio of lineage B to lineage A infections on 1 January 2020 in a multi-introduction scenario. The proportion of the posterior with more lineage B infections than lineage A is reported in the grey box. Color scheme is the same as Fig. 6.

Minimal cryptic circulation of SARS-CoV-2

We do not see evidence for substantial cryptic circulation before December 2019 in our epidemic simulations (Fig. 7). By the tMRCA of lineage B on 13 December (95% HPD: 29 November to 23 December), 99.1% of our simulated epidemics have only 20 cumulative lineage B infections (supplementary text). Further, our simulated epidemics reject the possibility of substantially elevated hospitalizations before mid-December, with >1 hospitalization on 1 December 2019 occurring in only 1.5% simulated epidemics. These results are thus in accordance with multiple lines of evidence described previously—from thousands of SARS-CoV-2-negative sample from cases of influenza-like illness in Wuhan in late 2019, to SARS-CoV-2 seronegative stored serum samples from HIV patients in Wuhan—that only extremely low levels of SARS-CoV-2 could have been present in Wuhan prior to December 2019 (*13*).

Zoonoses without sustained transmission can be easily missed

Although the empirical genomic diversity is most consistent with two successful introductions (lineages A and B), our simulations indicate there were likely multiple failed introductions of SARS-

CoV-2 that went extinct. The extinction rate of our simulated epidemics (*i.e.*, those simulations that did not produce self-sustaining transmission chains) was approximately 67%, matching our previous results (*9*). Each failed introduction produces a mean of only 1.85 additional infections and only 0.015 hospitalizations, indicating failed introductions would likely go unnoticed. If we treat each SARS-CoV-2 introduction, failed or successful, as a Bernoulli trial and simulate introductions until we see two successfully introduced epidemics, we estimate that 5 (95% HPD: 2-15) total introductions would be required to lead to the establishment of both lineages A and B in humans.

Discussion

The most probable explanation for the introduction of SARS-CoV-2 into humans involves zoonotic jumps from an as-yet undetermined, intermediate host animal at the Huanan Seafood Wholesale Market (*13*, *36*, *37*). However, the genomic diversity of SARS-CoV-2 during the early pandemic presents a paradox: lineage B viruses predominated early in the pandemic, particularly at the Huanan market, but lineage B is two mutations further from bat coronaviruses than the lineage A viruses.

Here, we demonstrate that no virus transitional between lineages A and B has been sampled in humans. Our phylodynamic inference indicates the MRCA of SARS-CoV-2 was either the lineage A, lineage B, or a transitional C/C haplotype. However, our epidemic simulations show that a single zoonotic introduction of any of these plausible ancestral haplotypes is unlikely to produce the two large phylogenetic polytomies separated by two mutations that characterize lineages A and B. Additionally, the ancestral haplotypes for both lineages A and B persisted through the early months of the pandemic (Fig. S20), reminiscent of successful introductions of SARS-CoV-2 into locations outside China early in the pandemic (*16*, *28*, *29*). Therefore, the empirical genomic data is best explained by at least two distinct zoonotic transmissions, whereby lineage A and B viruses both circulated in a non-human host (Fig. 8, S25).

Our findings demonstrate that the MRCA of the pandemic virus phylogeny was most likely in animals. The SARS-CoV-2 phylogeny is therefore reflective of multiple cross-species transmissions, as there were very likely at least two introductions of SARS-CoV-2 into humans. We acknowledge that if the MRCA was in an animal where we have no evidence of a C-to-T mutational bias, the T/T ancestral haplotype is also possible.

Through late-2019 the Huanan market sold animals both susceptible to SARS-CoV-2 infection and capable of transmitting it (*e.g.*, raccoon dogs, *Nyctereutes procyonoides*) (*37*, *38*). The presence of potential animals reservoirs at the Huanan market, coupled with the timing of the lineage B primary case and the exclusive association of the Huanan market human cases with lineage B viruses, strongly suggests that lineage B was an independent jump into humans at the Huanan market in late-November or early-December 2019.

The earliest lineage A genome (IME-WH01) belongs to an individual who resided just to the south of the Huanan market (*7*). The next earliest lineage A genome, and only other from a patient with COVID-19 onset prior to 2020 (WH04) was from a person who stayed at a hotel near the Huanan market (*5*).

Critically, these two cases are more closely positioned to the Huanan market than randomly chosen pairs sampled from the Wuhan population or from COVID-19 cases in Wuhan in early 2020 (*37*). This geographic proximity is consistent with lineage A emerging from the same location as lineage B, but starting at a later point in time. Hence, the Huanan market is the most likely source of the lineage A introduction as well, likely in early-to-mid December 2019.

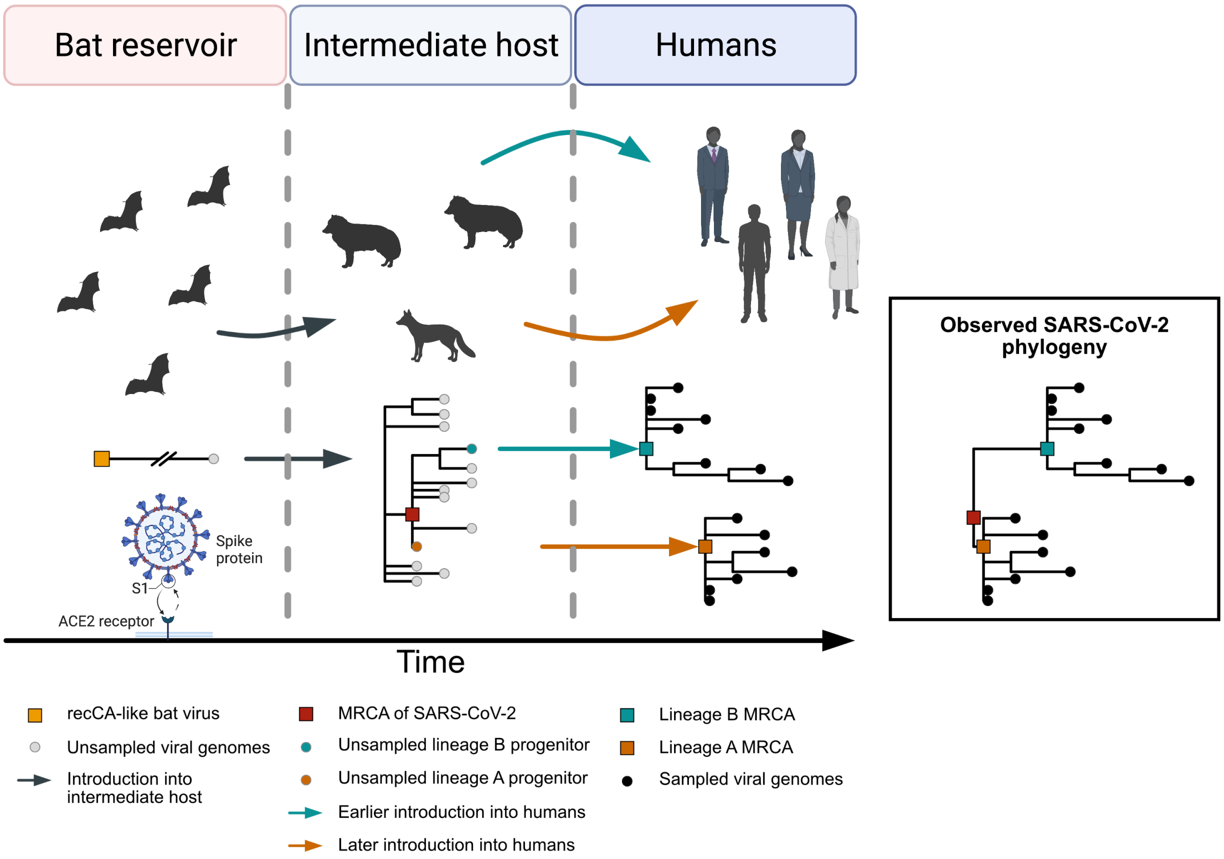
Bat reservoir
Intermediate host
Humans
Observed SARS-CoV-2 phylogeny
Spike protein
S1
ACE2 receptor
Time
■ recCA-like bat virus
○ Unsampled viral genomes
→ Introduction into intermediate host
■ MRCA of SARS-CoV-2
● Unsampled lineage B progenitor
● Unsampled lineage A progenitor
→ Earlier introduction into humans
→ Later introduction into humans
■ Lineage B MRCA
■ Lineage A MRCA
● Sampled viral genomes

Figure 8. Schematic depicting the multiple zoonotic origin of SARS-CoV-2. A recCA-like virus was circulating in bats, and likely after gaining the ability to bind ACE2, jumped into an intermediate host. Therein, lineages A and B appeared and were separately introduced into humans shortly thereafter. An example phylogeny of viruses in the intermediate host is depicted, leading to separate phylogenies for lineages A and B. The resulting SARS-CoV-2 phylogeny from the combined lineage A and B viruses is presented in the black box. This scenario depicts a lineage A ancestral haplotype. See Figure S25 for intermediate and lineage B ancestral haplotypes.

The high extinction rate inherent in SARS-CoV-2 transmission chains reveals that the two zoonotic events establishing lineages A and B were likely accompanied by additional introductions of SARS-CoV-2 that failed to establish in humans. However, these additional 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. 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, at which time the susceptible host animals that had been documented at the market during the previous months were no longer found in the Huanan market (*13*).

Other recent coronavirus epidemics and outbreaks in humans were the result of repeated introductions from animal hosts, including SARS-CoV-1, MERS-CoV, and, most recently, porcine deltacoronavirus in Haiti (*39*–*41*). 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. Like SARS-CoV-1, SARS-CoV-2 is inextricably linked with a wildlife market (*13*, *37*). However, the genomic diversity within the putative SARS-CoV-2 animal reservoir at the Huanan market was almost certainly more shallow relative to diversity seen in SARS-CoV-1 and MERS-CoV reservoirs (*39*, *40*, *42*). As a result, lineages A and B had nearly identical haplotypes. As seen with the early imports of SARS-CoV-2 to Washington State, two genetically similar clades can have their MRCA elsewhere, in the case of SARS-CoV-2 in an intermediate host reservoir (Fig. 8; Fig. S18, supplementary text) (*16*).

Successful transmission of both lineage A and B viruses after independent zoonotic events indicates that evolutionary adaptation within humans was evidently not needed for SARS-CoV-2 to spread (*43*). Much of the necessary evolution for infection of and transmission among humans, particularly the ability of the direct progenitor of SARS-CoV-2 to bind ACE2, likely occurred even before its introduction to the intermediate host reservoir (Fig. 8) (*19*). Analogously, 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 generalist nature (*44*–*49*).

Viruses that can infect humans and naturally circulate in animals, such as rabies, Ebola, Nipah, Lassa, and highly pathogenic avian influenza viruses, are rare, but of high consequence (*50*–*53*). However, once a virus is capable of human infection and transmission, the remaining barriers to spillover are pervasiveness of the virus among animal hosts and extended host-human contact (*54*). Thereafter, a single zoonotic transmission event portends additional jumps, because all previous conditions leading up to a successful zoonotic jump have necessarily been met. For example, the reverse-zoonosis of SARS-CoV-2 from humans to minks on Dutch fur farms was followed by repeated reintroduction of SARS-CoV-2 from minks to humans (*46*, *47*). Indeed, we can see multiple cross-species transmissions from minks to humans from a shallow genetic reservoir on multiple farms in the Netherlands (*37*, *46*). Further, multiple introductions of SARS-CoV-2 from a small population of imported Syrian hamsters to humans were recently reported in Hong Kong (*49*). Notably, the genomic diversity of the viral reservoir in these animals is shallow, likely matching the genomic diversity within the putative animal reservoir at the Huanan market (Fig. 8) (*46*, *49*).

Our findings show that it is highly unlikely that SARS-CoV-2 circulated in humans earlier than November 2019 or spread internationally before December 2019. With highly limited genetic diversity and likely fewer than 20 infections by the tMRCA of the first introduction, there could not have been substantial cryptic spread before the first detection of COVID-19 in Wuhan in mid-December 2019.

By the time SARS-CoV-2 was subsequently identified as the etiological agent of COVID-19 in late-December (*7*), the virus had clearly spread sufficiently such that containment was challenging.

People living in the vicinity of horseshoe bat species (*Rhinolophus spp.*) are more likely to show serological evidence of infection by SARS-related CoVs (*33*, *55*, *56*). This observation underscores the continued threat of emergence of novel viruses from animal reservoirs, with wildlife markets posing a particularly high risk (*57–60*). Although zoonoses leading to sustained epidemics will inevitably be discovered, those with little to no onward transmission are likely to occur without leaving any epidemiological trace (*61*), particularly with viruses associated with asymptomatic infections or mild disease, like SARS-CoV-2. Our results highlight the imperative of establishing intensive virus surveillance architectures. Such architectures will involve early detection of unexplained disease in humans, but must be complemented by a focus on wild, farmed, and traded animals where the risk of transmission to humans is greatest. The ability to more rapidly identify spillovers and novel pathogens will improve our capacity to contain pathogens with pandemic potential.

Note: A recent preprint by Gao *et al*. presented evidence of a single lineage A environmental sample in the Huanan market on 1 January 2020 (*62*), consistent with a separate and subsequent origin of lineage A at the Huanan market.

Acknowledgements

We thank 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. A complete list acknowledging the authors who submitted the data found in GISAID and analyzed in this study can be found in Data S1. The GISAID genomes are also available using the GISAID EPI_SET ID EPI_SET_20220305ud. We are greatly appreciative towards 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 8 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). 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.

Authors' 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 their institution unrelated to this research. MAS receives contracts and grants from the US Food & Drug Administration, the US Department of Veterans Affairs and Janssen Research & Development unrelated to this research. RFG. is co-founder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. MW, ECH, AR, MAS, and KGA have received consulting fees on SARS-CoV-2 and the COVID-19 pandemic.

Data and materials availability:

Supplementary Materials:

Materials and Methods

Supplementary Text

Tables S1-S9

Figs S1-S25

References (1-95)

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).

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.* . **133**, 1015–1024 (2020).

3. 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).

4. 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).

5. 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).

6. S. Lytras, J. Hughes, D. Martin, P. Swanepoel, A. de Klerk, R. Lourens, S. L. K. Pond, W. Xia, X. Jiang, D. L. Robertson, Exploring the natural origins of SARS-CoV-2 in the light of recombination. *Genome Biol. Evol.* (2022), doi:10.1093/gbe/evac018.

7. M. Worobey, Dissecting the early COVID-19 cases in Wuhan. *Science*. **374**, 1202–1204 (2021).

8. 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).

9. 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).

10. Clock and TMRCA based on 27 genomes. *Virological* (2020), (available at https://virological.org/t/clock-and-tmrca-based-on-27-genomes/347/6).

11. 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).

12. 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).

13. W. H. O. Headquarters, WHO-convened global study of origins of SARS-CoV-2: China Part (2021), (available at https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-

cov-2-china-part).

14. 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), (available at 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).

15. Issues with SARS-CoV-2 sequencing data. *Virological* (2020), (available at https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473).

16. 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).

17. 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).

18. 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).

19. S. Temmam, K. Vongphayloth, E. B. Salazar, 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* (2022), doi:10.1038/s41586-022-04532-4.

20. 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).

21. 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).

22. 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** (2020), doi:10.1128/mSphere.00408-20.

23. 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).

24. 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).

25. 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).

26. 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).

27. 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).

28. 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, Seattle Flu Study Investigators, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome, Cryptic transmission of SARS-CoV-2 in Washington state. *Science*. **370**, 571–575 (2020).

29. 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).

30. N. Moshiri, CoaTran: Coalescent tree simulation along a transmission network. *bioRxiv* (2020), p. 2020.11.10.377499.

31. Further musings on the tMRCA. *Virological* (2020), (available at https://virological.org/t/further-musings-on-the-tmrca/340).

32. J. Giesecke, Primary and index cases. *Lancet*. **384**, 2024 (2014).

33. 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).

34. 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).

35. J. Ma, First Chinese coronavirus cases may have been infected in October 2019, says new research. *South China Morning Post* (2021), (available at https://www.scmp.com/news/china/science/article/3126499/first-chinese-covid-19-cases-may-have-been-infected-october-2019).

36. 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).

37. 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), doi:10.5281/zenodo.6299116.

38. 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).

39. 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).

40. G. Dudas, L. M. Carvalho, A. Rambaut, T. Bedford, MERS-CoV spillover at the camel-human interface. *Elife*. **7** (2018), doi:10.7554/eLife.31257.

41. 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).

42. 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).

43. 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).

44. 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. 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* (2021), doi:10.1038/s41586-021-04353-x.

45. 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** (2021), doi:10.1073/pnas.2114828118.

46. L. Lu, R. S. Sikkema, F. C. Velkers, D. F. Nieuwenhuijse, E. A. J. Fischer, P. A. Meijer, N. Bouwmeester-Vincken, A. Rietveld, M. C. A. Wegdam-Blans, P. Tolsma, M. Koppelman, L. A. M. Smit, R. W. Hakze-van der Honing, W. H. M. van der Poel, A. N. van der Spek, M. A. H. Spierenburg, R. J. Molenaar, J. de 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).

47. 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).

48. 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** (2022), doi:10.1073/pnas.2121644119.

49. 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 (Variant Delta) from Pet Hamsters to Humans and Onward Human Propagation of the Adapted Strain: A Case Study (2022), (available at https://papers.ssrn.com/abstract=4017393).

50. Z. L. Grange, T. Goldstein, C. K. Johnson, S. Anthony, K. Gilardi, P. Daszak, K. J. Olival, T. O'Rourke, S. Murray, S. H. Olson, E. Togami, G. Vidal, Expert Panel, PREDICT Consortium, J. A. K. Mazet, University of Edinburgh Epigroup members those who wish to remain anonymous, Ranking the risk of animal-to-human spillover for newly discovered viruses. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021), doi:10.1073/pnas.2002324118.

51. M. E. Khalifa, L. Unterholzner, M. Munir, Structural and Evolutionary Insights Into the Binding of Host Receptors by the Rabies Virus Glycoprotein. *Front. Cell. Infect. Microbiol.* **11**, 736114 (2021).

52. A. Islam, J. Ferdous, S. Islam, M. A. Sayeed, S. Dutta Choudhury, O. Saha, M. M. Hassan, T. Shirin, Evolutionary Dynamics and Epidemiology of Endemic and Emerging Coronaviruses in Humans, Domestic Animals, and Wildlife. *Viruses*. **13** (2021), doi:10.3390/v13101908.

53. E. R. Rush, E. Dale, A. A. Aguirre, Illegal Wildlife Trade and Emerging Infectious Diseases: Pervasive Impacts to Species, Ecosystems and Human Health. *Animals (Basel)*. **11** (2021), doi:10.3390/ani11061821.

54. R. K. Plowright, C. R. Parrish, H. McCallum, P. J. Hudson, A. I. Ko, A. L. Graham, J. O. Lloyd-Smith, Pathways to zoonotic spillover. *Nat. Rev. Microbiol.* **15**, 502–510 (2017).

55. N. Wang, S.-Y. Li, X.-L. Yang, H.-M. Huang, Y.-J. Zhang, H. Guo, C.-M. Luo, M. Miller, G. Zhu, A. A. Chmura, E. Hagan, J.-H. Zhou, Y.-Z. Zhang, L.-F. Wang, P. Daszak, Z.-L. Shi, Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol. Sin.* **33**, 104–107 (2018).

56. H. Li, E. Mendelsohn, C. Zong, W. Zhang, E. Hagan, N. Wang, S. Li, H. Yan, H. Huang, G. Zhu, N. Ross, A. Chmura, P. Terry, M. Fielder, M. Miller, Z. Shi, P. Daszak, Human-animal interactions and bat coronavirus spillover potential among rural residents in Southern China. *Biosaf Health*. **1**, 84–90 (2019).

57. M. Ruiz-Aravena, C. McKee, A. Gamble, T. Lunn, A. Morris, C. E. Snedden, C. K. Yinda, J. R. Port, D. W. Buchholz, Y. Y. Yeo, C. Faust, E. Jax, L. Dee, D. N. Jones, M. K. Kessler, C. Falvo, D. Crowley, N. Bharti, C. E. Brook, H. C. Aguilar, A. J. Peel, O. Restif, T. Schountz, C. R. Parrish, E. S. Gurley, J. O. Lloyd-Smith, P. J. Hudson, V. J. Munster, R. K. Plowright, Ecology, evolution and spillover of coronaviruses from bats. *Nat. Rev. Microbiol.* (2021), doi:10.1038/s41579-021-00652-2.

58. W.-T. He, X. Hou, J. Zhao, J. Sun, H. He, W. Si, J. Wang, Z. Jiang, Z. Yan, G. Xing, M. Lu, M. A. Suchard, X. Ji, W. Gong, B. He, J. Li, P. Lemey, D. Guo, C. Tu, E. C. Holmes, M. Shi, S. Su, Total virome characterizations of game animals in China reveals a spectrum of emerging viral pathogens. *bioRxiv* (2021), p. 2021.11.10.467646.

59. L.-J. Chen, X.-D. Lin, W.-P. Guo, J.-H. Tian, W. Wang, X.-H. Ying, M.-R. Wang, B. Yu, Z.-Q. Yang, M. Shi, E. C. Holmes, Y.-Z. Zhang, Diversity and evolution of avian influenza viruses in live poultry markets, free-range poultry and wild wetland birds in China. *J. Gen. Virol.* **97**, 844–854 (2016).

60. Y. Guan, B. J. Zheng, Y. Q. He, X. L. Liu, Z. X. Zhuang, C. L. Cheung, S. W. Luo, P. H. Li, L. J. Zhang, Y. J. Guan, K. M. Butt, K. L. Wong, K. W. Chan, W. Lim, K. F. Shortridge, K. Y. Yuen, J. S. M. Peiris, L. L. M. Poon, Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science*. **302**, 276–278 (2003).

61. E. E. Glennon, F. L. Jephcott, O. Restif, J. L. N. Wood, Estimating undetected Ebola spillovers. *PLoS Negl. Trop. Dis.* **13**, e0007428 (2019).

62. 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 (2022), , doi:10.21203/rs.3.rs-1370392/v1.

63. S. Khare, C. Gurry, L. Freitas, M. B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R. T. Lee, W. Yeo, G. C. Curation Team, S. Maurer-Stroh, GISAID’s Role in Pandemic Response. *China CDC Wkly*. **3**, 1049–1051 (2021).

64. 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).

65. Masking strategies for SARS-CoV-2 alignments. *Virological* (2020), (available at https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480).

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

67. 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).

68. 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).

69. *gofasta* (Github; https://github.com/virus-evolution/gofasta).

70. G. Dudas, *baltic: baltic - backronymed adaptable lightweight tree import code for molecular phylogeny manipulation, analysis and visualisation. Development is back on the evogytis/baltic branch (i.e. here)* (Github; https://github.com/evogytis/baltic).

71. 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).

72. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*. **4**, vex042 (2018).

73. 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).

74. A. Rambaut, *figtree* (Github).

75. 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).

76. 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).

77. 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).

78. N. Moshiri, *FAVITES-COVID-Lite: A simplified (and much faster) simulation pipeline specifically for COVID-19 contact + transmission + phylogeny + sequence simulation* (Github; https://github.com/niemasd/FAVITES-COVID-Lite).

79. N. Moshiri, NiemaGraphGen: A memory-efficient global-scale contact network simulation toolkit. *arXiv [physics.soc-ph]* (2022), (available at http://arxiv.org/abs/2201.04625).

80. A. L. Barabasi, R. Albert, Emergence of scaling in random networks. *Science*. **286**, 509–512 (1999).

81. S. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Modelling disease outbreaks in realistic urban social networks. *Nature*. **429**, 180–184 (2004).

82. 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).

83. 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).

84. X. Yang, Y. Yu, J. Xu, H. Shu, J. ’an 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).

85. 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).

86. 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).

87. N. Moshiri, TreeSwift: A massively scalable Python tree package. *SoftwareX*. **11**, 100436 (2020).

88. 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** (2021), doi:10.1128/mSphere.00019-21.

89. 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).

90. Issues with SARS-CoV-2 sequencing data. *Virological* (2020), (available at https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/6).

91. 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).

92. 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. *bioRxiv* (2021), , doi:10.1101/2021.07.27.21261148.

93. 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** (2014), doi:10.1098/rspb.2014.0732.

94. J. Ma, Coronavirus: China's first confirmed Covid-19 case traced back to November 17. *South China Morning Post* (2020), (available at https://www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back).

95. 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).

Materials and Methods

Sequence data. We queried the GISAID database SARS-CoV-2 viral genome alignment for sequences collected by 14 February 2020 (*63*). We selected this date to have a reasonable data set size that can be used for Bayesian phylogenetic analyses (*i.e.*, under 1000 genomes). We restricted our data set to sequences that (i) were ≥29,000 nucleotides, (ii) had high coverage with ≤0.5% unique amino acid mutations, (iii) had fewer than 1% 'N's, (iv) were not identified as potentially problematic via NextStrain (*64*), and (v) had a year-month-day sampling date reported. Genomes described in the WHO report (*13*) to have erroneous mutations were updated, and if the virus was sequenced multiple times, the corrections belonging to the genome with the highest coverage were used. The first 88 and last 195 nucleotides of each genome were masked due to poor evidence of homology, and an additional 105 sites were masked based on the work by De Maio *et al.* (*65*), leading to a total of 388 masked sites. Genomes with an ambiguous nucleotide (*e.g.*, Y or N) at site 8782 or 28144 were excluded. We excluded an additional 20 genomes from subsequent analyses, because the sequences contained either C8782T or T28144C (compared to the reference genome Wuhan Hu-1) for reasons described in the main text and De Maio et al. (2020) (*15*). The final dataset comprised 787 taxa. A list of GISAID and GenBank accessions are available in Data S1 and S2, respectively.

Examining shared mutations between early C/C and T/T genomes and lineages A and B. The 787 SARS-CoV-2 genome dataset and the 20 genomes with either C8782T or T28144C were aligned with MAFFT v7.453 (*2*) (options --auto --keeplength --addfragments) to reference genome Hu-1 (GenBank accession MN908947.3; GISAID accession EPI_ISL_402125). We then looked for pairs of genomes comprising an intermediate genome (C/C or T/T) and a major lineage (lineage A or B) that shared derived mutations.

We repeated this analysis for all complete, high-coverage SARS-CoV-2 consensus genomes collected by 28 February 2020 and submitted by 31 December 2020 to GISAID. We excluded any genomes that had an incomplete collection date, leaving 1716 genomes (including the initial 807). The genomes used are listed in Data S1.

Sequencing quality of specific early C/C and T/T genomes. We aligned reads FASTQ files belonging to EPI_ISL_413017, a C/C genome from South Korean, and EPI_ISL_462306, a T/T genome from Singapore, using Minimap2 (*66*), sorted the subsequent SAM files using samtools (*67*), and called variants using iVar (*68*). The variant calls were then manually inspected for depth and indeterminacy at specific sites, including 8782 and 28144. The raw data for EPI_ISL_413017 and EPI_ISL_462306 are available at https://www.ncbi.nlm.nih.gov/sra, with project IDs PRJNA806767 and PRJNA802993, respectively.

Finding C/C and T/T genomes in San Diego. The SEARCH consortium has sequenced over 35,000 genomes from San Diego during the course of the pandemic, by 3 February 2022. We generated a multiple sequence alignment of these genomes using Minimap2 and gofasta (*69*). We queried this alignment for all consensus genomes with C8782T and T28144, and validated these mutations by checking the read depth and allele frequency in the original alignment files.

The consensus genomes and associated BAM files are publicly available at https://github.com/andersen-lab/HCoV-19-Genomics, and the genome accessions are listed in Data S2. The tree figures were rendered using baltic (*70*).

Finding T/T genomes in NYC. Molecular surveillance conducted by the New York City Public Health Laboratory, part of the Department of Health and Mental Hygiene, has sequenced >5000 SARS-CoV-2 genomes through the end of 2021. We queried these data for genomes with C8782T and T28144 in the consensus sequence and validated the consensus sequence by checking the read depth and allele frequency from the primer removed BAM files. The genome accessions are listed in Data S1. The tree figure was rendered using baltic.

Constructing the recombinant common ancestor. We used the aligned sarbecovirus genomes and breakpoints provided by Sarah Temmam and Marc Eloit (*19*) to infer the phylogenetic history of each non-recombinant region. We inferred a maximum likelihood tree of the animal viruses in the alignment and Hu-1 for each non-recombinant region using IQ-TREE 2 (*71*) under a GTR+F+G+I substitution model. We midpoint-rooted each maximum likelihood tree and used TreeTime (*72*) to perform ancestral sequence reconstruction for each fragment. The genome belonging to the parent node of Hu-1 for each non-recombinant region was extracted, and then they were all strung together to construct the recombinant common ancestor (recCA) of SARS-CoV-2 in the sarbecovirus clade. The rooting of each tree occurs sufficiently upstream of the parent node of SARS-CoV-2, indicating that this recCA would not be sensitive to rooting.

There are 382 substitutions between Hu-1 and the recCA, one of which is masked. Since 387 of the remaining 29,521 sites were masked, there are 29,134 sites identical between Hu-1 and the recCA. When we used a different SARS-CoV-2 genome (*e.g.*, WH04, WA1) to construct the recCA, the recCA sequence was consistent, indicating the recCA can reliably be used as an ancestor of SARS-CoV-2.

We created a simplot of the sarbecovirus genomes and the recCA against Hu-1 using RDP4 (*73*), and our phylogenies were visualized using FigTree (*74*). The genome accessions are listed in Data S1 for genomes from GISAD and Data S2 for genomes from GenBank.

Early pandemic reversion analysis. Here, we define mutations away from the Hu-1 reference genome toward the recCA, such as C8782T and T28144C, as reversions. The 787 genomes were aligned with MAFFT to Hu-1 (GISAID accession EPI_ISL_402125). The phylogenetic history of SARS-COV-2 in China was first inferred in a maximum likelihood framework in IQ-TREE 2 using a GTR+F+I model. We used TreeTime to perform ancestral state reconstruction on the maximum likelihood tree of 787 genomes, rooted on Hu-1. We determined which branches had reversions. Each unique reversion and non-reversion substitution was only counted once to account for phylogenetic uncertainty. The tree figure was rendered using baltic.

Reversion analysis through the pandemic. We extended the reversion analysis from above to the following variants: Alpha (PANGO lineage B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Epsilon 1 (B.1.427), Epsilon 2 (B.1.429), Zeta (P.2), Iota (B.1.526), and Kappa (B.1.617.1). To match the 52-day window for sequence collection from the early pandemic, we found the earliest 52-day window in GISAID for each variant that contained at least 750 sequences. The genomes for each lineage were then aligned using MAFFT to reference genome Hu-1. We next used IQ-TREE 2 and TreeTime as above to generate a phylogeny and perform ancestral state reconstruction. We followed the same protocol as above with the recCA to determine which branches had reversions. The tree figures were rendered using baltic.

To examine reversions across a subsample of the entire pandemic, we extracted the global tree and associated public genomes from Nextstrain (*64*) on 14 January 2022 and constrained it to sequences before November 2021. We aligned the genomes using MAFFT as above, used TreeTime with the NextStrain tree and alignment to perform ancestral state reconstruction, and determined reversions with the recCA. The genomes used for the clade and global pandemic reversion analysis can be found in Data S1 and S2, respectively.

Phylogenetic inference. Molecular clock analysis was conducted using a Bayesian approach in BEAST v1.10.5 (*75*). For the primary analysis, we developed and employed a non-reversible, random-effects substitution model (described below), a strict molecular clock, a non-parametric skygrid prior with 20 grid points and a cut off of 0.37, which translates to 05 October 2019. To facilitate Markov chain Monte Carlo (MCMC) chain convergence, (i) we used our previous results of $7.9\text{x}10^{-4}$ and $6.8\text{x}10^{-5}$ substitutions/site/year as the mean and standard deviation, respectively, of a normal prior for the clock rate, and (ii) we initiated the MCMC sampling using the maximum likelihood phylogeny that had been transformed into a chronogram via TempEst v1.5.3 (*76*). We ran four independent chains of 400 million generations, sub-sampling every 25 thousand iterations to continuous parameter log files, 100 thousand iterations for the tree file, and 100 thousand iterations for the ancestral state reconstruction of the most recent common ancestor (MRCA); the first 15% of the chains were discarded as burnin. Convergence and mixing was assessed in Tracer v1.7.1 (*77*) and all 4 chains were combined in LogCombiner, such that all relevant effective sample size (ESS) values were >200. The accession IDs can be found in Data S1 and S2, and the XML files can be found at https://github.com/sars-cov-2-origins/multi-introduction. The outputs can be found at doi:10.6076/D1Q59K.

Random-effects substitution model. To accommodate the mutational bias in SARS-CoV-2 for C-to-T transitions (*21–23*), we developed a random-effects substitution model. This model employs a standard phylogenetic substitution model as a base model, incorporated as fixed effects, while the random effects allow each individual mutation rate (*e.g.*, C-to-T, separate from T-to-C) to be elevated (or decreased) relative to that model. Note that this makes the model non-reversible. We use HKY as the base model, which is defined by the parameters κ (governing the relative rate of transitions versus transversions) and $\boldsymbol{\pi}$ (governing the root and stationary frequencies). Working on the log-scale and denoting the random effect ϵ_{ij}, our HKY+RE model gives the log of the substitution rate matrix entries as,

$$log(q_{ij}) \ = log(\pi_j) \ + \ log(\kappa) \ + \epsilon_{ij} \ + log(\xi) \text{ if i>j is a transition}$$

or,

$$log(q_{ij}) \ = log(\pi_j) \ + \epsilon_{ij} + log(\xi) \text{ if i>j is a transversion,}$$

where ξ is a normalizing constant.

To accommodate among-site rate variation, we employ a proportion of invariable sites in the model, with a Uniform(0,1) prior. We place independent and identical Normal(mean=0,SD=σ_ϵ) priors on ϵ_{ij}, an improper infinite uniform prior on σ_ϵ, and a Normal(mean=0,SD=1.25) prior on log(κ). We fix $\boldsymbol{\pi}$ to the frequencies observed in the alignment.

Ancestral state reconstruction Bayes factors. We assume equally likely prior probabilities for each sequence in the ancestral state reconstruction posterior. Therefore, the Bayes factor (BF) in favor of sequence S_1 against another sequence S_0, given the data D, can be expressed as follows:

$$B_{10} = \frac{P(S_1|D)}{P(S_0|D)}$$

Note that all BFs were calculated with the sequence comprising the highest posterior probability as S_1, and BFs for each phylodynamic model were calculated separately.

BEAST sensitivity analyses. We performed four sensitivity analyses for the primary BEAST analysis: (i) using a GTR+F+I model; (ii) unmasked sequences; (iii) excluding 15 market-associated genomes; and (iv) excluding all genomes sampled from Wuhan.

Phylogenetic inference with constrained roots and recCA. In a standard phylogenetic model, the sequences at all internal nodes (and the root) are integrated out. A prior distribution is required for the root sequence in order to compute the likelihood or sample ancestral states at nodes. Typically, the prior distribution for the root sequence assumes every site is drawn identically and independently from some multinomial distribution defining the prior probability of an A, C, G, or T nucleotide at any (and every) site. Here, we consider two alternatives. The first alternative constrains the character state of the MRCA of all human SARS-CoV-2 sequences to be identical to a specific hypothesized ancestral haplotype (Fig 4B). We consider 6 ancestral haplotypes that match the sequences belonging to Wuhan Hu-1, the C/C haplotype (Hu-1 with T28144C), the T/T haplotype (Hu-1 with C8782T), WH04, 20SF2012, and WA1. Note that, in this model, the root of the tree is the MRCA of all 787 human SARS-CoV-2 sequences, and the aforementioned genomes remain taxa–the tree is not rooted on them. The second alternative places a per-site prior distribution on the ancestral haplotype. We do this by adding a branch ancestral to the MRCA and fixing the sequence at the root to be the sequence of the recCA (Fig 4D). This approach places a prior distribution on the ancestral haplotype, as the root and MRCA are distinct in this model. At each site, this prior is determined by the nucleotide present in the recCA, the length of the branch leading to the MRCA, and the substitution model parameters.

Epidemic simulation. To explore the evolutionary dynamics during the beginning of the COVID-19 pandemic, we developed FAVITES-COVID-Lite (*78*), a simplified simulation pipeline based on FAVITES (*24*), and performed a series of epidemic simulations. First, we generated static contact networks comprising 5 million individuals (nodes) using NiemaGraphGen (*79*) under a preferential-attachment model using the Barabási–Albert algorithm (*80*). We used this network algorithm because its scale-free properties recapitulate infectious disease spread (*81*). We chose to simulate a static contact network because our focus is on the number of people infected at the beginning of the epidemic, and we used an intermediate value of 16 contacts per day (mean degree), based on Mossong et al. (*82*). This approach captures highly connected individuals responsible for superspreading (*9*, *16*).

We extended the SAPHIRE [Susceptible (S)-Ascertained (I)-Presymptomatic (P)-Hospitalized (H)-Not Ascertained (A)-Removed (R)-Exposed (E)] model developed by Hao et al. (*27*), and implemented in our previous study on the timing of the primary case (*9*), to have two ascertained compartments. Individuals from the first ascertained compartment can either enter the recovered compartment or an "ascertained-pre-hospitalization" (I_H) compartment, where they eventually transition to the hospitalized compartment. We extended the model with the I_H compartment to decouple the proportion of people hospitalized with the amount of time until hospitalization. We did not include the travel component of the original SAPHIRE model (*i.e.*, individuals flying into and out of Wuhan), because our focus was on the early dynamics of the pandemic before its spread outside of Wuhan. The dynamics of these compartments across time (*t*) are described by the following set of ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{bS(\alpha P + \alpha A + I + I_h)}{N} \\
\frac{dE}{dt} &= \frac{bS(\alpha P + \alpha A + I + I_h)}{N} - \frac{E}{D_e} \\
\frac{dP}{dt} &= \frac{E}{D_e} - \frac{P}{D_p} \\
\frac{dA}{dt} &= \frac{(1-r)P}{D_p} - \frac{A}{D_i} \\
\frac{dI}{dt} &= \frac{rP}{D_p} - \frac{I}{D_i} \\
\frac{dI_h}{dt} &= \frac{hI}{D_i} - \frac{I_h}{D_q - D_i} \\
\frac{dH}{dt} &= \frac{I_h}{D_q - D_i} - \frac{H}{D_h}
\end{aligned}$$

$$\frac{dR}{dt} = \frac{A + (1-h)I}{D_i} + \frac{H}{D_h}$$

We performed forward simulations using this extended model to generate a viral transmission network using GEMF (*83*). Simulated epidemics started with a single seed infection among our 5 million susceptible individuals. The parameters were primarily determined by Hao et al. (*27*) (Table S8), except we required half the ascertained population to become hospitalized with an average of 11 days between symptom onset and hospitalization, matching reports of hospitalization of the early pandemic in Wuhan (*84–86*). Each simulation was run for 70 days and produced an output documenting when individuals transitioned from one compartment to another throughout the entire simulation. We used these to determine the number of individuals in a given compartment (*e.g.*, total infections, ascertained infections, unascertained infections, and hospitalized individuals) across each day in the simulation.

Once the forward simulations were complete, we subsampled the first 50,000 infected individuals. Then, we sample a sequencing time for each ascertained individual from a uniform distribution that starts when they enter the first ascertained compartment and ends when they're recovered. To match real-world data from December 2019, we only include individuals with sequencing times that occur after the first hospitalization. Additionally, the primary case (*i.e.*, first infected individual in the simulation) is sampled regardless of ascertainment status. We sample the primary case regardless of ascertainment or hospitalization status to properly determine stable coalescence (described below). All unascertained individuals are not sampled.

Lastly, we provide the genome sampling times and transmission network to CoaTran (*79*), which uses a coalescent process to generate time trees. We then use a constant substitution rate of 9.2x10^{-4} substitutions/site/year (inferred from our primary BEAST results) to convert the branch lengths from years to substitutions per site to create mutation trees.

For the primary analysis, we ran epidemic simulations until there were 1100 successful simulations, defined as those simulations in which ≥400 people had become infected and ≥1 person was still infectious at the end of the simulation. Failed epidemics were simulations that did not become established (*i.e.*, 0 infectious people at the end of the simulation) or had fewer than 400 people infected over the entire simulation; 2204 (66.7%) simulations failed to reach this epidemic threshold after 70 days. Our epidemics had a median doubling time of 2.65 days (95% range: 1.51-4.14), lower than the doubling times from Hao et al. (*27*) to match estimates of the growth rate from before 01 January 2020 in Wuhan (*25*, *26*). Twenty of the 3304 simulations (0.6%) did not reach 400 infections but were still persisting at the end of the simulation. None of these simulated epidemics resulted in more than 40 infections. GIven their rarity, it is highly improbable that SARS-CoV-2 was characterized by these growth dynamics.

All inputs for the primary analysis can be found in https://github.com/pekarj/SC2_emergence, the outputs and post-processed results can find at doi:10.6076/D1Q59K, and the software used is available at https://github.com/niemasd/FAVITES-COVID-Lite/tree/no_seqgen.

Determining stable coalescence. As in our previous study (*9*), we define the stable coalescence as the tMRCA that does not shift forward in time by more than one day, even as new individuals become infected and previously infected individuals recover (*i.e.*, the time to the most recent common ancestor [tMRCA]). Therefore, the stable coalescence is reached the first day that the coalescence for the currently infected individuals is within one day of the time of MRCA after the 70 day simulation or once 50,000 total individuals have been infected.

We extracted the tMRCA of infected and sampled individuals every day across each simulation using TreeSwift 1.1.14 (*87*). This tMRCA was calculated for each day of the 70 days or until 50,000 individuals had been infected, whichever came first. We chose not to explore dynamics after 50,000 infections due to a slowing in exponential growth arising from the saturation of the contact network.

Sensitivity analysis–slower rate of infection. The same approach for epidemic simulation was undertaken to evaluate the dynamics of a more slowly spreading virus. We used the aforementioned parameters, except the infectiousness coefficient was decreased from 0.38 to 0.28 per day (0.74x) and the simulation time was increased from 70 to 100 days (1.43x). We produced 1100 successful simulations with at least 400 infected individuals and a median epidemic doubling time of 3.47 days (95% range: 1.53-5.78), and 3857 simulations failed to reach the epidemic threshold.

Combining epidemic simulations and BEAST via rejection sampling. We previously described rejection sampling using the timing of the earliest case. Here we extend our previous method to include conditioning on the timing of the earliest hospitalization. Our aim is to obtain a posterior distribution for the date X of the primary case (the first case resulting from a SARS-CoV-2 cross-species transmission) in Wuhan, conditioned on the available sequencing data D_S, the date of the first reported COVID-19 case D_C, and the date of the first hospitalization D_H due to COVID-19. We do this in a Bayesian framework by marginalizing over the date T of the tMRCA as follows:

$$P(X|D_S, D_C, D_H) \ = \int_T P(X|T, D_S, D_C, D_H) P(T|D_S, D_C, D_H) \, dT \qquad \text{Equation (1)}$$

We assume that the sequencing data are informative only for the tMRCA; *i.e.*, given T, X does not depend on D_S: $P(X|T, D_S, D_C, D_H) \ = \ P(X|T, D_C, D_H)$. We also assume that the first reported COVID-19 case and hospitalization data are not informative for the tMRCA: $P(T|D_s, D_c, D_H) \ = \ P(T|D_s)$. This gives:

$$P(X|D_S, D_C, D_H) \ = \int_T P(X|T, D_C, D_H) P(T|D_S) \, dT \qquad \text{Equation (2)}$$

We further note that $P(X|T, D_c, D_H) \ = \int_{Y_H} \int_{Y_C} P(X, Y_C, Y_H|T, D_C, D_H) \, dY_C dY_H$, where Y_C and Y_H are the first simulated COVID-19 ascertained case and hospitalization, respectively. We model $P(X, Y_C, Y_H|T, D_C, D_H)$ as proportional to $I(Y_C \leq D_C \, , Y_H \leq D_H) P(X, Y_C, Y_H|T)$, where $I(Y_C \leq D_C \, , Y_H \leq D_H)$ is an indicator function with a value of 1 when Y_C and Y_H are consistent with D_C and D_H, respectively, and 0 otherwise. This approach allows us to sample from the posterior distribution of Equation 2. The BEAST analysis provides values of T sampled from the distribution $P(T|D_S)$, which we can use in conjunction with FAVITES to sample corresponding values of X, Y_C, and Y_H from the distribution $P(X, Y_C, Y_H|T)$. We use a simple rejection sampling approach to continue sampling from $P(X, Y_C, Y_H|T)$ until a sample is obtained for which $I(Y_C \leq D_C\ , Y_H \leq D_H) = 1$. The resulting set of sample values for X then follow the posterior distribution $P(X|D_S, D_C, D_H)$.

We require the first simulated case to be ascertained (SAPHIRE stage: I) and assign D_C as 10 December 2019. However, we note that this first ascertained case can be the primary case themselves, unless a secondary or tertiary case progresses faster through the course of infection. We assign D_H as 16 December 2019. Importantly, the rate at which cases were ascertained in the SAPHIRE model is based on real-time patterns in COVID-19 diagnosis from 01 through 22 January 2020 and may not reflect the actions that led to the retrospective diagnosis of earliest cases of COVID-19. Further, stable coalescence (*i.e.*, the MRCA) can happen any time after the primary case is infected, and there is no requirement for stable coalescence to occur after the first ascertained and unascertained individuals. Justifications for dates used here and in the sensitivity analyses is discussed in the supplementary text.

Sensitivity analysis–earliest case date of 8 December. We also condition single-introduction analyses on an 8 December case date, which was previously discounted by the WHO and 16 December hospitalization date (see supplementary text for full discussion).

Sensitivity analysis–rejection sampling with hospitalization only. We remove the requirement for the first simulated case to be ascertained by a given date, and then condition analyses only on the tMRCA and date of the first hospitalization.

Sensitivity analysis–recCA and constrained roots. We explored the sensitivity of the timing of the primary case to the phylodynamic model choice. We applied rejection sampling to the inferred phylogenies constrained by the recCA and SARS-CoV-2 roots and the primary forward epidemic simulations, conditioning on the same dates as above.

Forward simulation clade analysis. We examined the clade structure of the posterior phylogenetic trees resulting from the epidemic simulations. For each tree, we simulated individual mutations down the branches of the subtree whose root is the stable coalescence and using a substitution rate of 9x2x10^{-4} substitutions/site/year. We then counted the clades that were one mutation from the root (*i.e.*, stable coalescence), clades that were two mutations from the root, the size of these clades, any basal taxa, and any taxa with a unique mutation(s) that were not part of any clades. Note that, for clarity, we display the subtree with stable coalescence as a root in Figure 5.

Rejection sampling for lineages A and B. We apply the above method to the tMRCA, first ascertained case date, and first hospitalized case date for each lineage to infer the timing of the primary case for each lineage. For lineage B, the earliest case and hospitalization dates are 13 December and 16 December, respectively. For lineage A, the earliest case date is 15 December and the earliest hospitalization date is 25 December (see supplementary text for full discussion). After performing rejection sampling for both lineages, we combine the number of individuals in each compartment for each day in the dated simulations.

Sensitivity analysis–earlier lineage B COVID-19 dates. We performed rejection sampling for lineage B using a case date and hospitalization date of 10 December and 16 December, matching the SARS-CoV-2 index case, which does not have an associated genome. We also performed rejection sampling for lineage B using a case date and hospitalization date of 8 December and 18 December, respectively, in case the earliest lineage B case occurred on 8 December (see supplementary text for full discussion).

Simulating additional hospitalizations from failed introductions. Using our single introduction rejection sampling results, we combined simulations that failed to produce epidemics with a single successful introduction. For each combined epidemic simulation and BEAST tMRCA, we sampled up to 100 additional introduction times from a normal distribution centered at the time of the primary case with a standard deviation of 7 days. We then attached randomly sampled failed epidemics to the given introduction times and quantified the number of additional hospitalizations caused by these introductions.

Simulating cross-species transmissions to achieve two successful introductions. Our epidemic simulations had a success rate of approximately 33% (1100 successful introductions; 2204 failed introductions). To simulate the number of cross-species transmissions needed to achieve two successful introductions, we treated successful introductions as Bernoulli trials, with a success rate of 33% and simulated trials until there were two successful trials. We validated our result by determining the number of trials needed to achieve at least two successful trials using a negative binomial distribution (supplementary text).

Supplementary Text

C/C and T/T genomes through 28 February 2020. We extended our search for intermediates through 28 February 2020 after we determined the C/C and T/T genomes collected by 14 February 2020 are likely erroneous (Main Text). There were 28 C/C genomes collected by 28 February 2020, including the 16 genomes collected by 14 February 2020 (described in the Main Text). Of the 28 C/C genomes, 6 have no additional mutations from the Hu-1 reference genome other than T28144C. We identified 16 C/C genomes that share nucleotide substitutions also found in lineage A (Fig. S1), and 11 C/C genomes that share substitutions found within lineage B (Fig. S2). There are occasionally sets of taxa that share mutations with another set of taxa, except with a different pattern. For example, there are lineage B taxa with either C22444T or C26088T, but one C/C genome (EPI_ISL_539558) has both of these mutations (Fig. S2). These incongruencies are indicated by the brackets on the line connecting taxa in Figures S1 and S2. Notably, 9 of the C/C genomes share substitutions with both A and B lineages, whereas 4 contain substitutions not seen in other lineages.

There were 10 T/T genomes collected by 28 February 2020, including the 4 genomes collected by 14 February 2020. Two high coverage T/T genomes sampled after 14 February 2020 were from the Diamond Princess cruise ship outbreak, but these genomes were descendants of the lineage B virus that initiated this outbreak (*88*, *89*) (Fig. S3). Of the remaining 4 T/T genomes, EPI_ISL_418251 and EPI_ISL_418247 both have additional mutations C3037T and A23403G, which are common among later lineage B sequences and suggest convergent evolution or bioinformatics artifact. The remaining two genomes do not have any additional mutations.

Lastly, genome EPI_ISL_413017, which was collected on 6 February 2020, is in Fig. 3A and Fig. 3B; in addition to low coverage at 28144, this genome likely either suffers from contamination or is a recombinant (*90*).

Overall, although the Diamond Princess cruise ship outbreak represents convergent evolution at 8782, the remainder of the C/C and T/T genomes through 28 February likely appear as intermediate haplotypes because of sequencing or bioinformatics artifacts.

T/T genomes in New York City. We found 3 SARS-CoV-2 genomes with both C8782T and T28144 in the NYC Public Health Laboratory surveillance data set. We placed these genomes on a global tree of 3 million genomes (v2022-01-21) using UShER, and all 3 appear to be descendants of lineage B (*91*). Two of these genomes (EPI_ISL_8953704 and EPI_ISL_8953705), which are in the B.1.526 *Iota* lineage, have identical sequences suggesting that a T/T genome may have been transmitted locally (Fig. S4). We note that although these genomes were sampled from different individuals in different parts of the city, their genomes were sequenced on the same sequencing plate. The third NYC T/T genome (EPI_ISL_1447116), in the B.1.2 lineage, falls on a different part of the phylogeny, indicating an independent C8782T mutation. This third genome was sequenced on a separate run than the other two NYC T/T genomes. All three genomes at site 8782 and 28144 have read depth >4,000x, with coverage in both directions and 100% support for the T allele.

T/T and C/C genomes in San Diego. We found 24 T/T genomes in the San Diego SEARCH data set, with collection dates between December 2020 and December 2021and 1 C/C genome that was collected on 15 January 2021. We placed these genomes on a global tree of 3 million genomes using UShER. Eight of the T/T genomes were classified as Delta sublineages: AY.26, AY.40, and AY.44; the remaining T/T genomes were classified as B.1 and its other descendant lineages (Fig. S5). The C/C genome was classified as B.1.1.432 (Fig. S6). Therefore, these and other T/T and C/C haplotypes are the result of convergent evolution at 8782 and 28144, respectively, and do not represent the ancestral or transitional forms between lineages A and B.

Justification for a non-reversible substitution model. We constructed a random-effects non-reversible substitution model for our phylogenetic inference because of the substantial C-to-T transition bias (*21–23*) and frequent C-to-T reversions (described in Main Text). To compare the ancestral haplotype inference between the random-effects model and a standard reversible substitution model, we repeated our Bayesian phylodynamic inference with a GTR substitution model with both the unconstrained and recCA-rooted analyses. We find that C/C and T/T ancestral haplotypes were less common under the GTR model than the random-effects model (Data S2). Notably, the difference in posterior support for C/C and T/T was negligible under the unconstrained GTR model, indicating the increased level of biological realism reflected in the random-effects substitution model inference. Alternative ancestral haplotypes such as A.1 (BF>150) and A+C29095T (BF>50) were poorly supported under the GTR model, just as in the unconstrained and the recCA random-effects model.

The tMCRA of SARS-CoV-2 is consistent across likely ancestral haplotypes. It has been suggested that a phylogenetic root in lineage A would produce older tMRCA estimates than a lineage B rooting (*18*). However, we find that SARS-CoV-2 tMRCA inference is generally robust to the rooting model and ancestral haplotype.

The unconstrained model (Fig. S24A), which favored a lineage B ancestral haplotype (Table S2), and produced a median tMRCA of 11 December 2019 [95% highest posterior density (HPD): 25 November to 20 December] and a mean substitution rate of $9.2\text{x}10^{-4}$ substitutions/site/year (95% HPD: $8.1\text{x}10^{-4}$ to $1.0\text{x}10^{-3}$). These tMRCA estimates are similar to our previous inference (*9*), although the substitution is slightly faster. This elevated rate is expected, given that shorter sampling windows are associated with more rapid substitution rate inference in SARS-CoV-2 (*9*, *92*, *93*). The recCA-constrained model (Fig. 4C), which favored a lineage A ancestral haplotype (Table 1), produced a median tMRCA of 6 December 2019 (95% HPD: 15 November to 19 December) and a mean substitution rate of $9.2\text{x}10^{-4}$ substitutions/site/year (95% HPD: $8.0\text{x}10^{-4}$ to $1.0\text{x}10^{-3}$).

To explicitly explore the effect of ancestral haplotype on the SARS-CoV-2 tMRCA, we employed our novel phylodynamic framework that fixes the MRCA of the SARS-CoV-2 phylogeny to ancestral haplotypes (Fig. 4D, see methods), rather than using sampled taxon (*e.g.*, Hu-1), an outgroup (*e.g.*, RaTG13), or their inferred ancestor (*e.g.*, recCA). We explored the plausible ancestral haplotypes (lineage A, lineage B, and C/C), as well A.1 (WA1) and A.1 + C29095T (20SF012) (see methods). The tMRCAs ranged between 12 December (95% HPD 27 November to 19 December) for a Lineage

B root to 4 December (95% HPD 16 November to 18 December) for an A.1 root (Fig. S24, Table S2). Therefore, ancestral haplotype has minimal impact on tMRCA inference.

Selecting index case dates. Although early reports suggested an index case (*i.e.*, first identified case) dating to either 17 November 2019 (*94*) or 1 December 2019 (*95*), the World Health Organization (WHO)-China report did not find evidence to support the veracity these cases and identified the earliest case as having an illness onset of 8 December (case S01 from Table 6 in the WHO report) (*13*). A subsequent analysis of the earliest COVID-19 cases suggested that the '8 December' patient actually became ill on 16 December and concluded that the index case was a vendor from the Huanan Seafood Market who became ill on 10 December (*7*). This shift in index case dates necessitates reexamining the timing of the primary case (*i.e.*, first case resulting from a cross-species transmission) of SARS-CoV-2, as case data is crucial to timing the first SARS-CoV-2 infection (*9*).

The seafood vendor who became ill on 10 December was hospitalized on 16 December (*7*); we condition on their symptom onset and hospitalization dates when assuming a single introduction of SARS-CoV-2.

When we perform rejection sampling on lineages A and B separately, we need to use case and hospitalization dates associated with each lineage. Although the SARS-CoV-2 index case does not have an associated published genome, we can reasonably assume this individual had a lineage B virus, as all viruses associated with the Huanan market, including an environmental sample from the stall this vendor operated (EPI_ISL_408512), were lineage B. We therefore use the illness onset and hospitalization dates belonging to the index case for the lineage B to time the primary case of lineage B.

The earliest lineage B case with a published genome (IPBCAMS-WH-01; GISAID accession EPI_ISL_402123; case S02 from Table 6 in the WHO report) became symptomatic and was hospitalized on 13 and 18 December, respectively (*2*). Although this individual was initially reported to have an illness onset date of 15 December, the WHO report subsequently determined he had an earlier onset of 13 December (*13*). We use the dates belonging to the earliest confirmed lineage B case for a sensitivity analysis and find that the timing of the primary case is robust (Table S3).

The earliest case and hospitalization dates for lineage A belong to 'Cluster 1' from the WHO report (Annex E2) (*13*). The lineage A genome belongs to the individual in the cluster with illness onset on 26 December (case S13 from Table 6 in the WHO report; IME-WH01; GISAID accession EPI_ISL_529213). However, the spouse of this individual was infected as well, becoming symptomatic on 15 December and hospitalized on 25 December. We can therefore reasonably assume the spouse was also infected with a lineage A virus and subsequently infected the earliest confirmed lineage A case. We thus use the dates belonging to the spouse when timing the primary case of lineage A.

For our additional sensitivity analyses, we examine the effect of conditioning on 8 December as an index case date for both the single- and lineage-B-introductions to reflect the previously discounted case date of the '8 December' patient (case S01 from the WHO report). The timings of the primary case for both a single- and lineage-B-introductions were robust to this earlier index case date (Table S9).

Robustness of the timing of the primary cases and epidemiological dynamics of multiple introductions when assuming a slower doubling time. Although the doubling time of SARS-CoV-2 was likely less than 3 days before it was identified as the etiological agent of COVID-19 and non-pharmaceutical interventions were implemented (*25*, *26*), we sought to validate our multi-introduction results using a greater doubling time (*i.e.*, slower growth rate). When using a doubling time of 3.47 days (95% range: 1.53-5.78) and the unconstrained model, we time the primary cases for lineages A and B to 25 November (95% HPD: 31 October to 13 December) and 21 November (24 October to 11 December), respectively. These results (Fig. S22) are approximately 1 week earlier than using the faster doubling time described in the Main Text (Table S9). The timing of the primary case of lineage B consistently occurs before that of lineage A (Fig. S22). Due to the slightly earlier primary case, the epidemiological dynamics of the slower doubling time simulations match those of the simulations in the Main Text in December 2019 and January 2020, with lineage B similarly predominating in the early pandemic (Fig. S23). Therefore, our primary case timing and the relationship between lineages A and B in the simulations are robust to a slower growth rate.

Timing the primary case of a single SARS-CoV-2 introduction scenario. As a counterfactual scenario, we examined the timing of the primary case for a single introduction. Here, we condition on the SARS-CoV-2 tMRCA (described in the Main Text) and the ascertainment (10 December) and hospitalization (16 December) dates belonging to the SARS-CoV-2 index case. Under this scenario, we infer the primary case to have acquired SARS-CoV-2 on 25 November 2019 (95% HPD: 3 November to 8 December). Across the plausible ancestral haplotypes of lineage A, lineage B, and C/C, the inferred timing of this primary case is unchanged. Even when using the ancestral haplotype producing the oldest tMRCA (A.1), the timing of the primary case remains consistent (Fig. S24; Table S9). Lastly, the timing of the primary case in a single-introduction scenario is robust to the increased doubling time of 3.47 days (95% range: 1.53-5.78), resulting in the timing of the primary case to shift approximately 7 days earlier: 18 November (95% HPD: 22 October to 8 December). As the symptom onset and hospitalization dates are the same for a single introduction and the lineage B introduction, the slightly earlier timing of the primary for a single introduction is due to the earlier overall tMRCA relative to the lineage B tMRCA (Table S2). In sum, even if we assume that SARS-CoV-2 was the result of a singular introduction into humans, the primary case of SARS-CoV-2 was likely infected in November, matching the timing of the primary case of lineage B (Table S9, S4).

Epidemiological dynamics of a single SARS-CoV-2 introduction scenario. The dated infections and hospitalizations resulting from a single introduction are very similar to the combined infections and hospitalizations from lineages A and B (Fig. 7A, S21A). In fact, the slightly earlier timing of the primary case in a single-introduction scenario, coupled with exponential growth, can cause a single introduction to lead to more infections and hospitalizations than multiple introductions (Fig. 7A, S21A, Table S6, S7). Lastly, the numbers of infections and hospitalizations are also consistent across ancestral haplotypes (Fig. S24), as would be expected with a consistent timing of the primary case.

Consistent timing of the primary case when conditioning only on hospitalization. As the index case date has been uncertain, we tested our rejection sampling approach by conditioning just on the tMRCA and the date of the earliest hospitalization. The timing of the primary case was robust in both single- and multi-introduction scenarios (*i.e.*, lineages A and B), indicating that the index case dates did not bias our results toward earlier dates (Table S9-5). Additionally, if the identified hospitalizations are among the earliest hospitalizations in the pandemic, consistent results between approaches including and excluding the index case dates suggests cases were not extensively missed prior to the index case.

Minimal cryptic circulation before December 2019. Although we do not see any evidence for substantial cryptic circulation before December 2019 with the epidemic simulations (Fig. 7), we can quantify the expected number of infections before the tMRCA. To do so, we calculate the cumulative number of infections in the epidemic simulations once they reach stable coalescence, the point in time at which basal lineages cease to be lost. Importantly, the time to stable coalescence is the equivalent to the tMRCA for the epidemic simulations. There were, at most, 40 cumulative infections by the time of stable coalescence in the simulated epidemics, and 99.1% of simulated epidemics reached stable coalescence by 20 cumulative infections. With the tMRCA of lineage B, likely the first lineage of SARS-CoV-2 introduced into humans, estimated to 13 December (95% HPD: 29 November to 23 December), we would not expect more than a few dozen infections before 10 December, the date of the SARS-CoV-2 index case.

Similarities to WA1 and WA outbreak clades. To understand the phylogenetic signal of a hypothetical singular introduction of SARS-CoV-2 into humans, it is helpful to seek an analogy with the introduction of SARS-CoV-2 into North America. The first confirmed case of SARS-CoV-2 in the U.S. was associated with a virus strain ('WA1') isolated in Washington State from a traveler who returned from Wuhan, China, on 15 January 2020. There was subsequently an outbreak (henceforth, 'WA outbreak clade') in Washington State, with cases confirmed starting in February 2020 (*28*). As we have previously shown, although the MRCA of the WA outbreak clade differed from WA1 by only two substitutions, the WA1 WA outbreak clades were, in fact, separate introductions into Washington State (*16*). The WA outbreak clade showcases a basal polytomy, and although the WA1 introduction was contained, onward transmission would have likely led to a basal polytomy as well, as shown by the hypothetical polytomy in Fig. S18. Therefore, this pattern is remarkably similar to that seen with lineages A and B, with the exception of a successful prevention of onward transmission from the WA1 case: the MRCA of WA1 and the WA outbreak clades was in China, and as we have shown here, the MRCA of lineages A and B was likely in the intermediate host reservoir (Fig. 8). Both scenarios show introductions of SARS-CoV-2 from a prior location: China in the case of WA1 and the WA outbreak clades, and the intermediate host reservoir in the case of lineages A and B. Similarly, both scenarios lead to (or would lead to, in the case of WA1) basal polytomies from the onward transmission.

Therefore, lineages A and B look like separate introductions because introductions with sustained onward transmission result in large basal polytomies.

Number of introductions needed for two successful introductions. We validated our simulation results estimating the total number of introductions needed for two successful ones (described in the Main Text) using the negative binomial distribution. The probability of both introductions being successful when there are only two introductions is 0.11, 5 introductions leading to at least two successful introductions has a probability of 0.54, and 15 introductions leading to at least two successful introductions has a probability of 0.98. These results match our simulations, which had a median number of 5 total introductions leading to two successful introductions (95% CI: 2–15).

Table S1. Variant calls at positions 8782 and 28144 for three SARS-CoV-2 genomes with a T/T sequence[1].

GISAID accession	8782									28144								
	Depth	Count				Proportion				Depth	Count				Proportion			
		A	C	G	T	A	C	G	T		A	C	G	T	A	C	G	T
EPI_ISL_493179	64	0	39	1	24	0.000	0.609	0.016	0.375	61361	121	3784	195	57261	0.002	0.062	0.003	0.933
EPI_ISL_493180	40	0	24	1	15	0.000	0.600	0.025	0.375	95374	226	5709	293	89146	0.002	0.060	0.003	0.935
EPI_ISL_493182	29	0	10	0	19	0.000	0.345	0.000	0.655	69369	153	4051	232	64933	0.002	0.058	0.003	0.936

[1] Variant calls and depths provided by Di Liu and Yi Yan.

Table S2. tMRCA inferences for SARS-CoV-2, lineage B, and lineage A under different rooting strategies.

Phylodynamic analysis	SARS-CoV-2[1]	Lineage B[1]	Lineage A[1]
Unconstrained	12-11 (11-25 to 12-20)	12-13 (11-29 to 12-23)	12-25 (12-17 to 12-30)
recCA	12-06 (11-15 to 12-19)	12-15 (12-05 to 12-23)	12-20 (12-05 to 12-29)
Lineage B	12-12 (11-27 to 12-19)	12-13 (11-29 to 12-21)	12-25 (12-18 to 12-29)
C/C	12-08 (11-19 to 12-19)	12-16 (12-06 to 12-23)	12-21 (12-12 to 12-29)
T/T	12-08 (11-19 to 12-19)	12-15 (12-06 to 12-23)	12-21 (12-12 to 12-29)
Lineage A	12-07 (11-18 to 12-19)	12-16 (12-08 to 12-23)	12-18 (12-04 to 12-28)
Lineage A + C29095T	12-05 (11-17 to 12-19)	12-17 (12-10 to 12-23)	12-05 (11-17 to 12-19)
Lineage A.1	12-04 (11-16 to 12-18)	12-16 (12-10 to 12-23)	12-04 (11-16 to 12-19)
No markets	12-13 (11-25 to 12-26)	12-17 (11-29 to 12-26)	12-25 (12-15 to 01-04)

[1] Median and 95% HPD in parentheses.

Table S3. Time of the lineage B primary case under different robustness analyses for different lineage B index case and hospitalization dates, conditioning only on hospitalization dates, and epidemic simulations with slower transmission rate.

Phylodynamic analysis	Robustness analysis[1]				
	Case: 13 Dec. Hosp: 18 Dec.	**Case: 10 Dec. Hosp: 16 Dec.**	**Case: 8 Dec. Hosp: 18 Dec.**	**Only hosp: 18 Dec.**	**Case: 13 Dec. Hosp: 18 Dec. Slow transmission**
Unconstrained	11-28 (11-05 to 12-11)	11-25 (11-04 to 12-08)	11-25 (11-03 to 12-07)	11-28 (11-05 to 12-12)	11-21 (10-24 to 12-11)
recCA	11-28 (11-06 to 12-11)	11-26 (11-05 to 12-08)	11-25 (11-04 to 12-07)	11-29 (11-06 to 12-12)	11-21 (10-25 to 12-11)
Lineage B	11-28 (11-05 to 12-11)	11-25 (11-04 to 12-08)	11-25 (11-04 to 12-07)	11-28 (11-05 to 12-12)	11-21 (10-24 to 12-11)
C/C	11-28 (11-06 to 12-11)	11-26 (11-05 to 12-08)	11-25 (11-04 to 12-07)	11-29 (11-07 to 12-13)	11-21 (10-25 to 12-11)
Lineage A	11-28 (11-06 to 12-11)	11-26 (11-05 to 12-08)	11-25 (11-04 to 12-07)	11-29 (11-07 to 12-12)	11-21 (10-24 to 12-11)

Hosp, Hospitalization.

[1] Median and 95% HPD in parentheses.

Table S4. Time of the lineage A primary case results under different robustness analyses, including conditioning only on hospitalization dates and epidemic simulations with slower transmission rate.

Phylodynamic analysis	**Robustness analysis**[1]		
	Case: 15 Dec. **Hosp: 25 Dec.**	**Only hosp: 25 Dec.**	**Case: 15 Dec.** **Hosp: 25 Dec.** **Slow transmission**
Unconstrained	12-02 (11-12 to 12-13)	12-07 (11-14 to 12-18)	11-25 (10-31 to 12-13)
recCA	12-02 (11-10 to 12-14)	12-05 (11-12 to 12-19)	11-25 (10-29 to 12-14)
Lineage B	12-02 (11-13 to 12-14)	12-07 (11-14 to 12-18)	11-25 (11-01 to 12-13)
C/C	12-02 (11-11 to 12-14)	12-06 (11-13 to 12-19)	11-25 (10-30 to 12-13)
Lineage A	12-01 (11-09 to 12-14)	12-05 (11-12 to 12-19)	11-25 (10-30 to 12-14)

Hosp, Hospitalization

[1] Median and 95% HPD in parentheses.

Table S5. Number of days the timing of the primary case of lineage A occurs after the timing of the primary case of lineage B.

Phylodynamic analysis	**Median difference (95% HPD)**[1]
Unconstrained	7.0 (-20.5 to 35.1)
recCA	5.6 (-22.8 to 33.7)
Lineage B	7.2 (-20.3 to 34.9)
C/C	6.3 (-21.6 to 34.3)
Lineage A	5.7 (-22.7 to 33.8)

[1] See Fig. 6D for graphical representation of the full distribution.

Table S6. Number of estimated infections on 1 December 2019 for lineage A, lineage B, lineages A and B combined, and single introduction simulations.

Phylodynamic analysis	**Introductions[1]**			
	Lineage A	**Lineage B**	**Lineages A and B combined**	**Single introduction**
Unconstrained	0 (0, 6)	2 (0, 12)	3 (0, 15)	2 (0, 15)
recCA	0 (0, 9)	1 (0, 11)	3 (0, 14)	3 (0, 104)
Lineage B	0 (0, 6)	1 (0, 12)	3 (0, 14)	2 (0, 13)
C/C	0 (0, 7)	1 (0, 11)	3 (0, 13)	3 (0, 37)
Lineage A	1 (0, 9)	1 (0, 11)	3 (0, 14)	3 (0, 53)

[1] Median and 95% HPD in parentheses.

Table S7. Number of estimated infections on 15 December 2019 for lineage A, lineage B, lineages A and B combined, and single introduction simulations.

Phylodynamic analysis	**Introductions**[1]			
	Lineage A	**Lineage B**	**Lineages A and B combined**	**Single introduction**
Unconstrained	4 (1, 14)	9 (1, 361)	16 (3, 373)	11 (2, 859)
recCA	5 (1, 46)	8 (1, 61)	15 (3, 207)	25 (2, 8278)
Lineage B	4 (1, 14)	10 (1, 358)	16 (2, 365)	10 (2, 444)
C/C	4 (1, 18)	7 (1, 48)	14 (3, 69)	16 (2, 3308)
Lineage A	5 (1, 53)	7 (2, 35)	14 (3, 129)	19 (2, 4411)

[1] Median and 95% HPD in parentheses.

Table S8. Simulation parameters, with all parameters except *b* and *h* based on Hao et al. (*27*). The *b* value listed is for the main analysis. The *h* parameter is not included in the Hao et al. model.

Parameter	Definition	Value
b	Transmission rate of ascertained cases	0.38
r	Ascertainment rate	0.15
a	Ratio of transmission for unascertained	0.55
h	Hospitalization rate	0.5
D_e	Latent period (days)	2.9
D_p	Presymptomatic infectious period (days)	2.3
D_i	Symptomatic infectious period (days)	2.9
D_q	Duration from illness onset to isolation (hospitalization) (days)	11
D_h	Isolation (hospitalization) period (days)	30

Table S9. Time of the primary case of a hypothetical single-introduction scenario under different robustness analyses for different index case dates, conditioning on only hospitalization dates, and epidemic simulations with slower transmission rate.

Phylodynamic analysis	Robustness analysis[1]			
	Case: 10 Dec. Hosp: 16 Dec.	Case: 8 Dec. Hosp: 16 Dec.	Only hosp: 16 Dec.	Case: 10 Dec. Hosp: 16 Dec. Slow transmission
Unconstrained	11-25 (11-03 to 12-08)	11-24 (11-02 to 12-06)	11-26 (11-03 to 12-10)	11-18 (10-22 to 12-08)
recCA	11-23 (10-30 to 12-08)	11-22 (10-29 to 12-06)	11-24 (10-30 to 12-10)	11-18 (10-20 to 12-08)
Lineage B	11-25 (11-03 to 12-08)	11-24 (11-03 to 12-06)	11-27 (11-04 to 12-09)	11-19 (10-23 to 12-09)
C/C	11-24 (11-01 to 12-08)	11-23 (10-31 to 12-06)	11-25 (11-01 to 12-10)	11-18 (10-20 to 12-08)
T/T	11-24 (11-01 to 12-08)	11-23 (10-31 to 12-06)	11-25 (11-01 to 12-10)	11-18 (10-20 to 12-08)
Lineage A	11-24 (10-31 to 12-08)	11-23 (10-31 to 12-06)	11-25 (10-31 to 12-10)	11-18 (10-20 to 12-08)
Lineage A + C29095T	11-23 (10-31 to 12-08)	11-22 (10-30 to 12-06)	11-24 (10-31 to 12-10)	11-18 (10-19 to 12-08)
Lineage A.1	11-23 (10-30 to 12-08)	11-22 (10-29 to 12-06)	11-24 (10-30 to 12-10)	11-18 (10-19 to 12-08)

Hosp, Hospitalization

[1] Median and 95% HPD in parentheses.

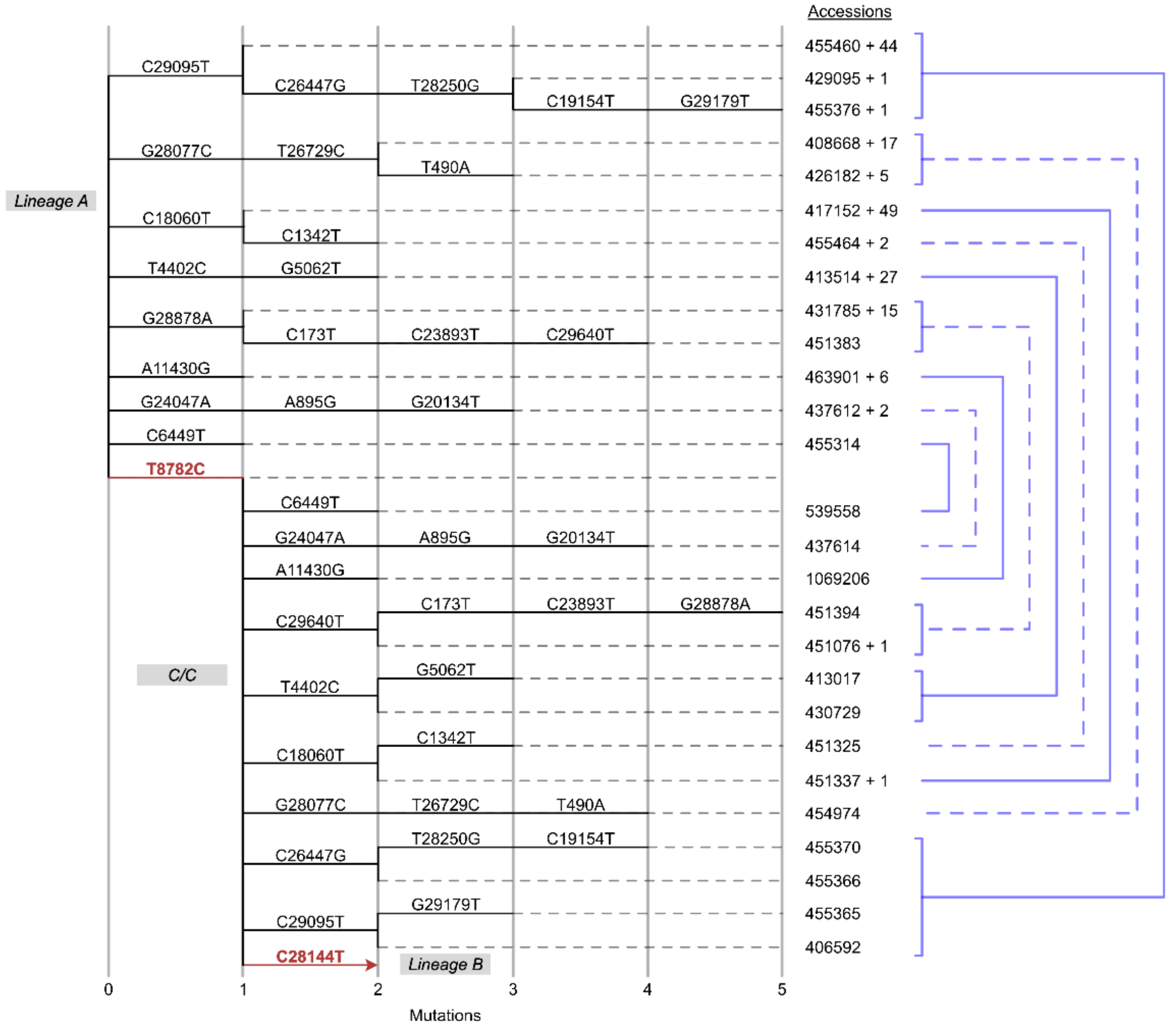

Figure S1. Phylogeny of SARS-CoV-2 intermediate C/C genomes and their shared mutations within lineage A through 28 February 2020. Mutations relative to the Hu-1 reference genome are shown above each branch, comparing lineage A and C/C. Lineage defining mutations are colored in red. Derived mutations that are not shared by both lineages are excluded. Branches are connected to taxon names with horizontal dashed lines. The taxon names are GISAID accession numbers, and in cases where more than one sequence is represented, the total number of additional matching homoplastic sequences is indicated after the "+" symbol. Sequences that share derived mutations are connected by the lines on the right, and brackets indicate that a group of sequences share the derived mutations that cannot be individually resolved. For clarity, lines are presented as alternating dashed and solid. Several lineage A genomes with ambiguous placement were also excluded for clarity.

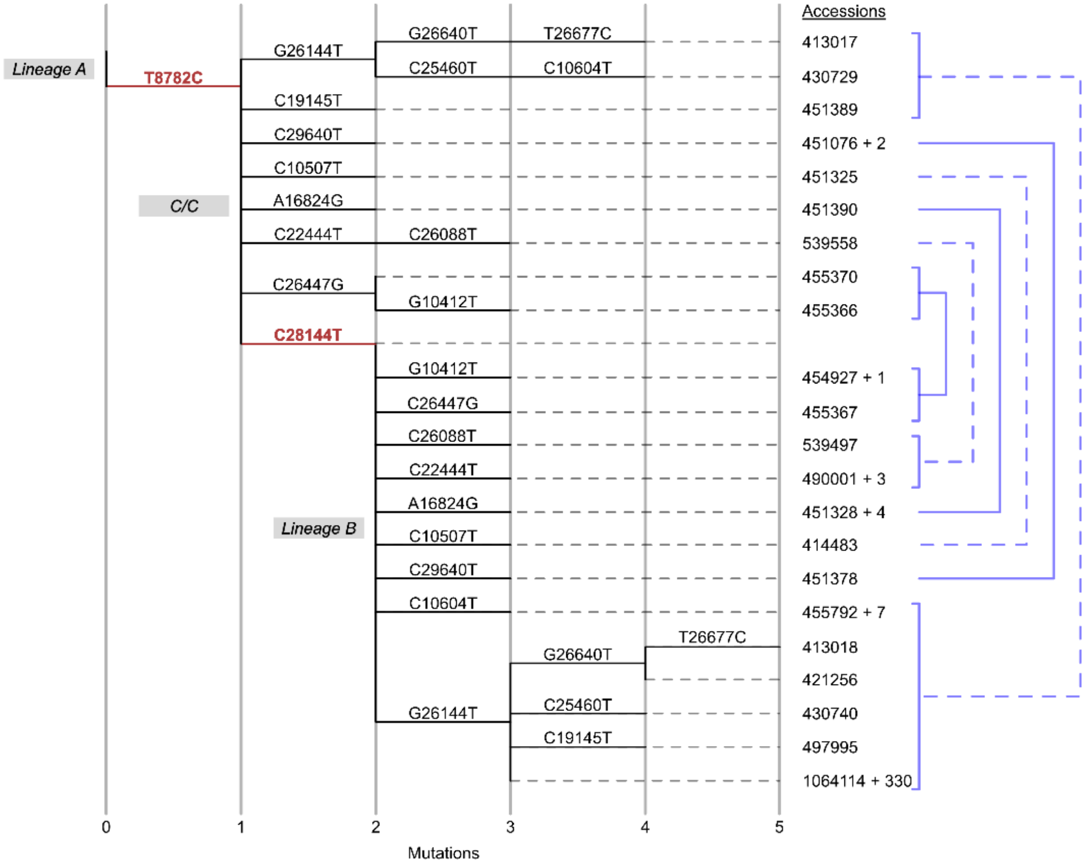

Figure S2. Phylogeny of SARS-CoV-2 intermediate C/C genomes and their shared mutations within lineage B through 28 February 2020. Mutations relative to the Hu-1 reference genome are shown above each branch, comparing lineage B and C/C. Refer to Fig. S1 for detailed explanation of format.

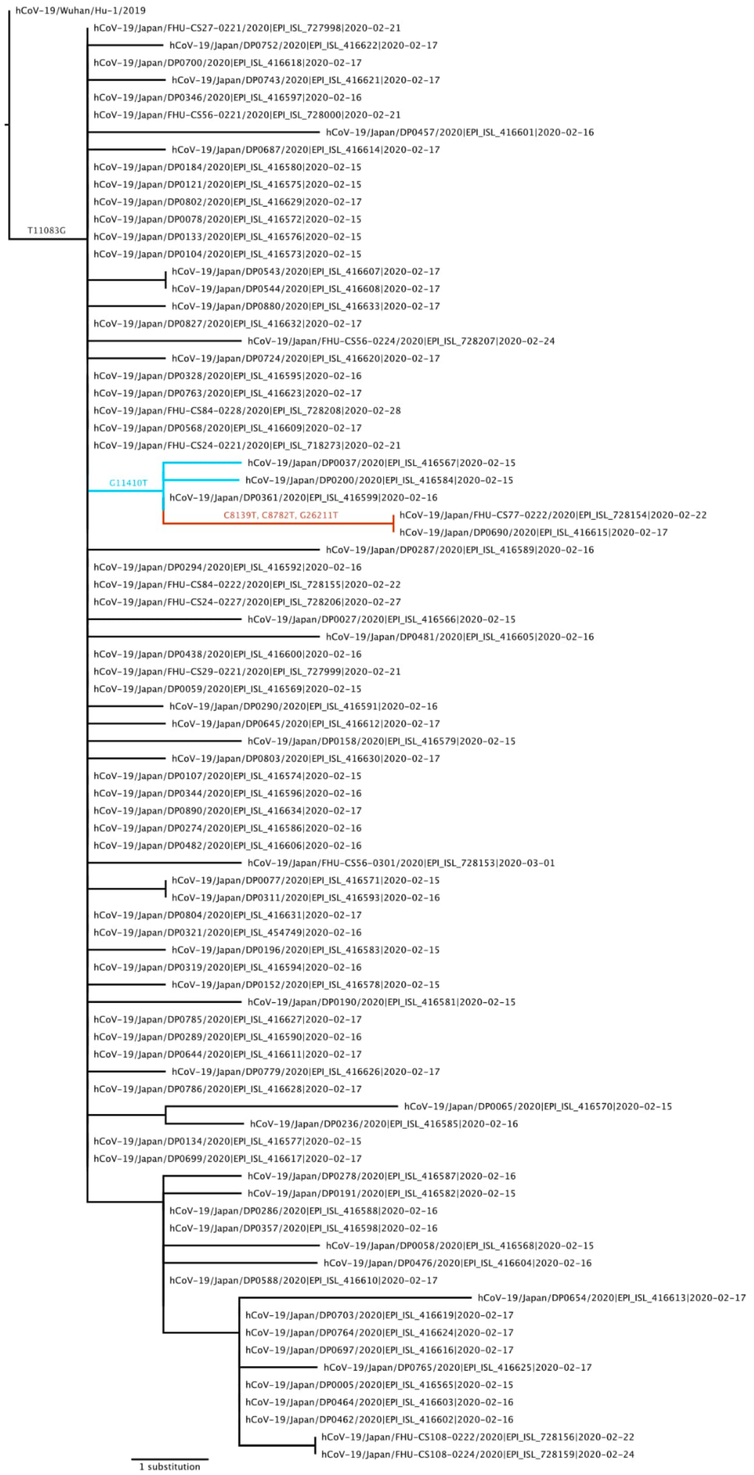
hCoV-19/Wuhan/Hu-1/2019
hCoV-19/Japan/FHU-CS27-0221/2020|EPI_ISL_727998|2020-02-21
hCoV-19/Japan/DP0752/2020|EPI_ISL_416622|2020-02-17
hCoV-19/Japan/DP0700/2020|EPI_ISL_416618|2020-02-17
hCoV-19/Japan/DP0743/2020|EPI_ISL_416621|2020-02-17
hCoV-19/Japan/DP0346/2020|EPI_ISL_416597|2020-02-16
hCoV-19/Japan/FHU-CS56-0221/2020|EPI_ISL_728000|2020-02-21
hCoV-19/Japan/DP0457/2020|EPI_ISL_416601|2020-02-16
hCoV-19/Japan/DP0687/2020|EPI_ISL_416614|2020-02-17
hCoV-19/Japan/DP0184/2020|EPI_ISL_416580|2020-02-15
hCoV-19/Japan/DP0121/2020|EPI_ISL_416575|2020-02-15
hCoV-19/Japan/DP0802/2020|EPI_ISL_416629|2020-02-17
hCoV-19/Japan/DP0078/2020|EPI_ISL_416572|2020-02-15
T11083G
hCoV-19/Japan/DP0133/2020|EPI_ISL_416576|2020-02-15
hCoV-19/Japan/DP0104/2020|EPI_ISL_416573|2020-02-15
hCoV-19/Japan/DP0543/2020|EPI_ISL_416607|2020-02-17
hCoV-19/Japan/DP0544/2020|EPI_ISL_416608|2020-02-17
hCoV-19/Japan/DP0880/2020|EPI_ISL_416633|2020-02-17
hCoV-19/Japan/DP0827/2020|EPI_ISL_416632|2020-02-17
hCoV-19/Japan/FHU-CS56-0224/2020|EPI_ISL_728207|2020-02-24
hCoV-19/Japan/DP0724/2020|EPI_ISL_416620|2020-02-17
hCoV-19/Japan/DP0328/2020|EPI_ISL_416595|2020-02-16
hCoV-19/Japan/DP0763/2020|EPI_ISL_416623|2020-02-17
hCoV-19/Japan/FHU-CS84-0228/2020|EPI_ISL_728208|2020-02-28
hCoV-19/Japan/DP0568/2020|EPI_ISL_416609|2020-02-17
hCoV-19/Japan/FHU-CS24-0221/2020|EPI_ISL_718273|2020-02-21
hCoV-19/Japan/DP0037/2020|EPI_ISL_416567|2020-02-15
hCoV-19/Japan/DP0200/2020|EPI_ISL_416584|2020-02-15
G11410T
hCoV-19/Japan/DP0361/2020|EPI_ISL_416599|2020-02-16
C8139T, C8782T, G26211T
hCoV-19/Japan/FHU-CS77-0222/2020|EPI_ISL_728154|2020-02-22
hCoV-19/Japan/DP0690/2020|EPI_ISL_416615|2020-02-17
hCoV-19/Japan/DP0287/2020|EPI_ISL_416589|2020-02-16
hCoV-19/Japan/DP0294/2020|EPI_ISL_416592|2020-02-16
hCoV-19/Japan/FHU-CS84-0222/2020|EPI_ISL_728155|2020-02-22
hCoV-19/Japan/FHU-CS24-0227/2020|EPI_ISL_728206|2020-02-27
hCoV-19/Japan/DP0027/2020|EPI_ISL_416566|2020-02-15
hCoV-19/Japan/DP0481/2020|EPI_ISL_416605|2020-02-16
hCoV-19/Japan/DP0438/2020|EPI_ISL_416600|2020-02-16
hCoV-19/Japan/FHU-CS29-0221/2020|EPI_ISL_727999|2020-02-21
hCoV-19/Japan/DP0059/2020|EPI_ISL_416569|2020-02-15
hCoV-19/Japan/DP0290/2020|EPI_ISL_416591|2020-02-16
hCoV-19/Japan/DP0645/2020|EPI_ISL_416612|2020-02-17
hCoV-19/Japan/DP0158/2020|EPI_ISL_416579|2020-02-15
hCoV-19/Japan/DP0803/2020|EPI_ISL_416630|2020-02-17
hCoV-19/Japan/DP0107/2020|EPI_ISL_416574|2020-02-15
hCoV-19/Japan/DP0344/2020|EPI_ISL_416596|2020-02-16
hCoV-19/Japan/DP0890/2020|EPI_ISL_416634|2020-02-17
hCoV-19/Japan/DP0274/2020|EPI_ISL_416586|2020-02-16
hCoV-19/Japan/DP0482/2020|EPI_ISL_416606|2020-02-16
hCoV-19/Japan/FHU-CS56-0301/2020|EPI_ISL_728153|2020-03-01
hCoV-19/Japan/DP0077/2020|EPI_ISL_416571|2020-02-15
hCoV-19/Japan/DP0311/2020|EPI_ISL_416593|2020-02-16
hCoV-19/Japan/DP0804/2020|EPI_ISL_416631|2020-02-17
hCoV-19/Japan/DP0321/2020|EPI_ISL_454749|2020-02-16
hCoV-19/Japan/DP0196/2020|EPI_ISL_416583|2020-02-15
hCoV-19/Japan/DP0319/2020|EPI_ISL_416594|2020-02-16
hCoV-19/Japan/DP0152/2020|EPI_ISL_416578|2020-02-15
hCoV-19/Japan/DP0190/2020|EPI_ISL_416581|2020-02-15
hCoV-19/Japan/DP0785/2020|EPI_ISL_416627|2020-02-17
hCoV-19/Japan/DP0289/2020|EPI_ISL_416590|2020-02-16
hCoV-19/Japan/DP0644/2020|EPI_ISL_416611|2020-02-17
hCoV-19/Japan/DP0779/2020|EPI_ISL_416626|2020-02-17
hCoV-19/Japan/DP0786/2020|EPI_ISL_416628|2020-02-17
hCoV-19/Japan/DP0065/2020|EPI_ISL_416570|2020-02-15
hCoV-19/Japan/DP0236/2020|EPI_ISL_416585|2020-02-16
hCoV-19/Japan/DP0134/2020|EPI_ISL_416577|2020-02-15
hCoV-19/Japan/DP0699/2020|EPI_ISL_416617|2020-02-17
hCoV-19/Japan/DP0278/2020|EPI_ISL_416587|2020-02-16
hCoV-19/Japan/DP0191/2020|EPI_ISL_416582|2020-02-15
hCoV-19/Japan/DP0286/2020|EPI_ISL_416588|2020-02-16
hCoV-19/Japan/DP0357/2020|EPI_ISL_416598|2020-02-16
hCoV-19/Japan/DP0058/2020|EPI_ISL_416568|2020-02-15
hCoV-19/Japan/DP0476/2020|EPI_ISL_416604|2020-02-16
hCoV-19/Japan/DP0588/2020|EPI_ISL_416610|2020-02-17
hCoV-19/Japan/DP0654/2020|EPI_ISL_416613|2020-02-17
hCoV-19/Japan/DP0703/2020|EPI_ISL_416619|2020-02-17
hCoV-19/Japan/DP0764/2020|EPI_ISL_416624|2020-02-17
hCoV-19/Japan/DP0697/2020|EPI_ISL_416616|2020-02-17
hCoV-19/Japan/DP0765/2020|EPI_ISL_416625|2020-02-17
hCoV-19/Japan/DP0005/2020|EPI_ISL_416565|2020-02-15
hCoV-19/Japan/DP0464/2020|EPI_ISL_416603|2020-02-16
hCoV-19/Japan/DP0462/2020|EPI_ISL_416602|2020-02-16
hCoV-19/Japan/FHU-CS108-0222/2020|EPI_ISL_728156|2020-02-22
hCoV-19/Japan/FHU-CS108-0224/2020|EPI_ISL_728159|2020-02-24
1 substitution

Figure S3. Maximum likelihood phylogeny of SARS-CoV-2 genomes from the Diamond Princess outbreak. The tree is rooted on Hu-1. Substitutions found in T/T genomes relative to Hu-1 annotated on branches. The G11410T clade is colored blue, with the branch leading to the T/T genomes colored red.

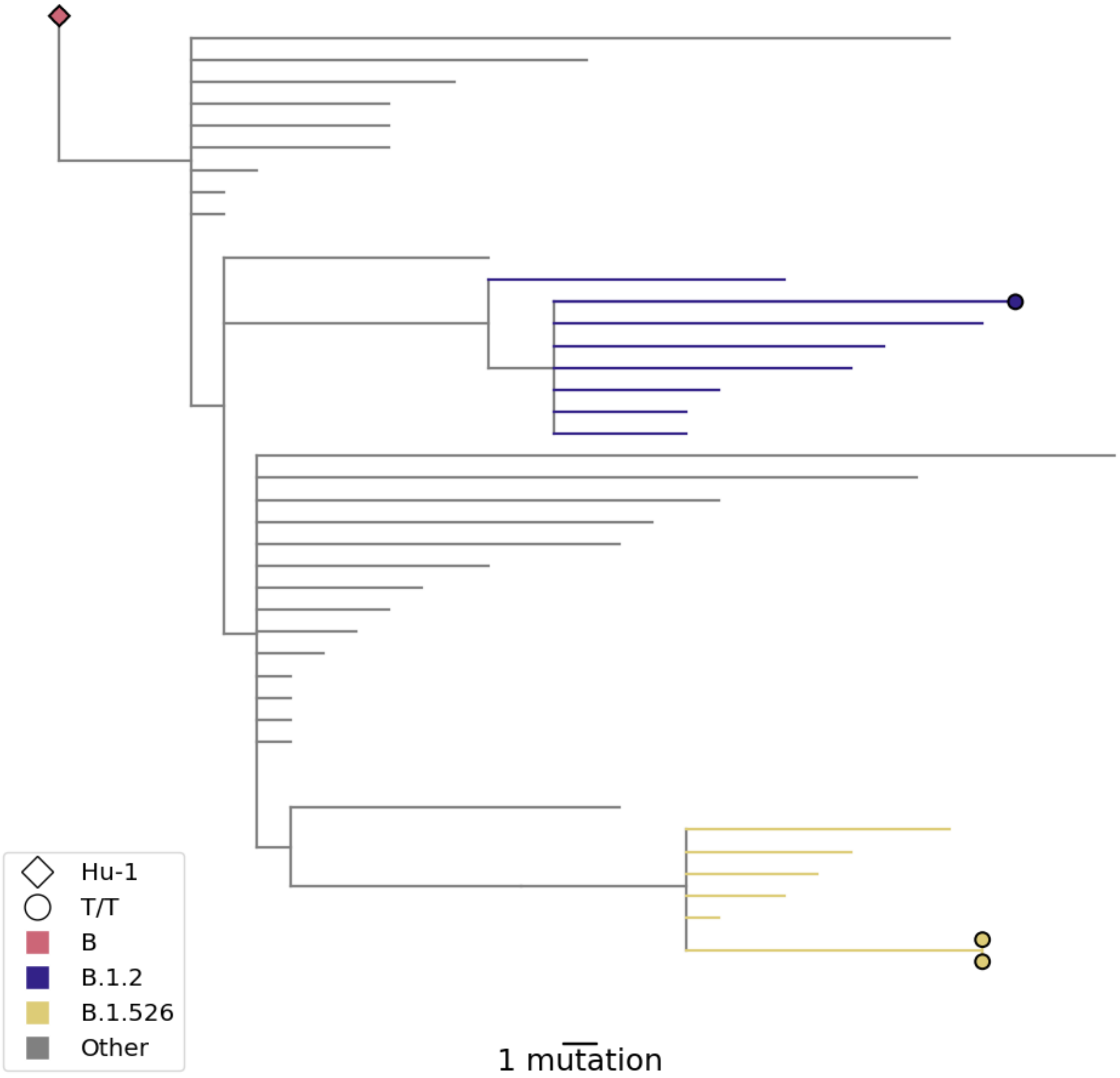

Figure S4. Subtree showing the placement of T/T genomes in the NYC Public Health Laboratories dataset. 3 T/T genomes were placed on a global tree of 3 million genomes (v2022-01-21) using UShER. The node branches are colored by the assigned PANGO lineage. The T/T genomes are highlighted using circles.

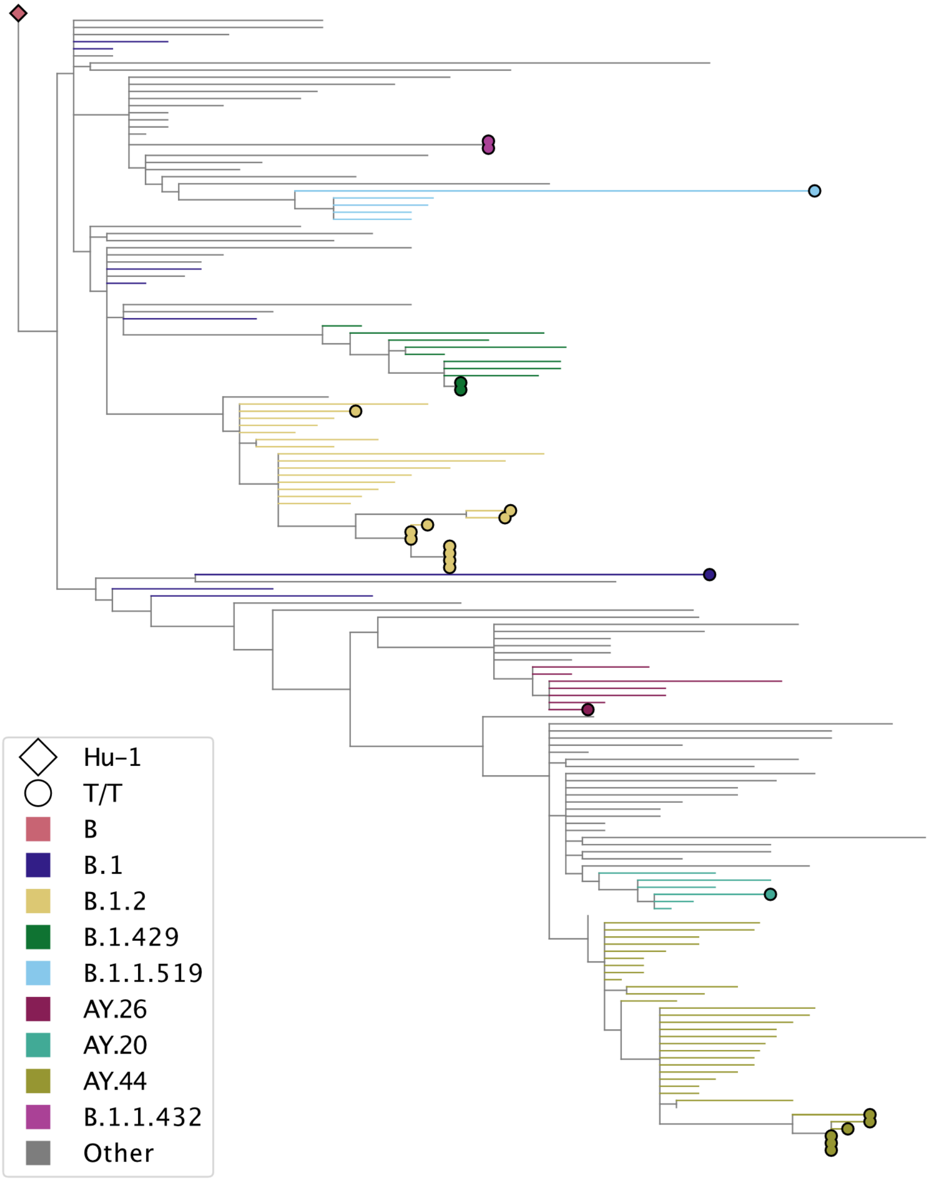

Figure S5. Subtree showing the placement of T/T genomes in the San Diego SEARCH dataset. 24 T/T genomes were placed on a global tree of 3 million genomes (v2022-01-21) using UShER. The node branches are colored by the assigned PANGO lineage. The T/T genomes are highlighted using circles.

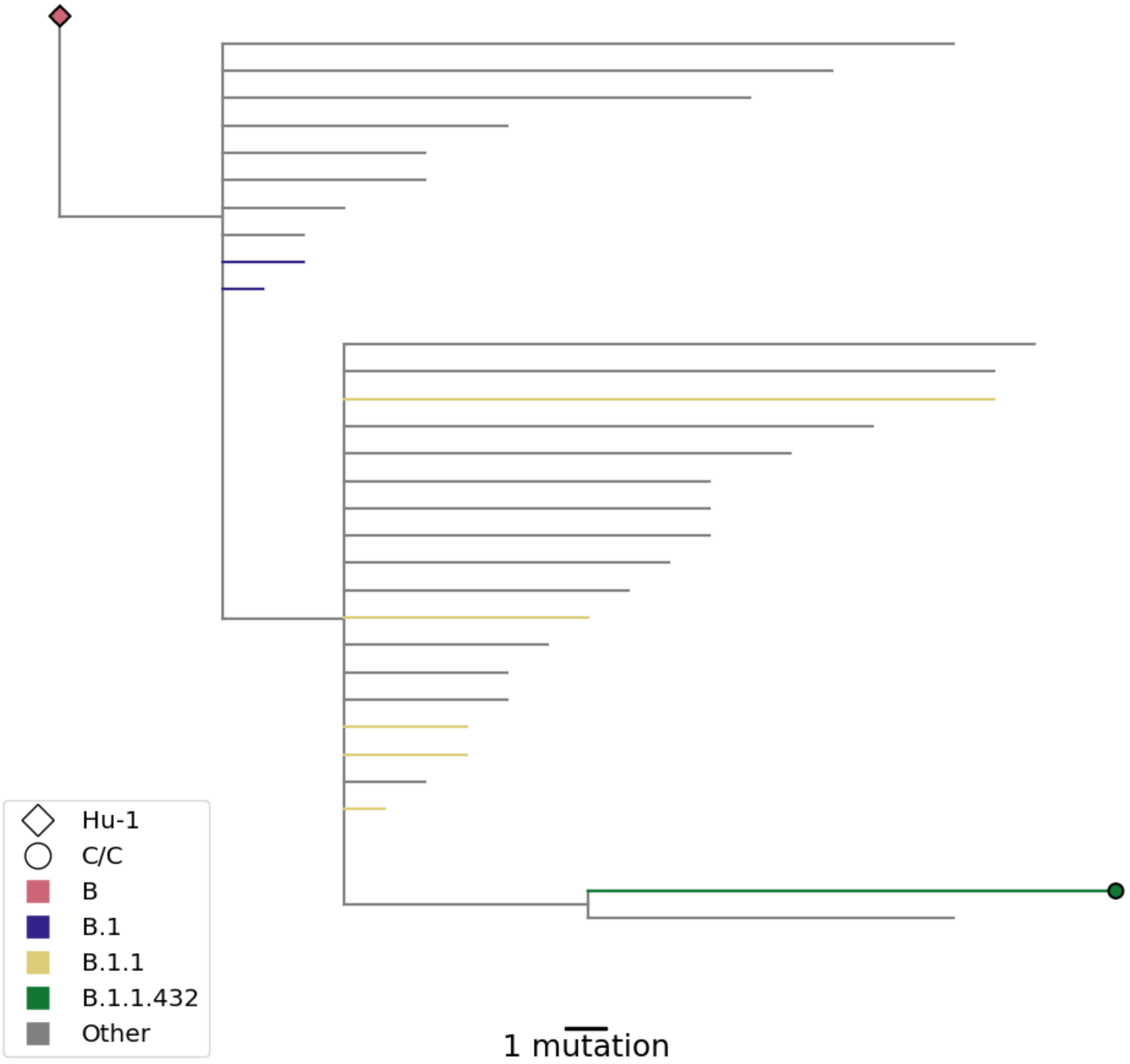

Figure S6. Subtree showing the placement of C/C genome in the San Diego SEARCH dataset. 1 C/C genome was placed on a global tree of 3 million genomes (v2022-01-21) using UShER. The node branches are colored by the assigned PANGO lineage. The C/C genome is highlighted using a circle.

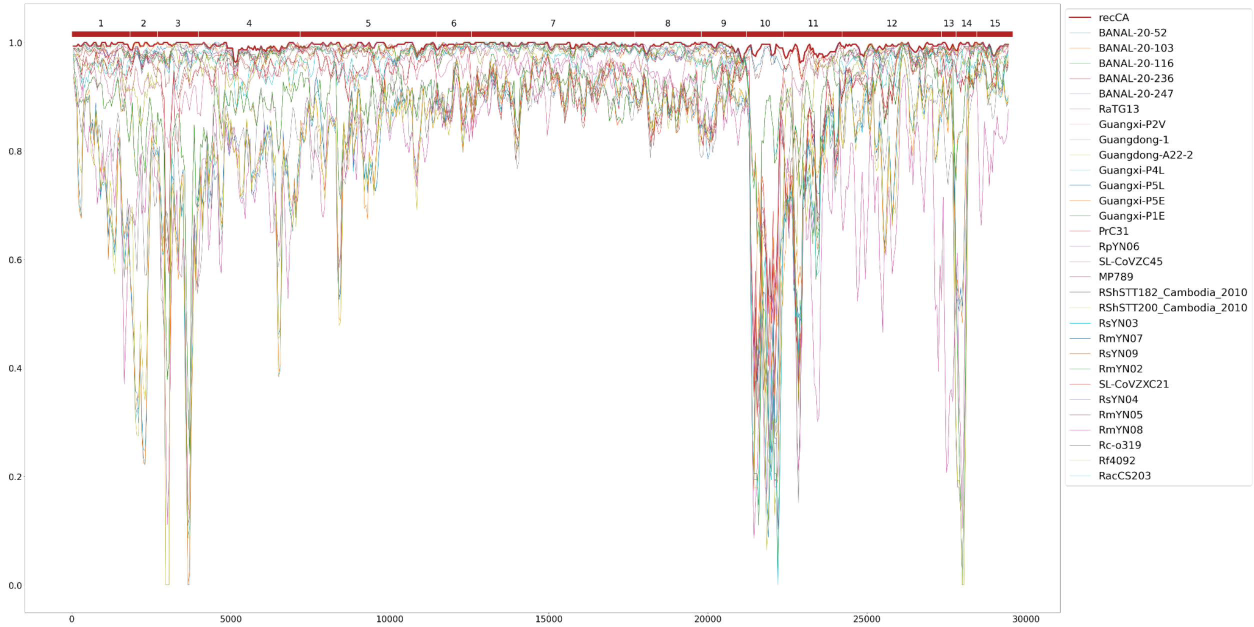

Figure S7. Simplot of closely related sarbecoviruses and recCA with Hu-1 as the reference.

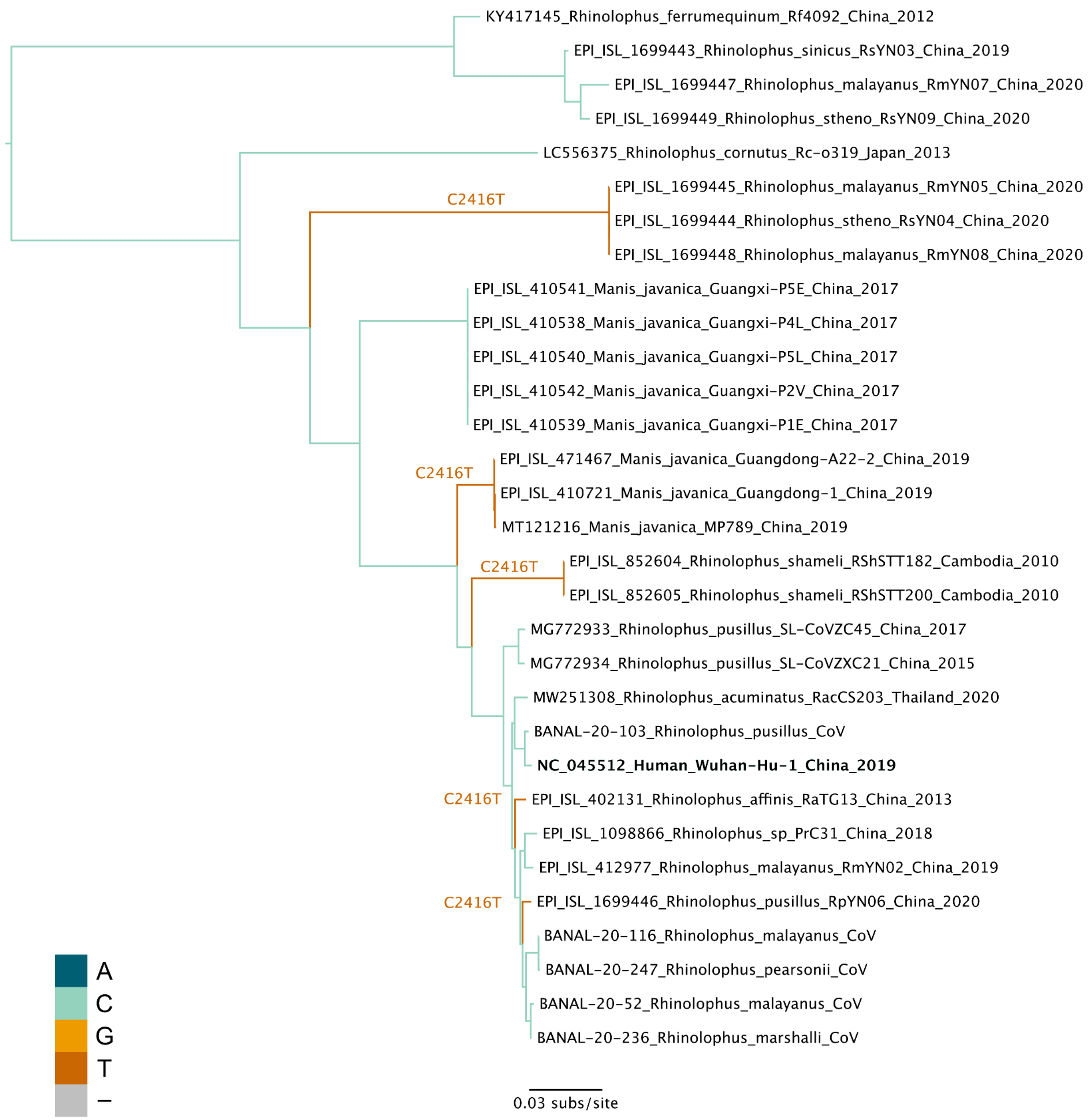
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
C2416T
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
C2416T
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
MT121216_Manis_javanica_MP789_China_2019
C2416T
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
BANAL-20-103_Rhinolophus_pusillus_CoV
NC_045512_Human_Wuhan-Hu-1_China_2019
C2416T
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
C2416T
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
BANAL-20-52_Rhinolophus_malayanus_CoV
BANAL-20-236_Rhinolophus_marshalli_CoV
A
C
G
T
–
0.03 subs/site

Figure S8. Maximum likelihood tree of non-recombinant region 2 with branches colored based on the nucleotide at position 2416. Some substitution labels shifted for clarity.

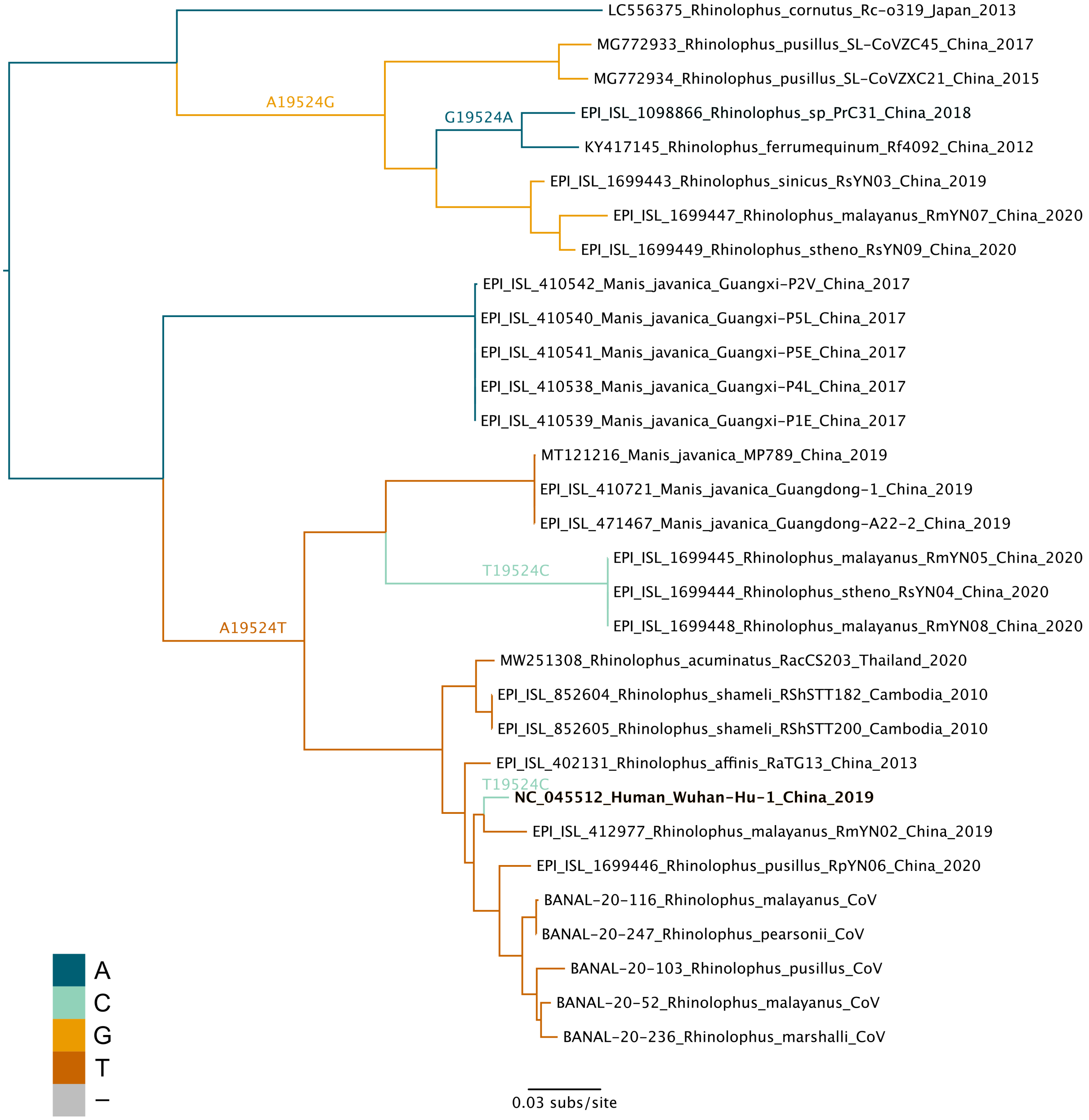
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
A19524G
G19524A
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
MT121216_Manis_javanica_MP789_China_2019
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
T19524C
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
A19524T
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
T19524C
NC_045512_Human_Wuhan-Hu-1_China_2019
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
BANAL-20-103_Rhinolophus_pusillus_CoV
BANAL-20-52_Rhinolophus_malayanus_CoV
BANAL-20-236_Rhinolophus_marshalli_CoV
A
C
G
T
–
0.03 subs/site

Figure S9. Maximum likelihood tree of non-recombinant region 8 with branches colored based on the nucleotide at position 19524. Some substitution labels shifted for clarity.

]

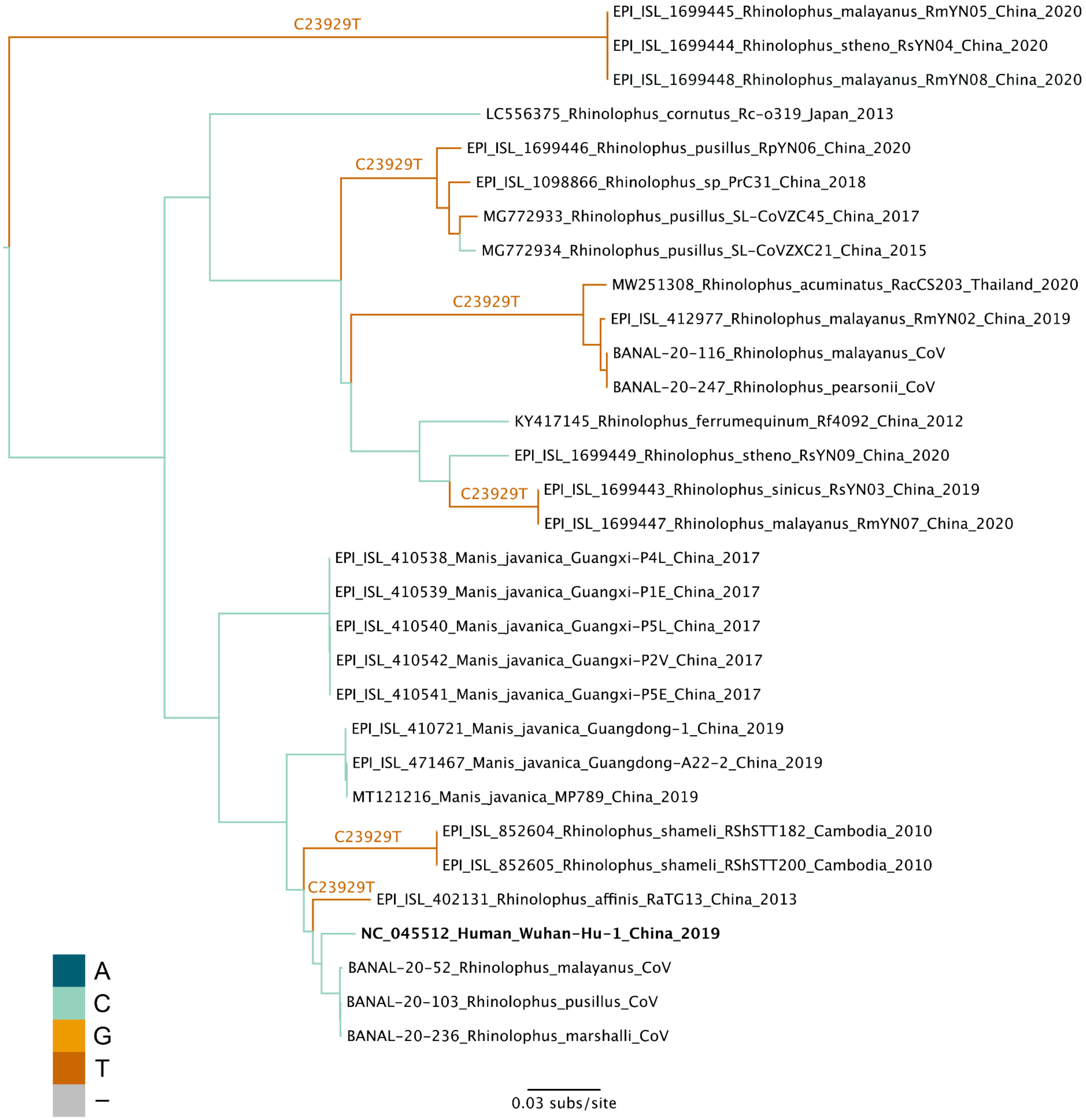
C23929T
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
C23929T
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
C23929T
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
C23929T
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
MT121216_Manis_javanica_MP789_China_2019
C23929T
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
C23929T
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
NC_045512_Human_Wuhan-Hu-1_China_2019
BANAL-20-52_Rhinolophus_malayanus_CoV
BANAL-20-103_Rhinolophus_pusillus_CoV
BANAL-20-236_Rhinolophus_marshalli_CoV
A
C
G
T
–
0.03 subs/site

Figure S10. Maximum likelihood tree of non-recombinant region 11 with branches colored based on the nucleotide at position 23929. Some substitution labels shifted for clarity.

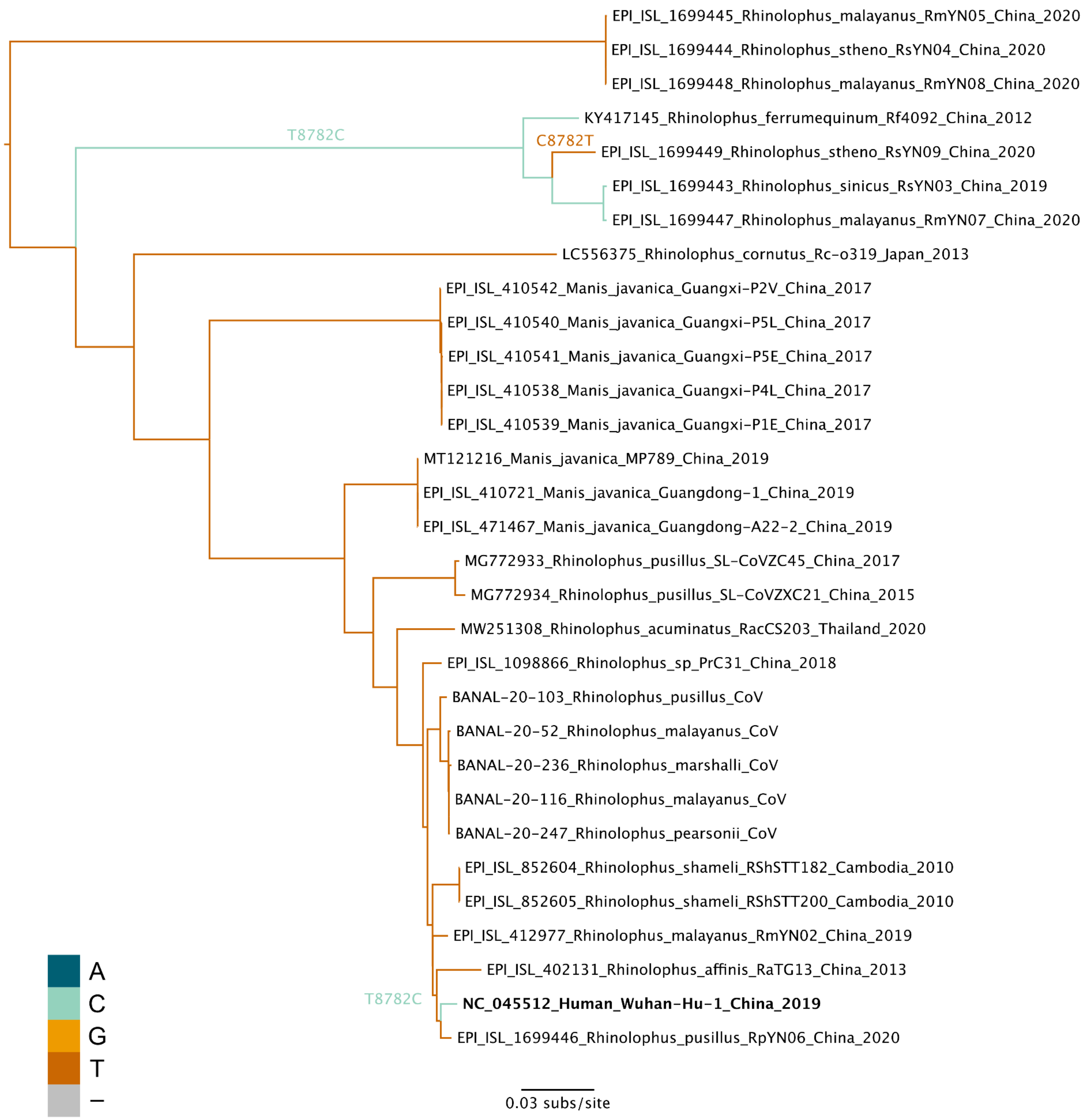
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
T8782C
C8782T
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
MT121216_Manis_javanica_MP789_China_2019
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
BANAL-20-103_Rhinolophus_pusillus_CoV
BANAL-20-52_Rhinolophus_malayanus_CoV
BANAL-20-236_Rhinolophus_marshalli_CoV
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
T8782C
NC_045512_Human_Wuhan-Hu-1_China_2019
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
A
C
G
T
–
0.03 subs/site

Figure S11. Maximum likelihood tree of non-recombinant region 5 with branches colored based on nucleotide at position 8782. Some substitution labels shifted for clarity.

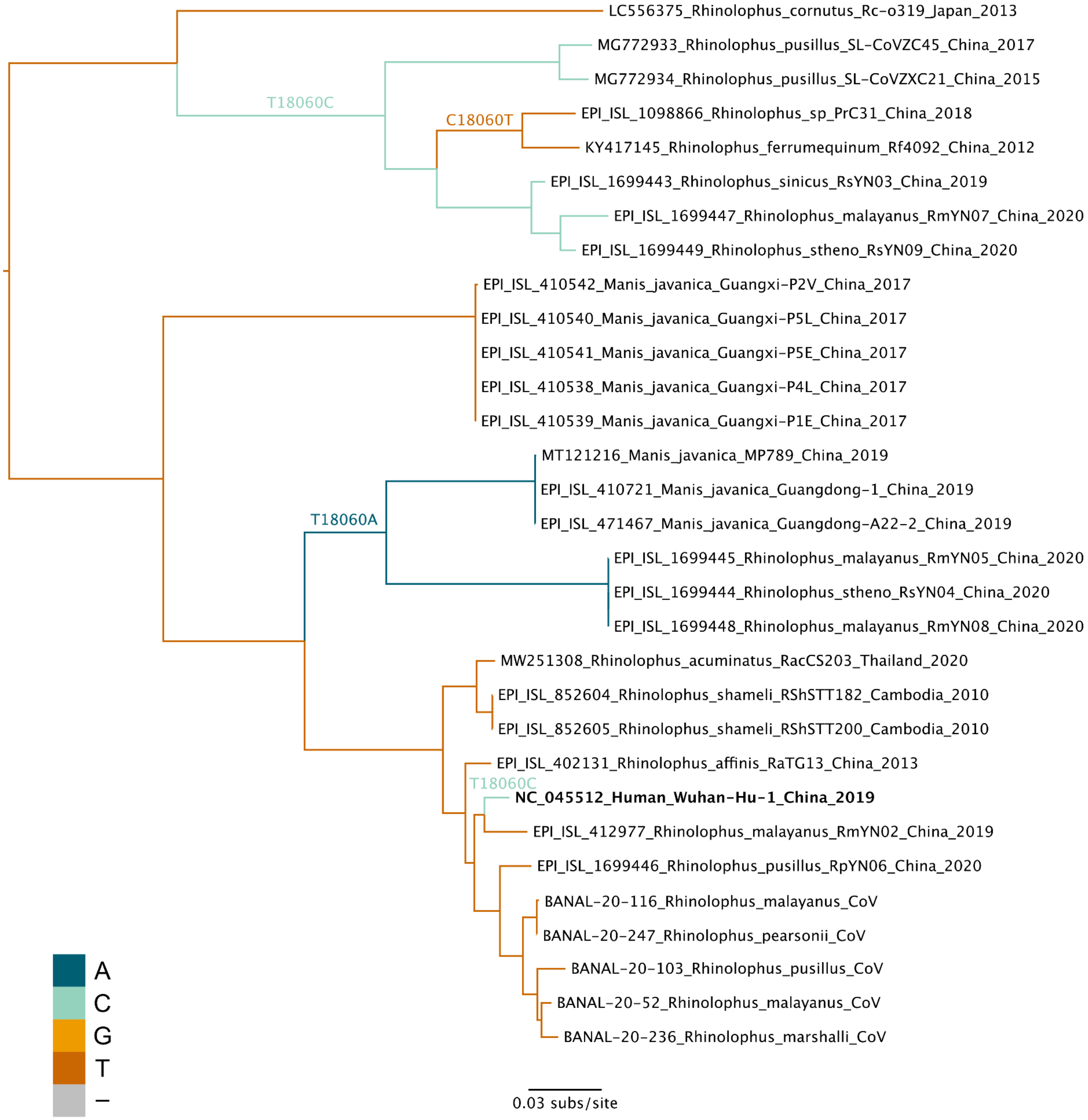
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
T18060C
C18060T
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
MT121216_Manis_javanica_MP789_China_2019
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
T18060A
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
T18060C
NC_045512_Human_Wuhan-Hu-1_China_2019
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
BANAL-20-103_Rhinolophus_pusillus_CoV
BANAL-20-52_Rhinolophus_malayanus_CoV
BANAL-20-236_Rhinolophus_marshalli_CoV
A
C
G
T
–
0.03 subs/site

Figure S12. Maximum likelihood tree of non-recombinant region 8 with branches colored based on the nucleotide at position 18060.

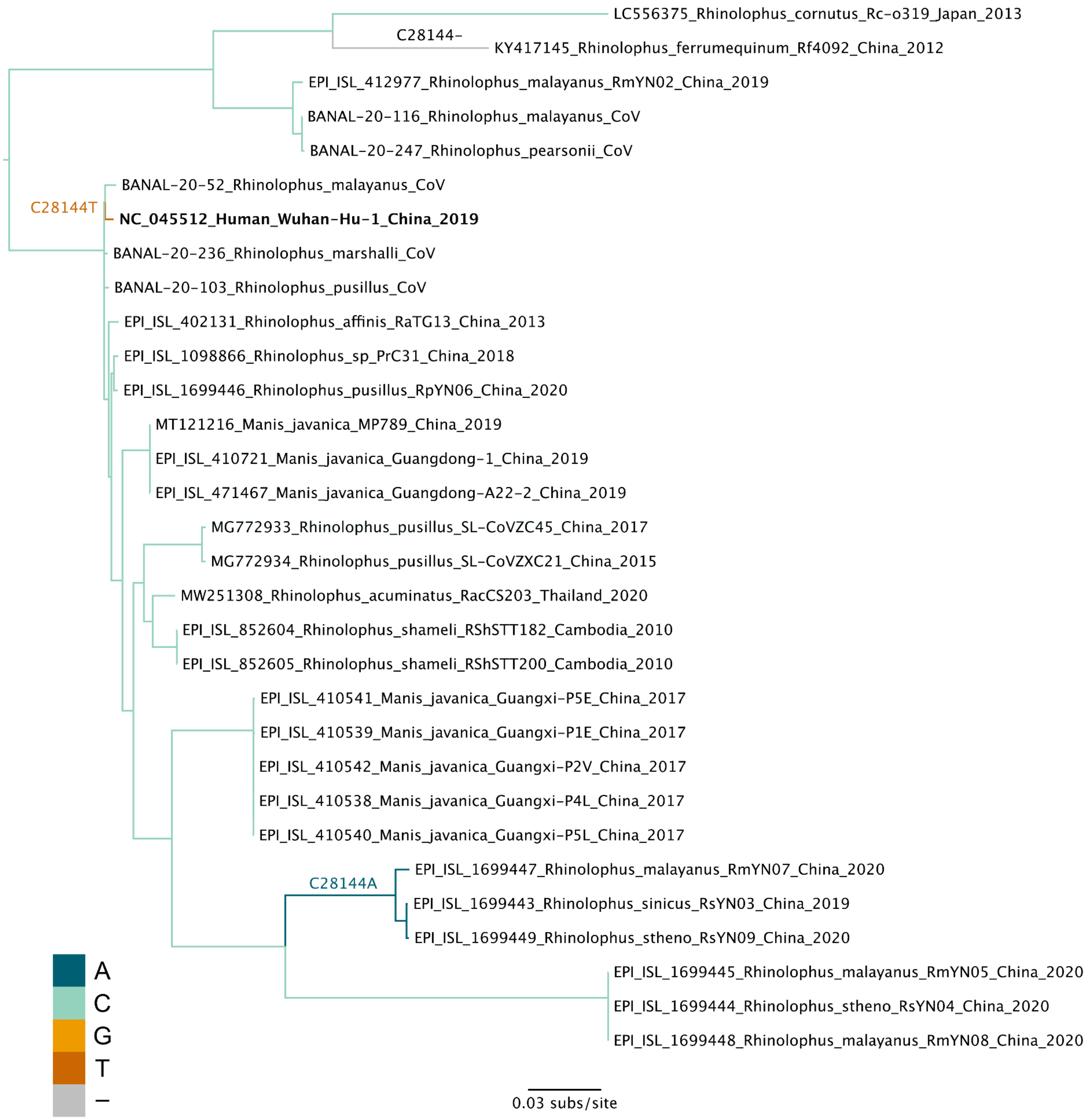
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
C28144-
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
BANAL-20-52_Rhinolophus_malayanus_CoV
C28144T
NC_045512_Human_Wuhan-Hu-1_China_2019
BANAL-20-236_Rhinolophus_marshalli_CoV
BANAL-20-103_Rhinolophus_pusillus_CoV
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
MT121216_Manis_javanica_MP789_China_2019
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
C28144A
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
A
C
G
T
–
0.03 subs/site

Figure S13. Maximum likelihood tree of non-recombinant region 14 with branches colored based on the nucleotide at position 28144. Some substitution labels shifted for clarity.

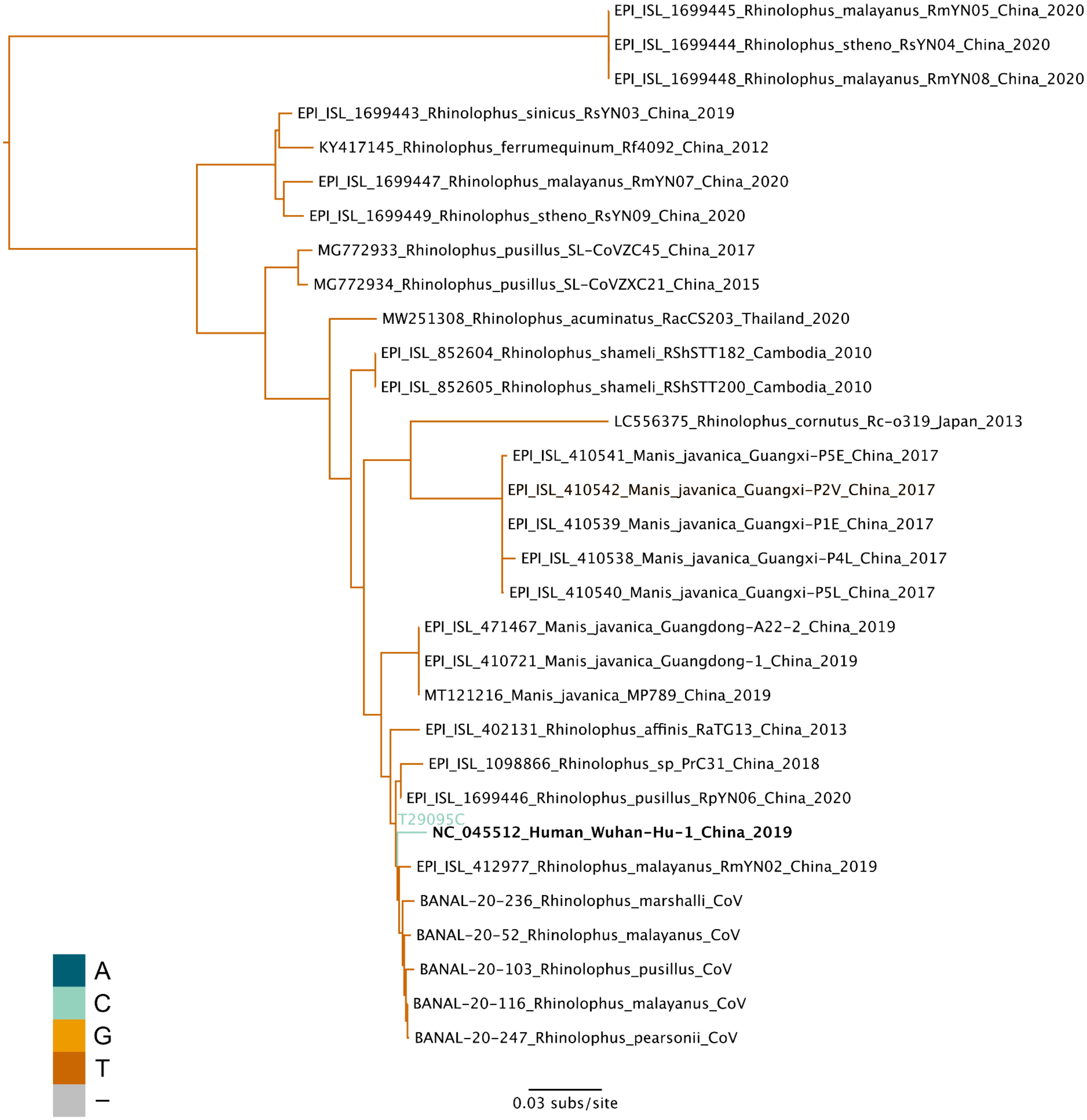
EPI_ISL_1699445_Rhinolophus_malayanus_RmYN05_China_2020
EPI_ISL_1699444_Rhinolophus_stheno_RsYN04_China_2020
EPI_ISL_1699448_Rhinolophus_malayanus_RmYN08_China_2020
EPI_ISL_1699443_Rhinolophus_sinicus_RsYN03_China_2019
KY417145_Rhinolophus_ferrumequinum_Rf4092_China_2012
EPI_ISL_1699447_Rhinolophus_malayanus_RmYN07_China_2020
EPI_ISL_1699449_Rhinolophus_stheno_RsYN09_China_2020
MG772933_Rhinolophus_pusillus_SL-CoVZC45_China_2017
MG772934_Rhinolophus_pusillus_SL-CoVZXC21_China_2015
MW251308_Rhinolophus_acuminatus_RacCS203_Thailand_2020
EPI_ISL_852604_Rhinolophus_shameli_RShSTT182_Cambodia_2010
EPI_ISL_852605_Rhinolophus_shameli_RShSTT200_Cambodia_2010
LC556375_Rhinolophus_cornutus_Rc-o319_Japan_2013
EPI_ISL_410541_Manis_javanica_Guangxi-P5E_China_2017
EPI_ISL_410542_Manis_javanica_Guangxi-P2V_China_2017
EPI_ISL_410539_Manis_javanica_Guangxi-P1E_China_2017
EPI_ISL_410538_Manis_javanica_Guangxi-P4L_China_2017
EPI_ISL_410540_Manis_javanica_Guangxi-P5L_China_2017
EPI_ISL_471467_Manis_javanica_Guangdong-A22-2_China_2019
EPI_ISL_410721_Manis_javanica_Guangdong-1_China_2019
MT121216_Manis_javanica_MP789_China_2019
EPI_ISL_402131_Rhinolophus_affinis_RaTG13_China_2013
EPI_ISL_1098866_Rhinolophus_sp_PrC31_China_2018
EPI_ISL_1699446_Rhinolophus_pusillus_RpYN06_China_2020
T29095C
NC_045512_Human_Wuhan-Hu-1_China_2019
EPI_ISL_412977_Rhinolophus_malayanus_RmYN02_China_2019
BANAL-20-236_Rhinolophus_marshalli_CoV
BANAL-20-52_Rhinolophus_malayanus_CoV
BANAL-20-103_Rhinolophus_pusillus_CoV
BANAL-20-116_Rhinolophus_malayanus_CoV
BANAL-20-247_Rhinolophus_pearsonii_CoV
A
C
G
T
–
0.03 subs/site

Figure S14. Maximum likelihood tree of non-recombinant region 15 with branches colored based on the nucleotide at position 29095.

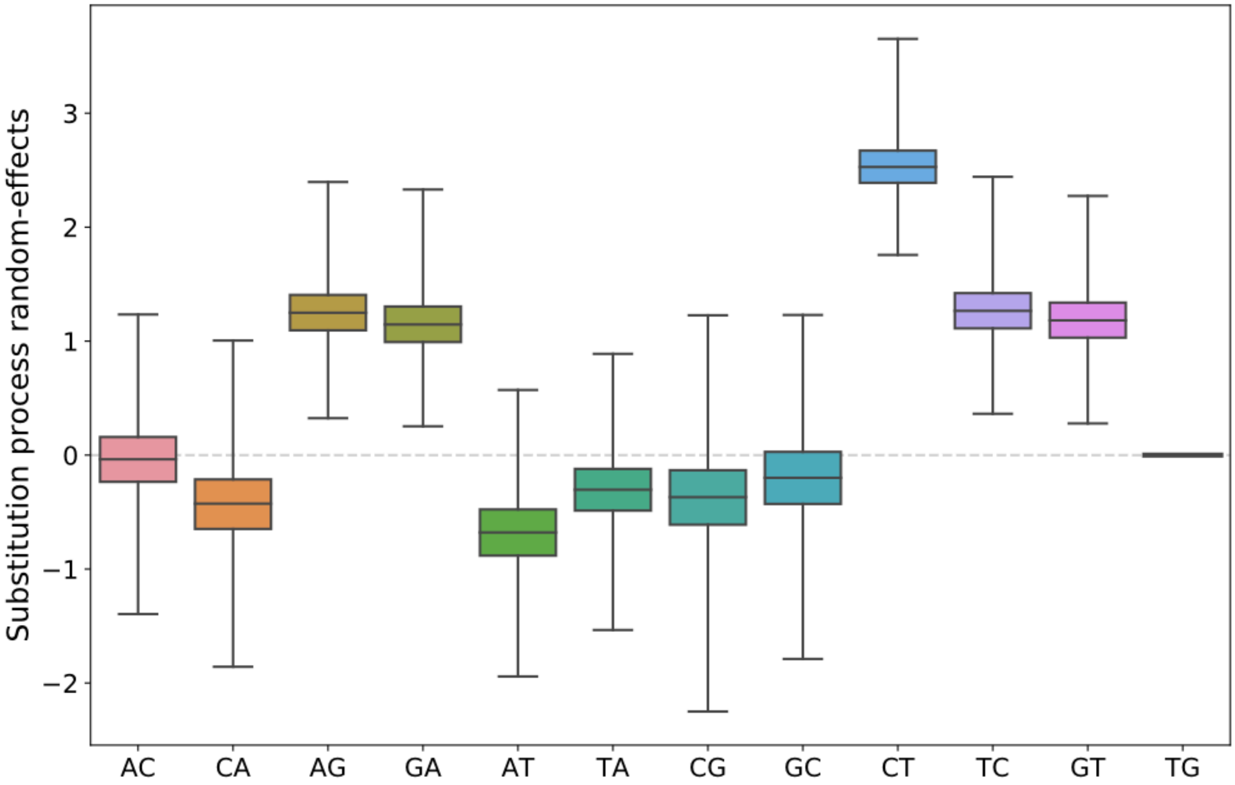

Figure S15. Substitution process random-effects for the unconstrained model. The random effects for transitions were rescaled with κ, and then all random effects were made relative to T-to-G (fixed to 0). The posterior probabilities that $e^{C\text{-}to\text{-}T} > e^{T\text{-}to\text{-}C}$ and $e^{G\text{-}to\text{-}T} > e^{T\text{-}to\text{-}G}$ is 1.00 for both, indicating the C-to-T transition and G-to-T transversion biases were present in every sample in the posterior.

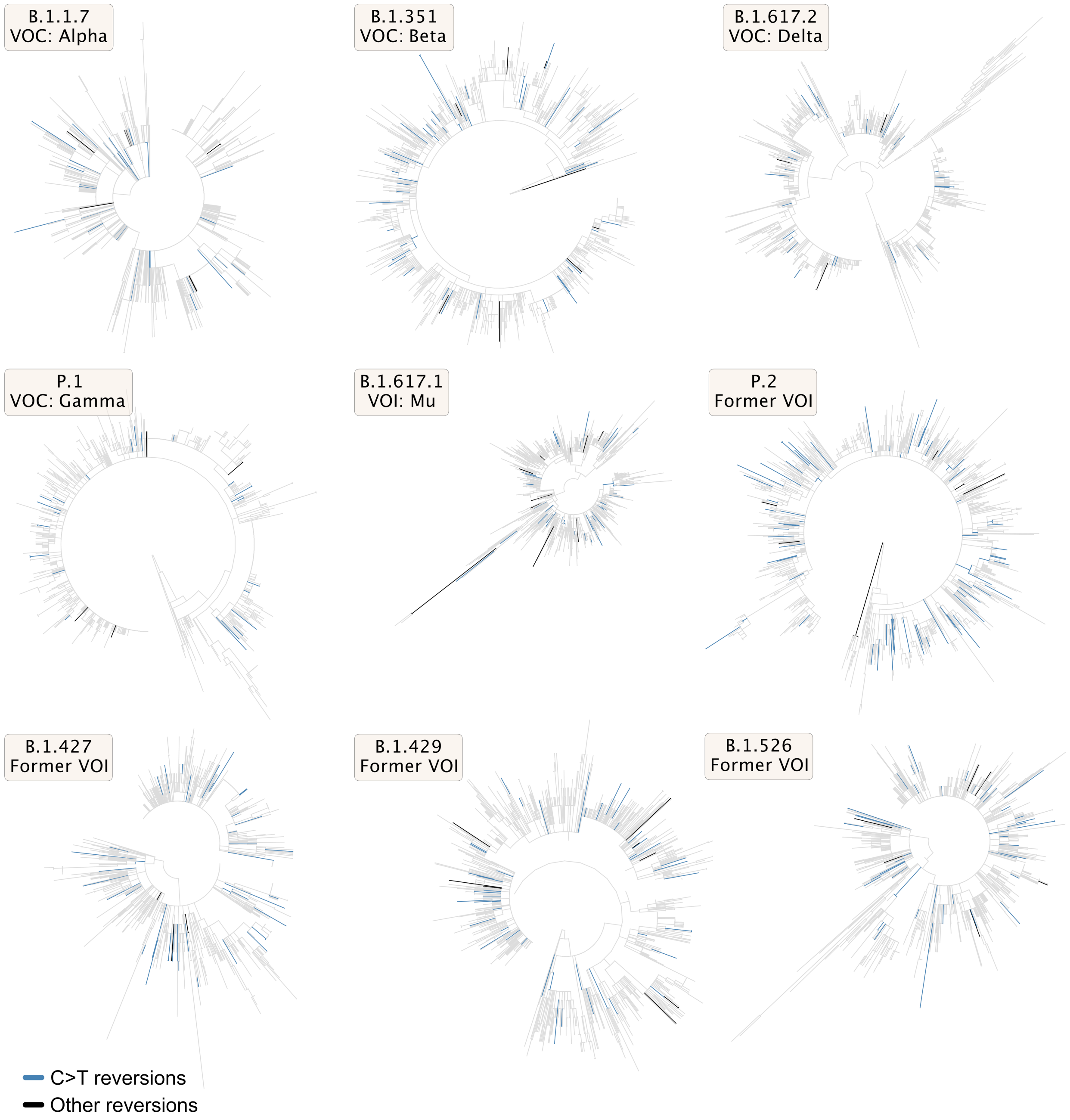

Figure S16. Maximum likelihood phylogenies of variants of concern (VOC) and variants of interest (VOI) with branches containing reversions colored.

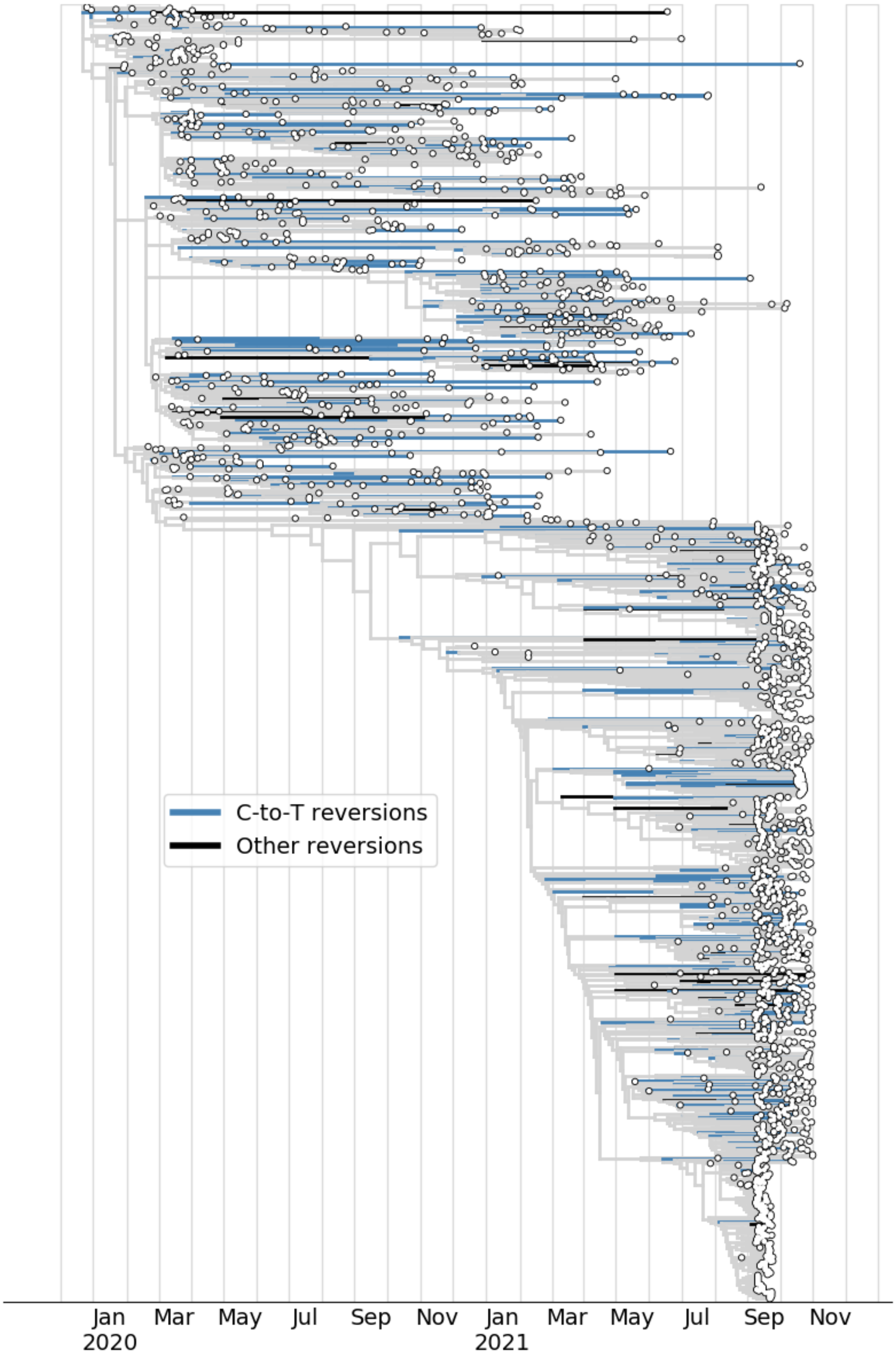

Figure S17. Subsampled global phylogeny showing reversions. Subsampled SARS-CoV-2 time-resolved phylogeny from Nextstrain, with reversions colored blue if a C-to-T reversion and black otherwise.

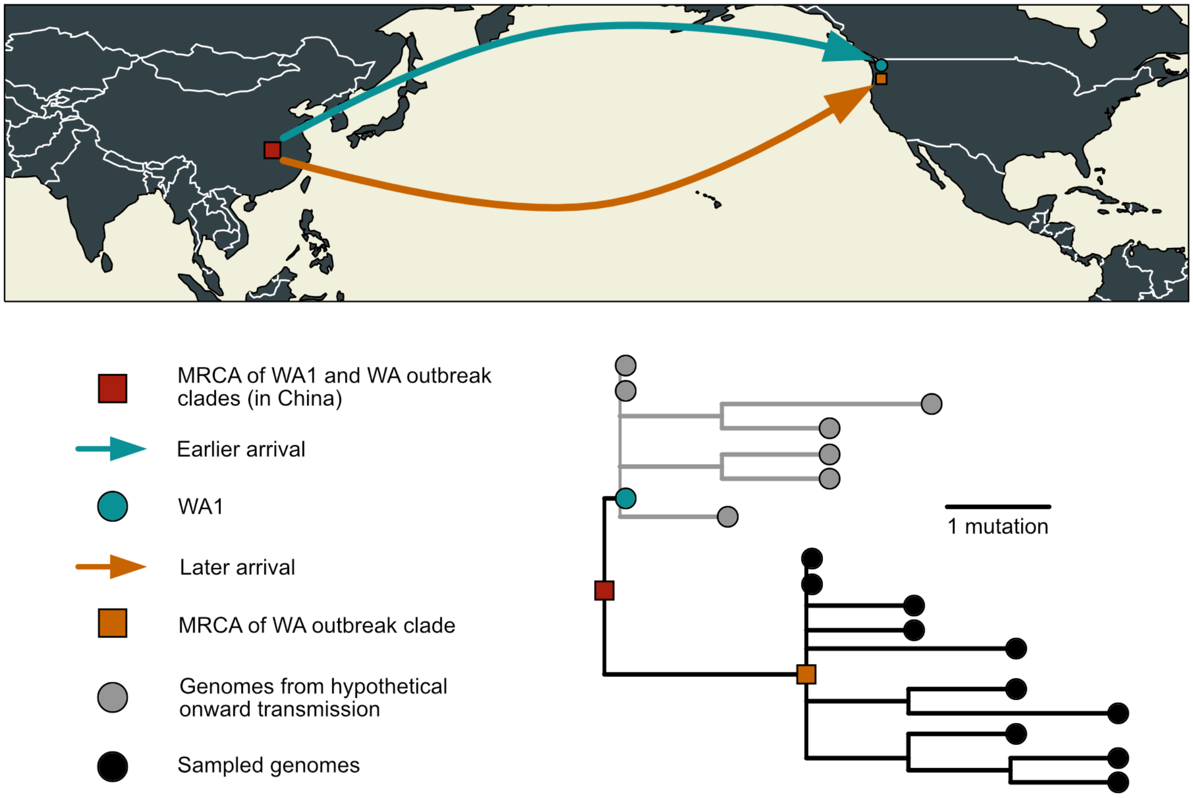
MRCA of WA1 and WA outbreak clades (in China)
Earlier arrival
WA1
Later arrival
MRCA of WA outbreak clade
Genomes from hypothetical onward transmission
Sampled genomes
1 mutation

Figure S18. Early SARS-CoV-2 introductions into Washington state. Similar phylogenetic structure to the origins of SARS-CoV-2 in Wuhan is observed in Washington state, with two separate introductions of SARS-CoV-2 from China differing by two mutations (with no intermediate genomes). Refer to the supplementary text for a discussion comparing the introductions to Washington State with the origins of SARS-CoV-2.

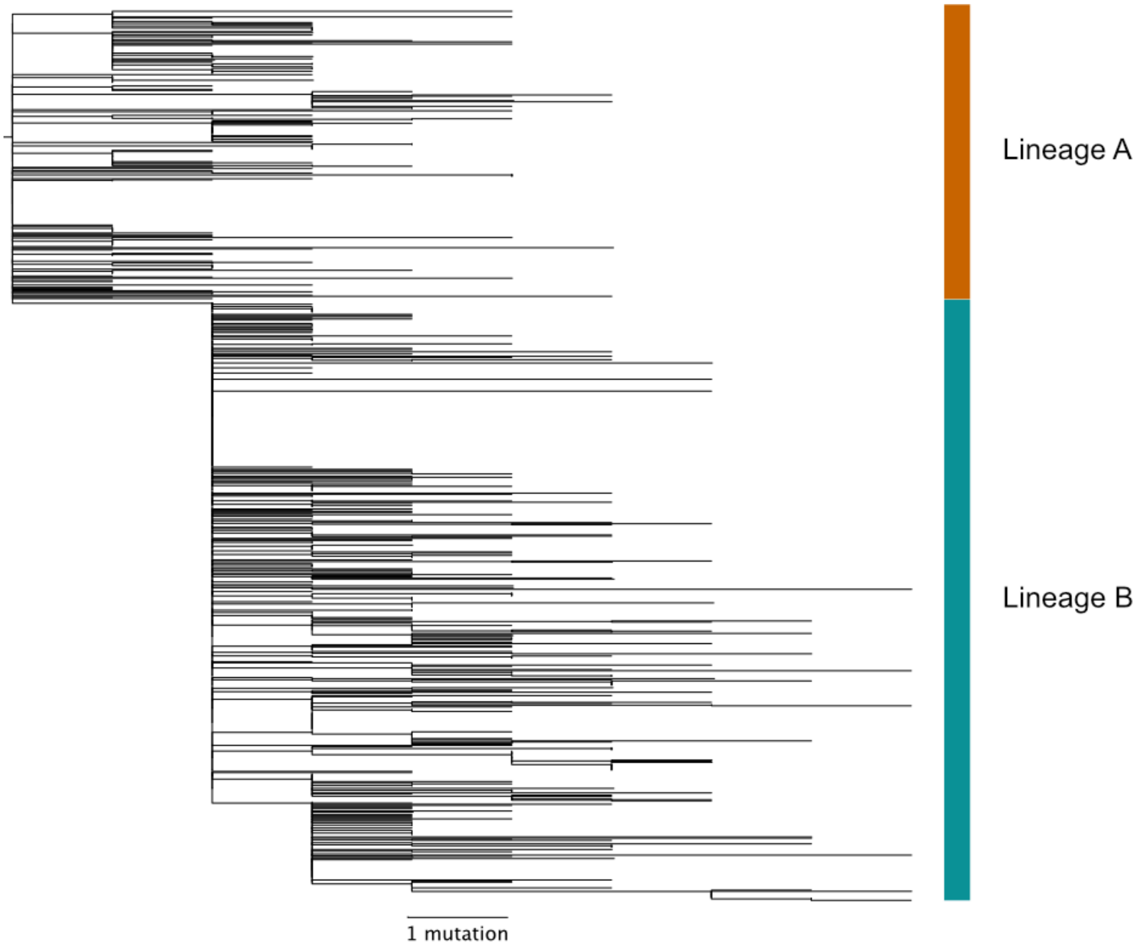

Figure S19. SARS-CoV-2 maximum likelihood tree rooted on lineage A (n=787 taxa, through 14 February 2020).

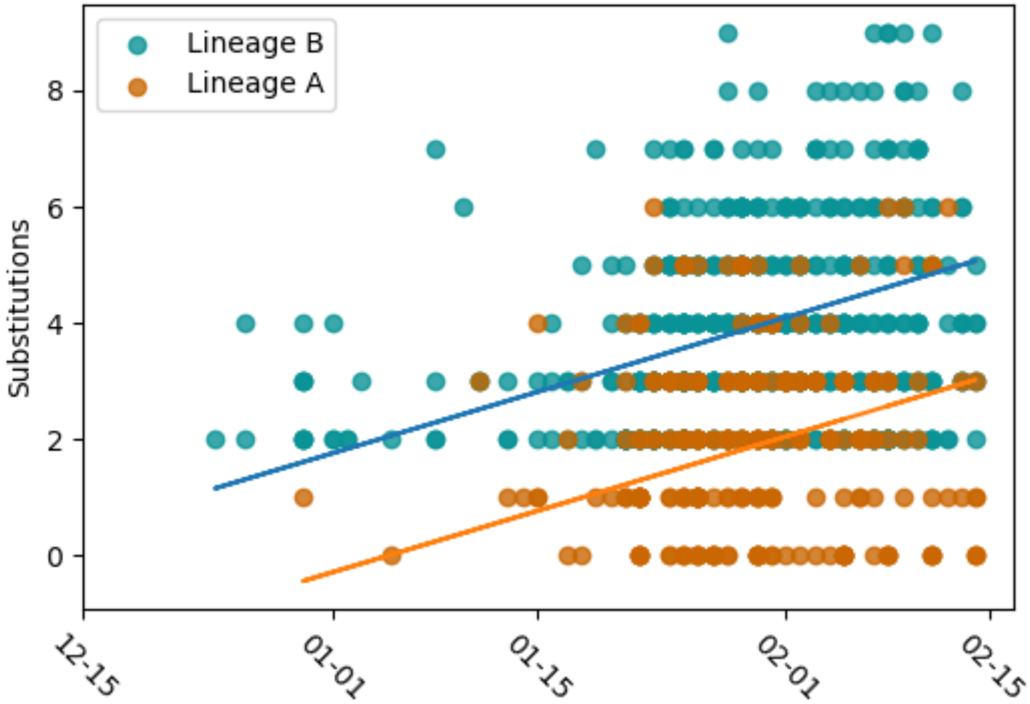

Figure S20. Substitution counts of SARS-CoV-2 genomes through 14 February 2020 from the root of the maximum likelihood tree when rooted on lineage A (Fig. S19). The plotted lines have a slope of 27.51 substitutions/year, are fit to their respective lineages, and are separated by 2.04 substitutions, showcasing the greater divergence of lineage B than lineage A when the tree is rooted on lineage A.

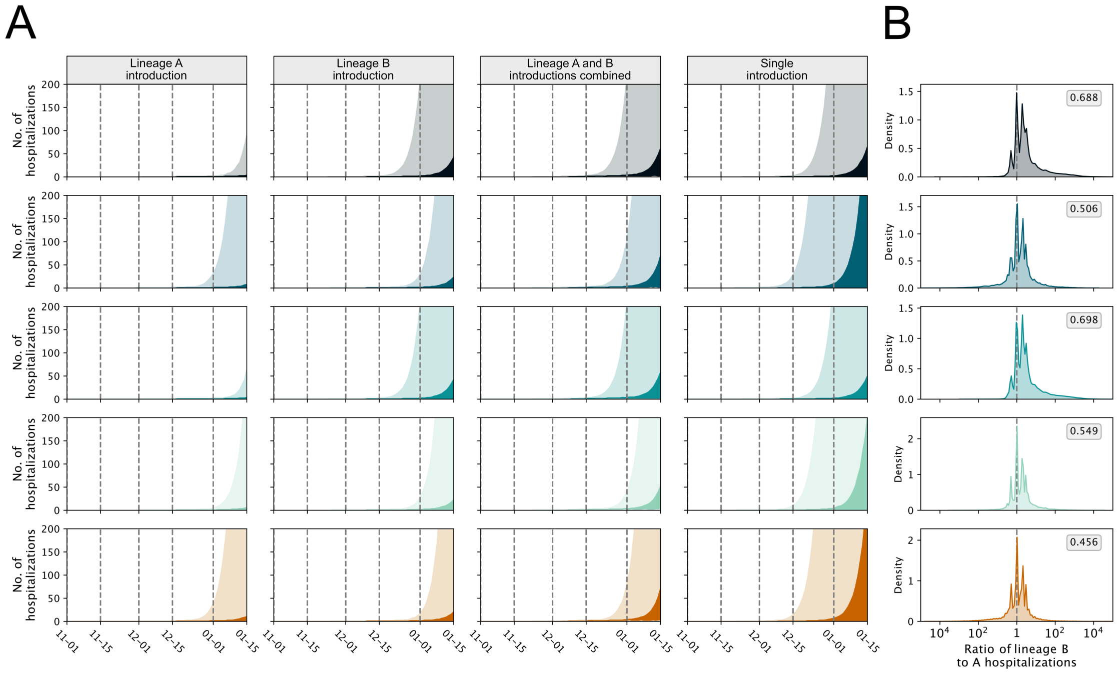

Figure S21. Dynamics of COVID-19 hospitalizations resulting from separate introductions of lineages A and B. (**A**) Estimated number of hospitalizations. The header of each column indicates whether the hospitalizations are caused by lineage A or B in a multi-introduction scenario, the two lineages together in multi-introduction scenario, or by a single introduction. (**B**) The log ratio of lineage B to lineage A hospitalizations on 1 January 2020 in a multi-introduction scenario. The proportion of the posterior with more lineage B infections than lineage A is reported in the grey box in (**B**). Note that the color scheme is the same as in Fig. 6 in the Main Text.

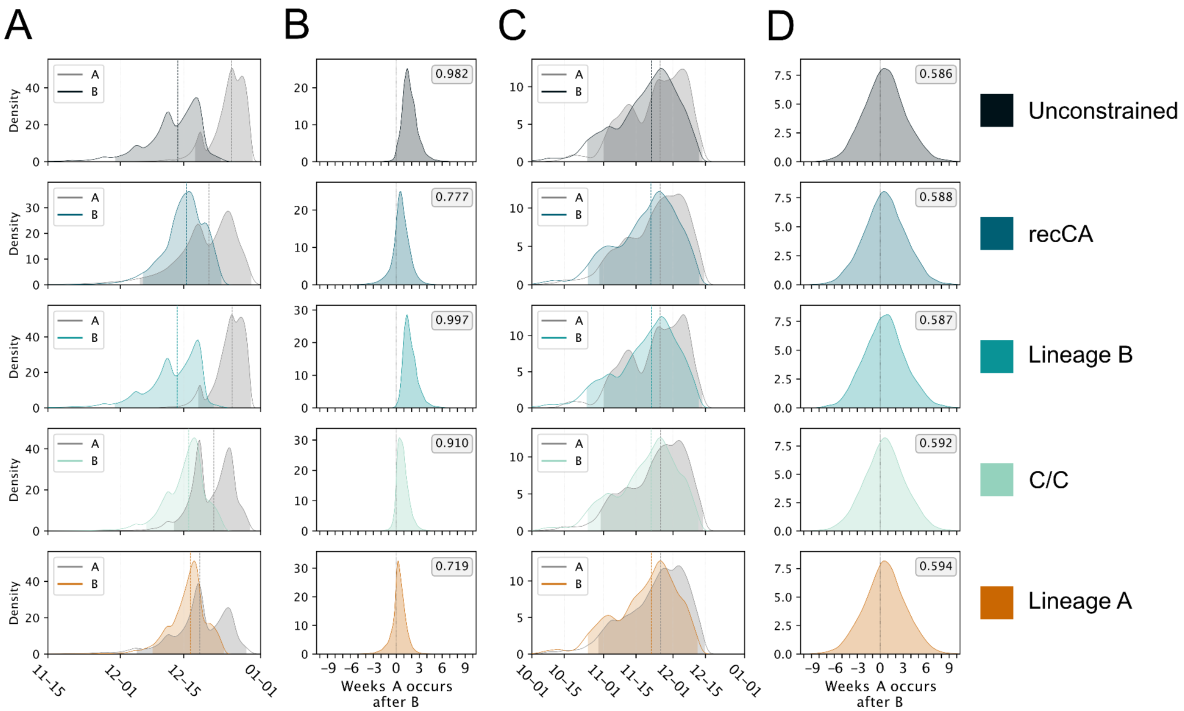

Figure S22. The timing of the MRCA and primary case for lineage A and lineage B with a slower transmission rate. (**A**) The tMRCA for lineages A and B. (**B**) The timing of the primary case for lineages A and B. (**C**) The number of weeks the tMRCA of lineage A occurs after the tMRCA of lineage B. (**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. In (**A, C**), dashed lines indicate the median and shading represents the 95% HPD for each distribution. The proportion of the posterior with lineage A occurring after lineage B in (**B**, **D**) is reported in the grey box. The legend indicates the phylodynamic model used.

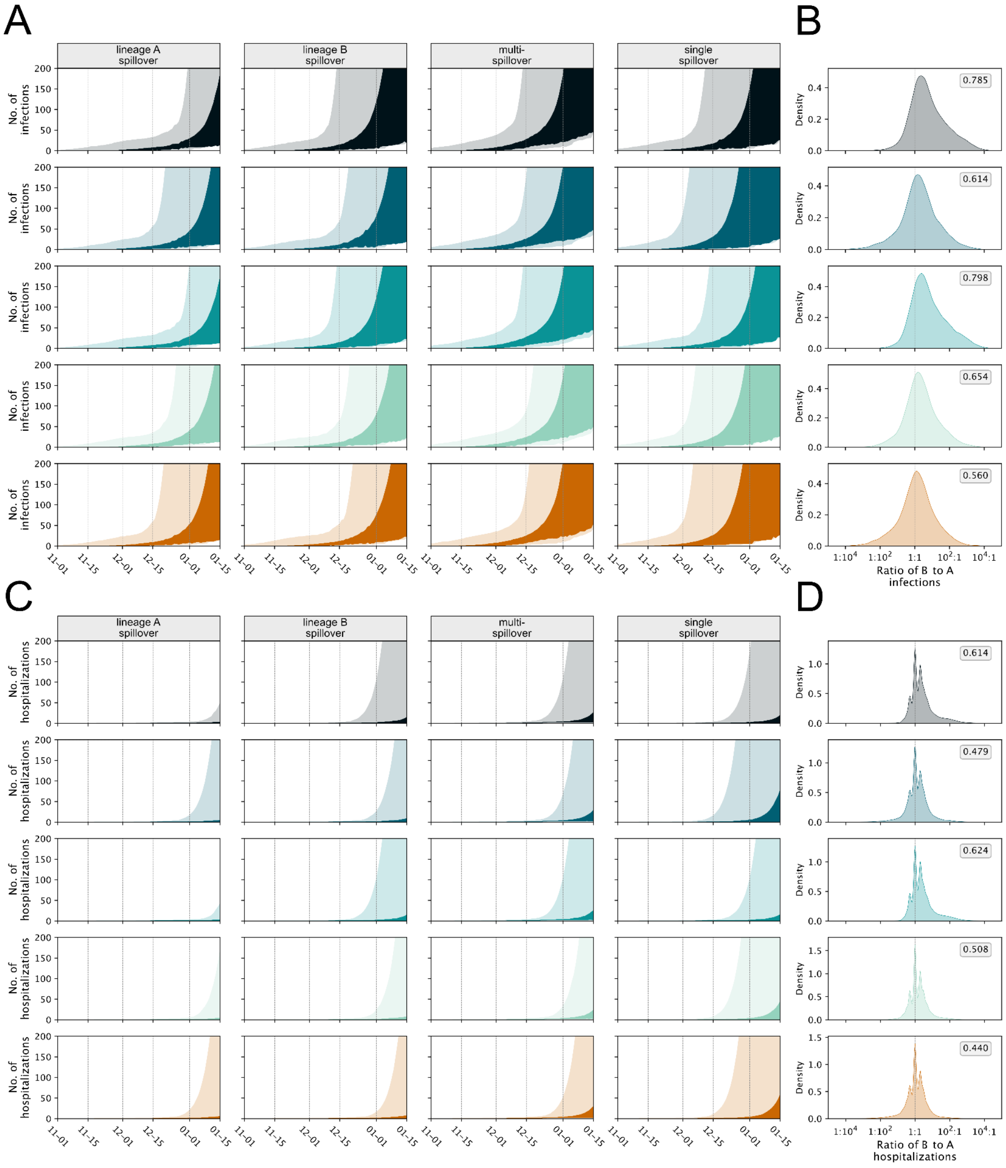

Figure S23. Dynamics of SARS-CoV-2 resulting from separate introductions of lineages A and B and a slower transmission rate. Note that the color scheme is the same as in Fig. S22. (**A**) Estimated number of infections. The header of each column indicates whether the infections are caused by lineage A or B in a multi-introduction scenario, the two lineages together in multi-introduction scenario, or by a single introduction. (**B**) The log ratio of lineage B to lineage A infections on 1 January 2020 in a multi-introduction scenario. (**C**) Estimated number of hospitalizations, with column headers identical to (**A**). (**D**) The log ratio of lineage B to lineage A hospitalizations on 1 January 2020 in a multi-introduction scenario. The proportion of the posterior with more lineage B infections or hospitalizations than lineage A in (**B**, **D**) is reported in the grey box.

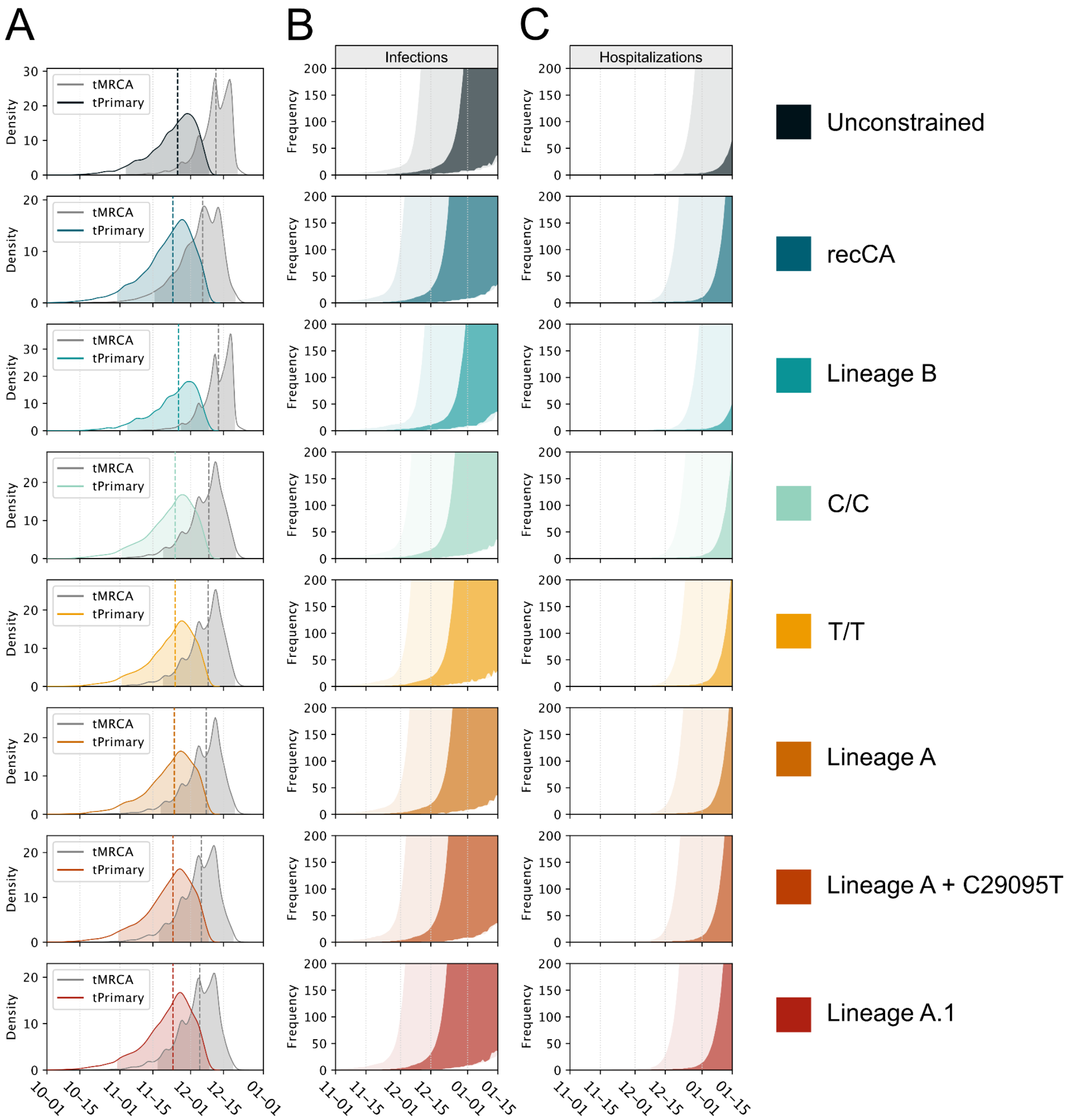

Figure S24. Single-introduction timing of the MRCA and primary case and subsequent epidemic growth. (**A**) Posterior distributions of the timing of the MRCA (tMRCA) and primary case (tPrimary), with dashed lines indicating the median and shading representing the 95% HPD for each distribution. (**B**) Estimated number of infections in late 2019. Darker shading represents 50% HPD; lighter shading represents 95% HPD. (**C**) Estimated number of hospitalizations in late 2019. The legend indicates the phylodynamic model used: the unconstrained model uses just the SARS-CoV-2 genomes; the recCA-constrained model constrains the ancestor of the MRCA of SARS-CoV-2 as the recCA; the remaining models constrain the MRCA of SARS-CoV-2 as a particular sequence (Fig. 4; see methods).

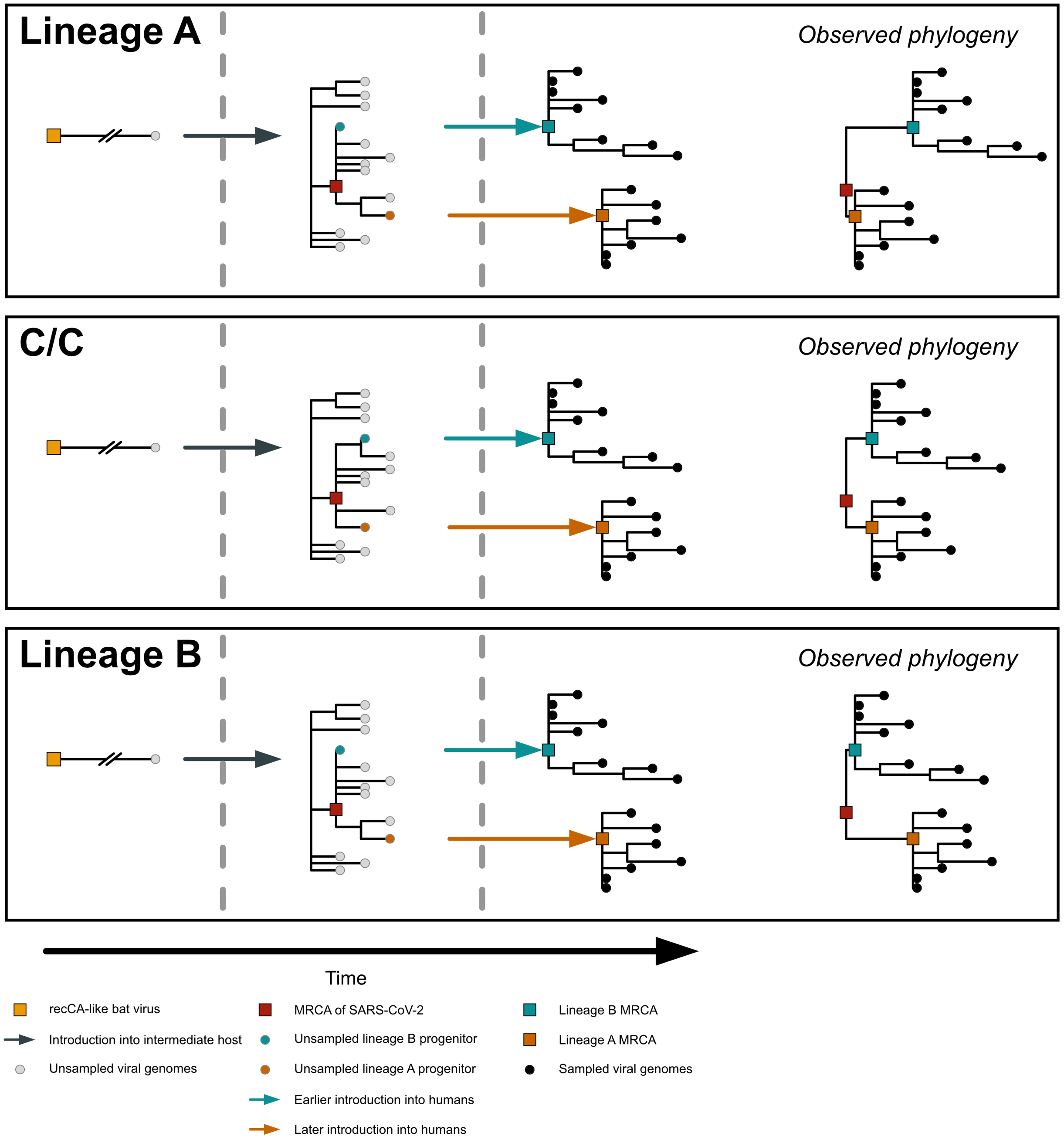

Figure S25. Rooting orientations of observed SARS-CoV-2 phylogenies resulting from different MRCAs and multiple introductions from the intermediate host (see Figure 8 for host depictions). The haplotype of the MRCA (red square in the left-center panel) is depicted in the upper left of each box.