science.sciencemag.org/cgi/content/full/science.abf8003/DC1

Supplementary Materials for

Timing the SARS-CoV-2 index case in Hubei province

Jonathan Pekar, Michael Worobey*, Niema Moshiri,
Konrad Scheffler, Joel O. Wertheim*

*Corresponding author. Email: worobey@arizona.edu (M.W.);
jwertheim@health.ucsd.edu (J.O.W.)

Published 18 March 2021 on *Science* First Release
DOI: 10.1126/science.abf8003

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S18
Tables S1 to S7
Caption for Data S1
References

Other Supplementary Materials for this manuscript includes the following:
(available at science.sciencemag.org/cgi/content/full/science.abf8003/DC1)

MDAR Reproducibility Checklist (.pdf)
Data S1 (.xlsx)

Materials and Methods

Sequence data. We queried the GISAID database (*45*) for SARS-CoV-2 viral genome alignment for sequences from mainland China that were not annotated as travel associated, as of 16 July 2020. We restricted our dataset to genomes that (i) were complete ≥29,000 nt, (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, and (v) had a day-month-year sampling date reported. The first 200 and last 299 nucleotides were removed due to poor evidence of homology. The final alignment comprised 583 taxa. List of GISAID IDs are available in Data S1.

Phylogenetic inference. The phylogenetic history of SARS-COV-2 in China was first inferred in a maximum likelihood framework in IQ-TREE 2 (*46*) using a GTR+F+I model, selected by model testing. Molecular clock analysis was conducted using a Bayesian Markov chain Monte Carlo (MCMC) approach in BEAST v1.10.4 (*20*). For the primary analysis, we employed a GTR+F+I substitution model, a strict molecular clock, a Bayesian skyline coalescent prior. To facilitate convergence, (i) a hard lower bound of $1x10^{-5}$ substitutions/site/year was placed on the clock rate and (ii) we initiated the MCMC using the maximum likelihood phylogeny that had been transformed into a chronogram via TempEst v1.5.3 (*47*). Four independent chains of 500 million generations were run, sampling every 25 thousand, and the first 15% were discarded as burnin. Convergence and mixing was assessed in Tracer v1.7.1 (*48*) and chains were combined in LogCombiner, such that all ESS values were >2070. The resulting posterior distribution comprised over 70,000 sampled trees to facilitate fuller exploration in the rejection sampling (see below). Evidence for shifting root tMRCAs after excluding the earliest sampled SARS-CoV-2 genomes was explored using TreeStat, a part of the BEAST package.

Robustness analysis was conducted using a relaxed clock with an uncorrelated lognormal distribution (ULD) and a Bayesian Skygrid coalescent prior (*49*), in which 4 and 2 independent chains of 500 million generations were run, respectively. We did not explore the effect of highly structured coalescent priors (e.g., constant size, exponential growth), because the Skyline dynamics from China depict a complicated history with stable, exponential, and decreasing effective population sizes (Fig. S15).

Epidemic Simulation. To explore the evolutionary dynamics at play during the beginning of the COVID-19 pandemic, we performed a series of epidemic simulations using FAVITES v1.2.6 (*23*). First, we generated static contact networks in FAVITES under a preferential-attachment model using the Barabási–Albert algorithm (*50*). We used this network algorithm, because its scale-free properties recapitulate infectious disease spread. We chose to simulate a static contact network, because our focus is on the number of people infected at the beginning of the epidemic. For the primary simulations, we selected an intermediate value of 16 contacts per day (mean degree), based on Mossong et al. (*51*), within a contact network comprising 100,000 individuals (nodes).

Across this contact network, we performed a forward simulation SAPHIRE (Susceptible-Ascertained-Presymptomatic-Hospitalized-Not Ascertained (I)-Removed-Exposed) model (*22*) to generate a viral transmission network using GEMF (*52*). 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. Simulated epidemics started with a single seed infection among our 100,000 susceptible individuals. The epidemic was propagated using the parameters determined by Hao et al. (*22*) (see Table S1 for values) for 100 days. Our epidemics had a median doubling time of 4.1 days (95% range: 2.7–6.7 days), corresponding to epidemic growth in Wuhan between 01 January 2020 and 23 January 2020 (*22*).

For the primary analysis, we ran 5000 epidemic simulations. Of these simulations, 29.7% successfully established epidemics, defined as those simulations in which ≥1000 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 1000 people infected over the entire simulation; 70.3% of simulations failed to reach this epidemic threshold after 100 days. For the successfully established simulations, we constructed the coalescent history using the VirusTreeSimulator (*53*) package in FAVITES. For each infected individual in these simulations, a single viral lineage was randomly sampled uniformly across the duration of infection to represent viral genotype sampling. If a simulation failed to converge, it was rerun until it successfully achieved coalescence. The final output generated by FAVITES is the viral time-based phylogeny (part 2b in Fig. S5). 1000 successfully established simulations were then randomly selected for further analysis. All inputs for the FAVITES primary analysis can be found in Table S7 and the JSON input files are available at https://github.com/pekarj/SC2_Index_Case.

Each FAVITES simulation with the SAPHIRE model produced an output documenting when individuals transitioned from one compartment to another throughout the entire simulation. We used these to determine the amount of individuals in a given compartment (e.g., total infections, ascertained infections, unascertained infections, and hospitalized individuals) across each day in the simulation.

Determining Stable Coalescence. Once we had the time-based phylogenies, we labeled the internal nodes using the FAVITES helper script *label_internal_nodes.py*. We then extracted the tMRCA of infected individuals every day across each simulation using TreeSwift 1.1.14 (*54*). This tMRCA was calculated for each day of the 100 days or until 10,000 individuals had been infected, whichever came first. We chose not to explore dynamics after 10,000 infections due to a slowing in exponential growth arising from the saturation of the contact network.

We defined the stable coalescence as the coalescence that does not shift forward in time by more than one day, even as new individuals become infected and previously infected individuals recover. 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 coalescence after the 100 day simulation or once 10,000 total individuals have been infected.

Sensitivity Analysis—Faster rate of infection. The same methods for epidemic simulations were performed to evaluate the dynamics of a more rapidly spreading virus. We used the aforementioned parameters, except the edge-based rates of transmission were increased (Table S2). We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each. Increasing the infectiousness coefficient to 0.55 per day produced a median epidemic doubling time of 3.1 days (95% range: 1.6-5.1).

Sensitivity Analysis—Slower rate of infection. The same methods for epidemic simulations were performed to evaluate the dynamics of a more slowly spreading virus. We used the aforementioned parameters, except the edge-based rates of transmission were increased (Table S2). We produced 7000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each. Decreasing the infectiousness coefficient to 0.3 per day produced a median epidemic doubling time of 5.3 days (95% range: 3.6-7.5) (Table S2).

Sensitivity Analysis—Higher average degree. The same methods for epidemic simulations were performed to evaluate the dynamics of a more densely connected network. We used the aforementioned parameters, except the average degree of the contact network was 26, and we rescaled the per-contact transmissibility in the SAPHIRE parameters by the mean degree such that the overall doubling time was maintained (Table S2). We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each.

Sensitivity Analysis—Lower average degree. The same methods for epidemic simulations were performed to evaluate the dynamics of a less densely connected network. We used the aforementioned parameters, except the average degree of the contact network was 10, and we rescaled the per-contact transmissibility in the SAPHIRE parameters by the mean degree such that the overall doubling time was maintained (Table S2). We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each.

Sensitivity Analysis—Higher probability of hospitalization. The same methods for epidemic simulations were performed to evaluate dynamics involving higher probabilities of hospitalization. We used the aforementioned parameters, except with either 2-times or 10-times the probability of hospitalization (D_h^{-1} in the SAPHIRE model) in the main analysis. We produced 5000 simulations to generate at least 1000 established replicates with at least 1000 infected individuals each.

Sensitivity Analysis—Small village. The same methods for epidemic simulations in FAVITES were performed to evaluate the dynamics of a less densely connected network, except without rescaling the transmission rate with the average degree. We performed two sets of 5000 simulations, decreasing the mean degree to 8 (50% reduction) and 4 (75% reduction). We kept constant all other parameters from the primary analysis.

Sensitivity Analysis—Random contact network. The same methods for primary epidemic simulations in FAVITES were performed to evaluate the extinction dynamics across a random network. We used the aforementioned parameters from the primary analysis, except transmission was propagated across an Erdős–Rényi network (mean degree = 16), to produce 5000 simulations.

Sensitivity Analysis—Impact of subsampling on stable coalescence. For each of the FAVITES epidemic simulations from the primary analysis, we randomly subsampled 50% of the lineages circulating at the time of stable coalescence and calculated the new time of coalescence. This process was repeated 100 times for each epidemic simulation, and we used the median time of coalescence. This approach was also applied using 5% and 25% of the circulating lineages at the stable coalescence.

Two-Phase Epidemics in FAVITES. We also used FAVITES to explore the evolutionary dynamics of a two-phase epidemic, in which the index case was infected with a less fit (i.e., less transmissible) variant that eventually went extinct (termed Phase 1 of the epidemic); however, before this less-fit variant went extinct, it gave rise to an adapted variant that became dominant in Wuhan and spread through China (termed Phase 2 of the epidemic). To simulate the first phase of this epidemic, we applied the same methods for epidemics simulations and parameters as in the primary analysis, except with a 50% lower transmission rate (*b* in Table S1). We generated 5000 simulations in FAVITES and randomly selected 2000 epidemics that do not establish (i.e., went extinct) for further analysis. These simulations (Phase 1) were combined via rejection sampling (see below) with the primary established simulations (Phase 2), which would persist until reaching a stable coalescence.

Sensitivity Analysis—Varied transmission rates for Phase 1. The same methods for the two-phase simulations were performed to evaluate dynamics involving higher and lower rates of transmission for Phase 1. We used the aforementioned parameters, except with transmission rates of 1/6, 2/6, 4/6, and 5/6 the rate of transmission during Phase 2. We produced 5000 Phase 1 simulations, each of these lower transmission rates.

Sensitivity Analysis—Higher ascertainment rate for Phase 1. The same methods for the epidemic simulations were performed to evaluate dynamics involving higher rates of ascertainment (Table S1; Table S6). Increased ascertained is a proxy for increased virulence, as ascertained individuals are more likely to transmit the virus and become hospitalized. We used the aforementioned parameters, except with 2-times, 4-times, or 6-times the rate of ascertainment for Phase 1, relative to Phase 2, to produce 5000 simulations.

Sensitivity Analysis—Higher two-phase probabilities of hospitalization. The same methods for the epidemic simulations were performed to evaluate dynamics involving higher probabilities of hospitalization for either just the Phase 1 epidemic, or the combined Phase 1 and Phase 2 epidemic. We used the aforementioned parameters, except with either 2-times or 10-times the probability of hospitalization for Phase 1 and Phase 2 simulations and 4-times the rate of ascertainment as the primary Phase 1 simulations.

Combining FAVITES and BEAST via Rejection Sampling. Our aim is to obtain a posterior distribution for the date X of the index case in Hubei province, conditioned on both the available sequencing data D_s and the date of the first reported COVID-19 case D_c. We do this in a Bayesian framework by marginalizing over the date Y of the first ascertained or unascertained COVID-19 case (see below) and the date Z of the tMRCA as follows:

$$P(X|D_s, D_c) = \int_Z P(X|Z, D_s, D_c)P(Z|D_s, D_c)\, dZ \qquad \text{Equation (1)}$$

We assume that the sequencing data are informative only for the tMRCA, i.e. given Z, X does not depend on D_s: $P(X|Z, D_s, D_c) = P(X|Z, D_c)$. We also assume that the first reported COVID-19 case data are not informative for the tMRCA: $P(Z|D_s, D_c) = P(Z|D_s)$. This gives:

$$P(X|D_s, D_c) = \int_Z P(X|Z, D_c)P(Z|D_s)\, dZ \qquad \text{Equation (2)}$$

We further note that $P(X|Z, D_c) = \int_Y P(X, Y|Z, D_c)\, dY$, where we model $P(X, Y|Z, D_c)$ as proportional to $C(Y)P(X, Y|Z)$, where $C(Y)$ is a “consistency function” with a value of 1 when Y is consistent with D_c and 0 otherwise. This approach allows us to sample from the posterior distribution of Equation 2. The BEAST analysis provides values of Z sampled from the distribution $P(Z|D_s)$. For each sampled value of Z, we sample corresponding values of X and Y from the distribution $P(X, Y|Z, D_c)$ using the FAVITES simulation (providing samples from the distribution $P(X, Y|Z)$ in conjunction with a simple rejection sampling-based strategy: sample values from $P(X, Y|Z)$ until a sample is obtained for which $C(Y) = 1$. The resulting set of sample values for X then follow the posterior distribution $P(X|D_s, D_c)$.

We assign Y as the first instance of an ascertained (SAPHIRE stage: I) or unascertained (SAPHIRE stage: A) (I/A: Fig. S6) case and D_c as 17 November 2019, where $C(Y) = 1$ when Y precedes D_c. However, we note that the first ascertained/unascertained case will often be the index case themselves, unless a secondary or tertiary case progresses faster through the course of infection. 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, coalescence can happen any time after the index case is first infected, and there is no requirement for coalescence to occur after the first ascertained and unascertained individuals.

Rejection Sampling—two-phase epidemics. Rejection sampling for the two-phase epidemics involved randomly sampling (i) a Phase 1 and (ii) a Phase 2 epidemic simulations and combining them at (iii) a time point during Phase 1, representing the moment at which a more fit variant arose. This time when the adaptive variant evolved was chosen uniformly across the Phase 1 epidemic, weighted by the number of circulating lineages (Fig. S9-2B).

The first ascertained or unascertained case can come from either Phase 1 or Phase 2—whichever occurs first—and is used for the consistency function $C(Y)$. The remainder of the rejection sampling method remains the same for the two-phase epidemic.

Sensitivity Analysis—ascertained and unascertained cases. The primary rejection sampling analysis was conditioned on an individual having an ascertained (SAPHIRE stage: I) or unascertained (SAPHIRE stage: A) infection prior to the date of the first reported case of COVID-19 on 17 November 2019. However, we also performed sensitivity analyses whereby the minimum date for the earliest case must be an ascertained case (Fig. S6). We performed rejection sampling using the first ascertained or unascertained case (I/A; Fig. S6) or solely the first ascertained case (I-only; Fig. S6).

Sensitivity Analysis—date of first COVID-19 case. We also explored the sensitivity of the rejection sampling approach to the date before which an ascertained or unascertained case must have existed. We performed rejection sampling using 17 November 2019 or 01 December 2019 as the minimum date for the earliest case.

Sensitivity Analysis—relaxed clock and coalescent priors. We explored the sensitivity of the rejection sampling approach to different molecular clocks in the BEAST inference. We performed separate rejection sampling analyses using tMRCAs inferred under a relaxed ULD clock and a Skygrid coalescent prior.

Sensitivity Analysis—shifting the tMRCA. We explored the sensitivity of the rejection sampling approach to the stability of the timing of coalescence in the BEAST analysis (Fig. S3A). The primary analysis used the inferred tMRCA of all viruses sampled on or after 01 January 2020, at which point the tMRCA had stabilized. Sensitivity analysis was conducted using the tMRCA of all sampled viruses (i.e., the root tMRCA). This tMRCA is within 1 calendar day in 78.5% of sampled trees (Fig. S3B) and represented by the same node in 74.6% of sampled trees (Fig. S3C).

Supplementary Text

Relaxed molecular clock. Strict molecular clocks can produce overly-precise tMRCA estimates when evolution occurred under a relaxed clock. Relaxing the molecular clock for SARS-CoV-2 in China using an uncorrelated lognormal distribution of rates (ULD) did produce a slightly wider tMRCA estimate, with a mean of December 6th (95% HPD: November 9th–December 22nd) (Fig. S1A) with a comparable rate of 8.45×10^{-4} (95% HPD: 7.05×10^{-4}–9.89×10^{-4}). However, the standard deviation of the ULD was 0.0009 (95% HPD: 0.00005–0.00013), suggesting strong (strict) clock-like evolution across the SARS-CoV-2 phylogeny in China. Furthermore, 91.7% of the posterior tMRCA estimate post-dated the earliest reported case on 17 November 2019.

Phylogenetic rooting. The position of the root in the Bayesian phylodynamic inference under a strict molecular clock was ambiguous. In 63.5% of the posterior trees, the root fell within the basal polytomy of 79 identical genomes, exemplified by the Wuhan-Hu-1 reference genome (GenBank Accession MN908947). This position aligns with the assumption of the NextStrain algorithm (https://nextstrain.org/ncov/global). In 21.8% of posterior trees, the root fell on a branch which led to a single virus: the earliest sampled genome IPBCAMS-WH-01 (GenBank Accession MT019529) in 15.0% of trees and another early genome WH01 (GenBank Accession MT291826) in 2.4% of trees. This hypothetical root orientation corresponds with the plurality of estimates from Pipes et al. (*21*)*,* though their approach did not make use of the molecular clock. Notably, we find very little support for a root position on branches corresponding to the T28114C (3.9%) or C8782T (1.4%) mutations, as previously suggested by Zhang et al. (*7*).

Rooting configurations were similar when using a relaxed ULD clock. The root fell among viruses with the Wuhan-Hu-1 reference genome sequence in 61.7% of the posterior sample, on the branch leading to IPBCAMS-WH-01 in 18.1% of samples, and on the branch leading to WH01 in 3.7% of samples. Again, there was little support for a root position on branches corresponding to the T28114C (2.5%) or C8782T (0.4%) mutations.

Rooting configurations were also similar when using a Skygrid coalescent prior. The root fell among viruses with the Wuhan-Hu-1 reference genome sequence in 68.2% of the posterior sample, on the branch leading to IPBCAMS-WH-01 in 9.7% of samples, and on the branch leading to WH01 in 1.9% of samples. Again, there was little support for a root position on branches corresponding to the T28114C (3.3%) or C8782T (0.7%) mutations.

Time to stable coalescence in simulations. Across the simulated epidemics, a stable coalescence (i.e., non-shifting tMRCA from that point in time until the end of the simulated epidemic) was established after a median of 16 days after the index case was first infected; 95% of simulations reached a stable coalescence by day 53 post-index case infection (Fig. S16A), indicating that the 100-day simulations were sufficient to determine time to stable coalescence. Importantly, the time at which stable coalescence was achieved represents when the tMRCA of the currently circulating lineages is within one day of the tMRCA of the lineages remaining at the end of the simulation. This point in time is when basal lineages ceased to be lost in the coalescent tree, not the tMRCA of the remaining viral lineages (Fig. S16B). At the time a stable coalescence was reached, a median of 16 people had been infected; 95% of simulations reached this stable coalescence after 1194 people had become infected. Recall that the empirical phylogenetic analysis from the Chinese epidemic appears to have reached a stable coalescence by 01 January 2020, when around 1000 people are believed to have been infected (*22*).

Epidemics extinction. Most epidemics that go extinct, that is they do not establish themselves in the population, do so after a single infection by day 9, the average duration of a single infection in the SAPHIRE model (Fig. 4C, Table S2). As the probability of epidemic establishment decreases, the maximum number of total infections during the simulated epidemic tends to increase. This pattern holds in networks with variable transmission rates and network connectivity (Table S2). As the probability of viral transmission decreases, more infections can occur before an epidemic goes extinct.

Subsampling lineages at the stable coalescent. To match the real-world scenario of unseen lineages when calculating a tMRCA, we subsampled the circulating lineages in the FAVITES simulations at the stable coalescent to determine any potential effects on the timing of the index case for a one-phase epidemic. The subsampling was done at 5%, 25%, and 50%. The subsampled lineages were used to determine a new time of stable coalescence for the FAVITES simulations, and then rejection sampling was performed as in the primary analysis. The point of coalescence shifted slightly forward in time as sampling completeness decreased (Fig. S17A), but the timing of the index case was robust and closely matched the primary analysis without subsampling (Fig. S17B).

Varying the rate of transmission for Phase 1 of a two-phase epidemic. Although a less-fit variant would have been less transmissible than the variant that spread through China in late-2019, it is unclear how much less transmissible it would have been. Our two-phase analysis (Fig. 3F to K) assumed a relative fitness of one-half. To explore the robustness of inferred date of the index case to this assumption, we explored additional two-phase simulations varying the relative fitness of this Phase 1 variant by sixths: from 1/6 through 5/6 the transmission rate of the Phase 2 variant (Table S1). The extinction probability varied substantially across these fitness variants (Table S6). When the relative fitness in Phase 1 was 5/6 of that in Phase 2, 77.3% of simulated epidemics ended in extinction; however, when the relative fitness was 1/6 of that in Phase 2, 4999 out of 5000 simulated epidemics went extinct. Remarkably, the inferred timing of the index case was robust across all these relative fitness values (Fig. S18). This consistency likely arises because our index case inference relied on only Phase 1 epidemics that went extinct, and the time from index case to extinction was similar across relative fitness values (Table S6).

Varying the ascertainment rates for Phase 1 of the epidemic. It is possible that the less-fit variant during Phase 1 of a two-phase epidemic was also more virulent. Therefore, we explored this scenario by increasing the ascertainment rate of the Phase 1 variant to 2- and 4-times that of the ascertainment rate of the Phase 2 variant (Table S1). The results of the index case timing were similar across all these ascertainment rates (Table S6; Fig. S18). The proportion of simulations that went extinct decreased slightly as the ascertainment rate increased, because ascertained cases have a higher transmission rate than unascertained cases.

Number of hospitalized individuals through time. There has also been speculation that hospitals in Hubei province were inundated with COVID-19 patients in October and November 2019 (*38*). However, our primary simulation analysis conditioning on the first case being identified by 17 November 2019 suggests that an increase in hospitalizations due to COVID-19 would not have been notable until mid- to late-December 2019. In fact, we do not observe enough total infected people, let alone ascertained of hospitalized people, in October and November 2019 to overwhelm hospitals. We note that our estimate of 0 to 36 hospitalizations as of 01 January 2020 in the primary analysis is less than the 42 hospitalizations that had been previously reported (*6*); however, this value is contained within the 95% HPD of many robustness analyses (Fig. S13).

To explore if increased probability of hospitalization early in the epidemic (due to different clinical practices or a more virulent virus), we explored increasing by two- and ten-fold the probability of hospitalization. The simulations also produced the 42 hospitalizations on 01 January 2020 in the 95% HPD (Fig. S13). However, even these extreme deviations from observed parameters failed to produce high rates of hospitalization prior to December 2019, even in the 99% HPD (Fig. S13).

Similar results were seen in two-phase epidemic simulations, including a scenario in which the Phase 1 variant had both higher rates of ascertainment and hospitalization. To test the most extreme case, we simulated a Phase 1 epidemic with 4-times the ascertainment rate as Phase 2 and a ten-fold higher probability of hospitalization for Phases 1 and 2, but there are consistently fewer than 10 hospitalized individuals prior to 01 December 2019 (Fig. S14).

It is important to acknowledge that our SAPHIRE model is based on parameters from Hao et al. during the period between 01 January and 22 January 2020, which occurred after SARS-CoV-2 was discovered and while there was still limited understanding of COVID-19 dynamics (*22*). Therefore, even though we tested a two-phase epidemic, higher rates of ascertainment, and higher probabilities of hospitalization, we have little confidence in the precision of our estimates regarding the number of hospitalized patients in late-2019. Nonetheless, even across the various robustness analyses we explored, we never observed a substantial number of hospitalized patients in November 2019, even in the 99% extreme of our estimates (Fig. S13, S14).

Table S1. Simulation parameters, with all parameters except for *b* based on Hao et al. (*22*)**.** The *b* value listed is for the primary analysis.

Parameter	**Meaning**	**Value**
b	Transmission rate of ascertained cases	0.385
r	Ascertainment rate	0.15
α	Ratio of transmission of unascertained to ascertained cases	0.55
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)	21
D_h	Isolation/hospitalization period (days)	30

Table S2. Parameterization and success/failure of one-phase compartmental epidemic simulations in primary analysis and robustness analyses.

Parameter	Primary analysis	Robustness analyses						
		Less infectious	More infectious	Densely connected	Sparsely connected	Small village 1	Small village 2	Erdos-Renyi
Mean No. of edges per node (degree)	16	16	16	10	26	8	4	16
Infectiousness (*b*)	0.385	0.3	0.55	0.385	0.385	0.385	0.385	0.385
Doubling time of established simulations[a]	4.1 (2.7-6.7)	5.3 (3.6-7.5)	3.1 (1.6-5.1)	4.2 (2.7-7.0)	4.2 (2.7-7.1)	8.5 (7.8-16.4)	11.4 (6.5-32.3)	13.8 (10.7-18.5)
Percentage failed	70.3	80.5	53.6	69.4	68.3	94.5	99.6	83.7
Days to extinction of failed simulations[a]	7.9 (1.5-40.4)	8.3 (1.6-45.7)	7.2 (1.4-34.2)	8.0 (1.5-42.7)	7.6 (1.4-40.1)	8.6 (1.7-51.3)	7.9 (1.8-39.9)	10.4 (1.7-75.0)
No. of infections of failed simulations[a]	1 (1-9)	1 (1-11)	1 (1-7)	1 (1-9)	1 (1-9)	1 (1-15)	1 (1-7)	2 (1-33)
Maximum No. of infections in failed simulations	44	69	20	29	33	110	565	101

[a]Median and 95% range in parentheses

Table S3. Estimates for existence of lineages in one- and two-phase epidemics.

Type of Epidemic				One-phase	Two-phase	Two-phase	Two-phase	Two-phase
Analysis Framework				Index case infected at tMRCA (%)	Index case infected at tMRCA (%)	First adapted case infected at tMRCA (%)	Phase 1 infections exist at tMRCA (%)	Earliest case due is Phase 1 infection (%)
Earliest Case	Clock	Ascertained	Modification					
Nov 17	Strict	I/A	None	1.1	2.2	12	17	61
Nov 17	Strict	I-only	None	1	4	20	33	70
Dec 01	Strict	I/A	None	7.9	8	25	22	63
Dec 01	Strict	I-only	None	6.9	9.1	30	39	74

Table S4. Number infected estimates for one-phase primary (first row) and robustness analyses, with the number infected reported on 17 November, 01 December, and 15 December 2019.

Analysis framework				**Median # infected (95% HPD)**		
Earliest Case	**Clock**	**Ascertained**	**Modification**	**17 Nov 2019**	**01 Dec 2019**	**15 Dec 2019**
Nov 17	Strict	I/A	None	4 (1-13)	9 (2-26)	21 (3-281)
Nov 17	Strict	I-only	None	7 (1-15)	13 (4-31)	24 (8-271)
Nov 17	Strict	I/A	No coalescent shift	4 (1-15)	10 (2-53)	23 (3-1052)
Nov 17	Strict	I/A	Faster transmission	3 (1-12)	7 (2-24)	17 (3-909)
Nov 17	Strict	I/A	Slower transmission	4 (1-17)	11 (2-38)	26 (4-241)
Nov 17	Strict	I/A	Densely connected	4 (1-15)	9 (2-29)	21 (4-268)
Nov 17	Strict	I/A	Sparsely connected	3 (1-16)	9 (2-32)	21 (4-345)
Nov 17	Strict	I/A	2x Hospitalization Probability	6 (1-18)	10 (3-26)	23 (6-250)
Nov 17	Strict	I/A	10x Hospitalization Probability	7 (2-23)	13 (4-40)	27 (8-390)
Nov 17	UCL	I/A	None	4 (1-13)	9 (2-29)	18 (3-402)
Nov 17	UCL	I-only	None	7 (1-15)	12 (4-33)	21 (8-401)
Dec 01	Strict	I/A	None	0 (0-9)	4 (1-20)	14 (2-276)
Dec 01	Strict	I-only	None	1 (0-12)	8 (1-25)	18 (2-325)
Dec 01	Strict	I/A	2x Hospitalization	1 (0-14)	6 (1-24)	18 (3-328)
Dec 01	Strict	I/A	10x Hospitalization	1 (0-14)	8 (1-31)	21 (3-322)
Dec 01	UCL	I/A	None	0 (0-9)	4 (1-24)	13 (2-392)
Dec 01	UCL	I-only	None	1 (0-12)	8 (1-28)	17 (2-451)

Table S5. Number infected estimates for two-phase primary (first row) and robustness analyses, with the number infected reported on 17 November, 01 December, and 15 December 2019.

Earliest Case	Clock	Ascertained	Phase 1 ascertainment rate (*r*)	Hospitalization probability ($1/D_h$)		Median # infected (95% HPD)		
				Phase 1	Phase 2	17 Nov 2019	01 Dec 2019	15 Dec 2019
Nov 17	Strict	I/A	1	1	1	3 (1-15)	7 (1-30)	17 (1-254)
Nov 17	Strict	I-only	1	1	1	5 (1-26)	10 (1-46)	20 (1-252)
Nov 17	Strict	I/A	4	1	1	3 (1-14)	7 (1-27)	17 (1-254)
Nov 17	Strict	I/A	4	2	1	3 (1-14)	8 (1-28)	17 (1-253)
Nov 17	Strict	I/A	4	2	2	3 (1-15)	8 (1-28)	18 (1-270)
Nov 17	Strict	I/A	4	10	1	3 (1-14)	8 (1-27)	18 (1-262)
Nov 17	Strict	I/A	4	10	10	3 (1-16)	8 (1-30)	19 (1-254)
Dec 01	Strict	I/A	1	1	1	0 (0-9)	3 (1-22)	12 (1-246)
Dec 01	Strict	I-only	1	1	1	1 (0-16)	5 (1-33)	14 (1-234)
Dec 01	Strict	I/A	4	1	1	0 (0-9)	3 (1-21)	12 (1-231)
Dec 01	Strict	I/A	4	2	1	0 (0-9)	3 (1-20)	12 (1-238)
Dec 01	Strict	I/A	4	2	2	0 (0-8)	3 (1-22)	12 (1-251)
Dec 01	Strict	I/A	4	10	1	0 (0-9)	3 (1-21)	12 (1-236)
Dec 01	Strict	I/A	4	10	10	0 (0-9)	3 (1-24)	13 (1-244)

Table S6. Parameterization and success/failure of two-phase compartmental epidemic simulations in primary analysis and robustness analyses.

Parameter	**Infectiousness robustness analysis**					**Ascertainment rate robustness analysis**		
Infectiousness	1/6	2/6	3/6	4/6	5/6	3/6	3/6	3/6
Ascertainment rate	1	1	1	1	1	2	4	6
Percentage failed	>99.9	98.6	93.9	85.5	77.3	91.8	89.7	87.2
Days to extinction of failed simulations[a]	7.8 (1.7-30.8)	8.2 (1.7-46.9)	8.6 (1.8-51.1)	8.4 (1.5-48.8)	8.1 (1.6-41.6)	8.7 (1.7-56.3)	8.5 (1.7-62.3)	9.0 (1.5-71.7)
Number of infections of failed simulations[a]	1 (1-4)	1 (1-12)	1 (1-13)	1 (1-13)	1 (1-10)	1 (1-13)	1 (1-11)	1 (1-11)

[a] Median and 95% range in parentheses

Table S7. FAVITES parameter values for the primary analysis. Rate values and end time are adjusted for years and density of contact network as needed.

Parameter[a]	Interpretation	Value
num_cn_nodes	Total number of nodes in network	100,000
num_seeds	Number of seeds (individuals in compartments E,P,I,A) at time 0	1
saphire_freq_s	Frequency of S individuals at time 0	99,999
saphire_freq_e	Frequency of E individuals at time 0	1
saphire_freq_p	Frequency of P individuals at time 0	0
saphire_freq_i	Frequency of I individuals at time 0	0
saphire_freq_a	Frequency of A individuals at time 0	0
saphire_freq_h	Frequency of H individuals at time 0	0
saphire_freq_r	Frequency of R individuals at time 0	0
saphire_s_to_e_seed	Target infection rate (S→E) from outside the contact network (i.e., seed infection)	0
saphire_e_to_p	Target rate of becoming presymptomatic (E→P)	365/2.9
saphire_p_to_i	Target rate of becoming ascertained (P→I)	0.15*365/2.3
saphire_p_to_a	Target rate of becoming unascertained (P→A)	(1-0.15)*365/2.3
saphire_i_to_h	Target rate of becoming hospitalized (I→H)	365/21
saphire_i_to_r	Target rate of becoming removed/recovered (I→R)	365/2.9
saphire_a_to_r	Target rate of becoming removed/recovered (A→R)	365/2.9
saphire_h_to_r	Target rate of becoming removed/recovered (H→R)	365/30
saphire_s_to_e_by_p	Target infection rate (S→E) by P	0.55*0.385*365/16
saphire_s_to_e_by_i	Target infection rate (S→E) by I	0.385*365/16
saphire_s_to_e_by_a	Target infection rate (S→E) by A	0.55*0.385*365/16
end_time	Time at which to end the transmission simulation	100/365
vts_model	Intrahost viral population growth model to use	constant
vts_growthRate	Target effective population size growth rate (used in exponential and logistic models)	0
vts_max_attempts	Maximum number of attempts to coalesce a single tree (e.g. 100) before FAVITES kills VirusTreeSimulator	1000
vts_n0	Target effective population size at time zero (used in all models)	1
vts_t50	Time point, relative to the time of infection in backwards time, at which the population is equal to half its final asymptotic value, in the logistic model	-99999

[a] FAVITES module choices and parameter values for robustness analyses can be found in the json files at https://github.com/pekarj/SC2_Index_Case.

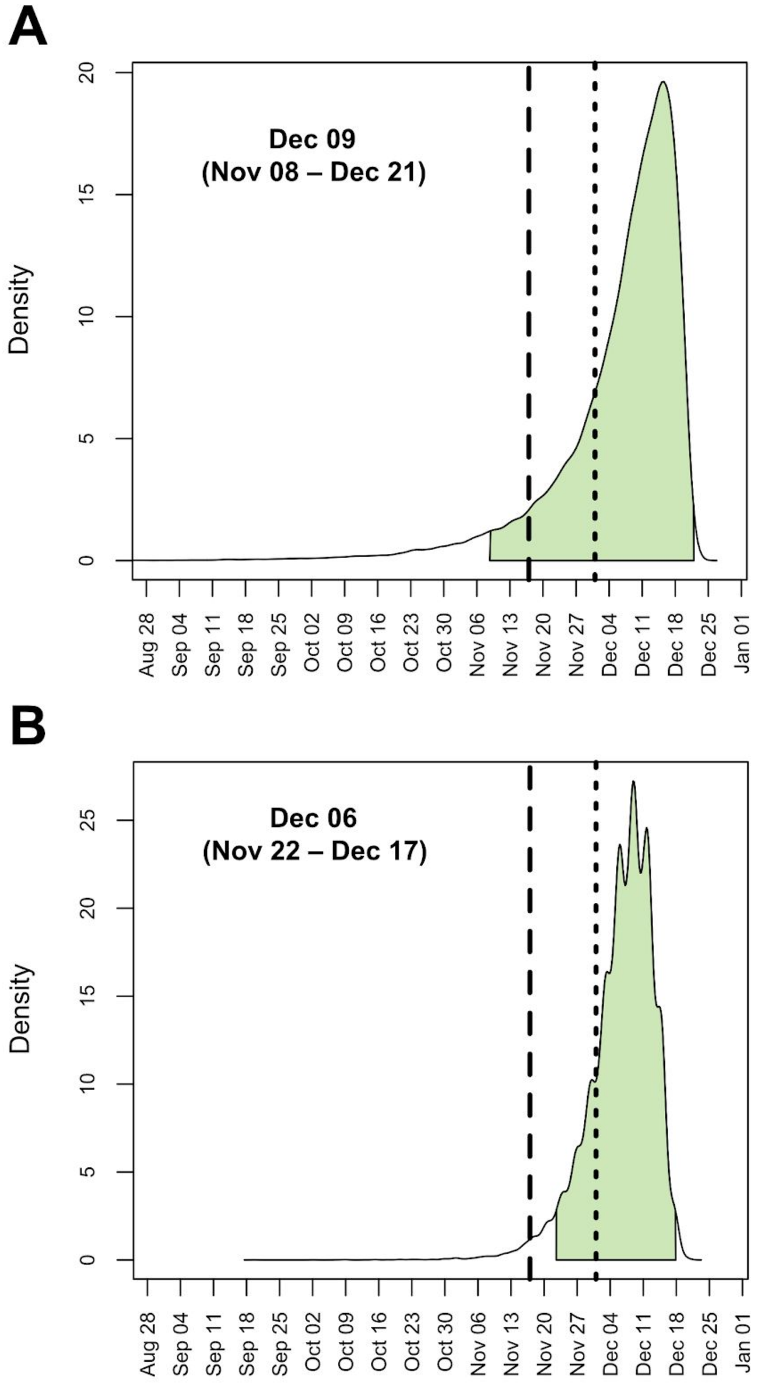

Fig. S1. Posterior distribution for the tMRCA of 583 sampled SARS-CoV-2 genomes circulating in China between December 2019 and April 2020 using (A) a relaxed molecular clock (ULD) and Skyline coalescent prior and (B) a strict clock model and Skygrid coalescent prior. Shaded area denotes 95% HPD. Long-dashed line is 17 November 2019, and short-dashed line is 01 December 2019.

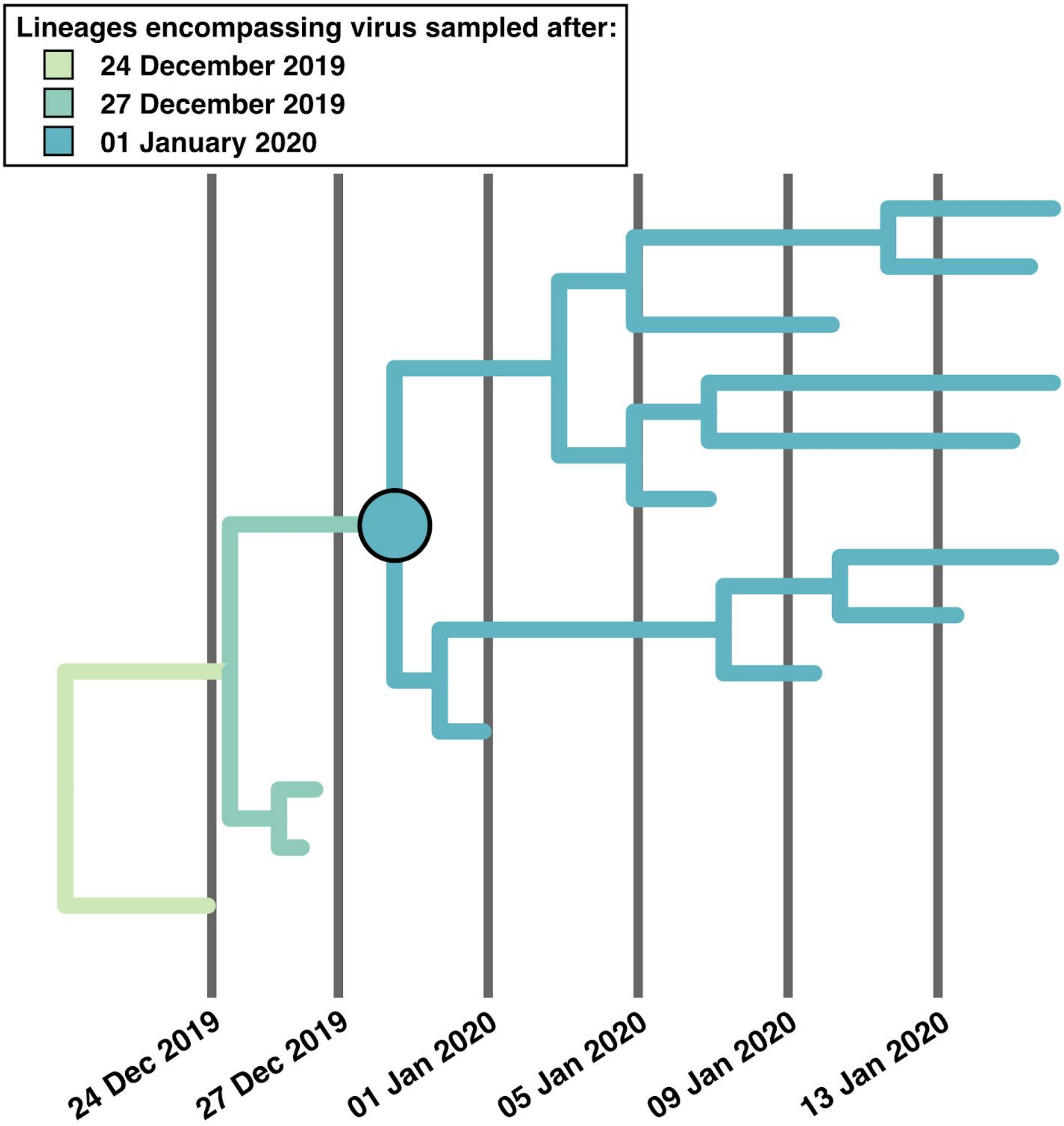

Fig. S2. Example of how coalescence of all sampled genomes can shift forward when analyzing viral genomes sampled in late-December 2019 and January 2020 as basal lineages are lost and cease to propagate. In this example, viruses sampled after 01 January 2020 have a stable tMRCA, indicated by the large circle.

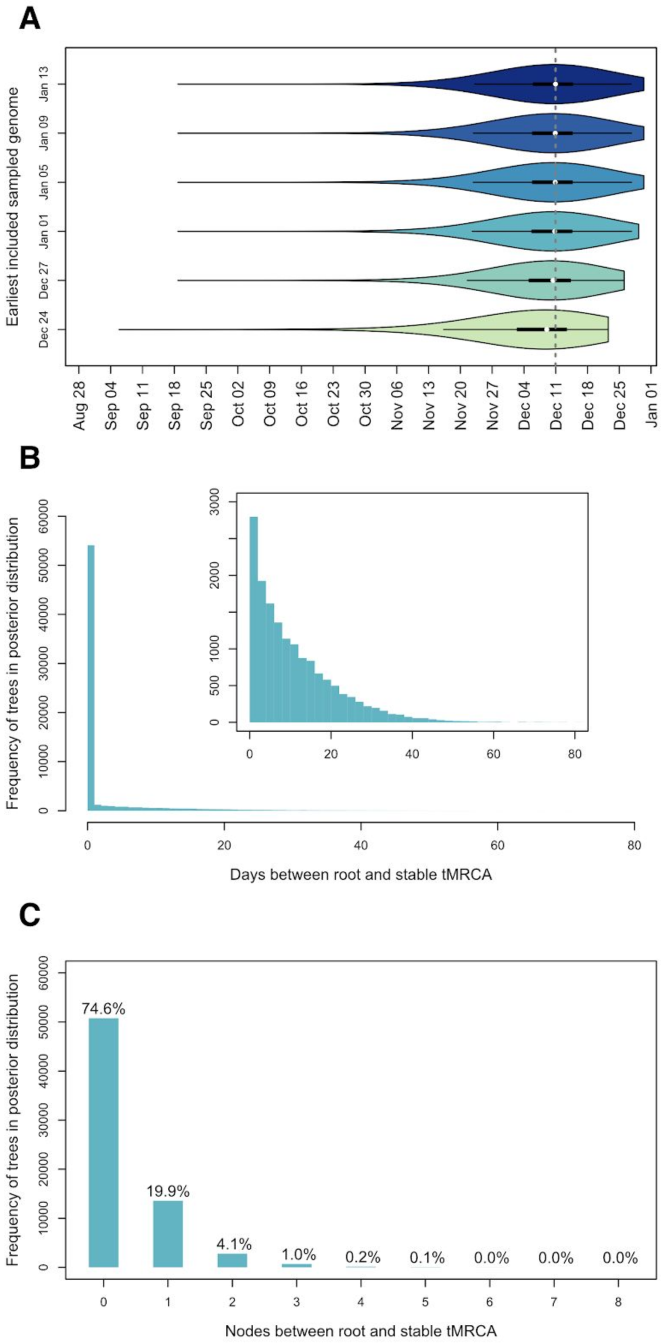

Fig. S3. Shifting of tMRCA in empirical molecular clock analyses. (A) Violin plots of tMRCA estimates upon excluding the earliest sampled genomes between 24 December 2019 and 13 January 2020 (darker colors exclude progressively more early sampled genomes). Mean tMRCAs are depicted with white dots. The dashed grey line represents the mean estimate when including only genotypes sampled on 13 January 2020 or later. (B) The number of days between the tMRCA of all genomes (i.e., root) and the tMRCA of genomes sampled on or after 01 January 2020. Inset excludes BEAST samples with <1 day shift. (C) Number of nodes between the tMRCA of all genomes (i.e., root) and the tMRCA of genomes sampled on or after 01 January 2020. Percentage denotes the number of BEAST samples represented in column.

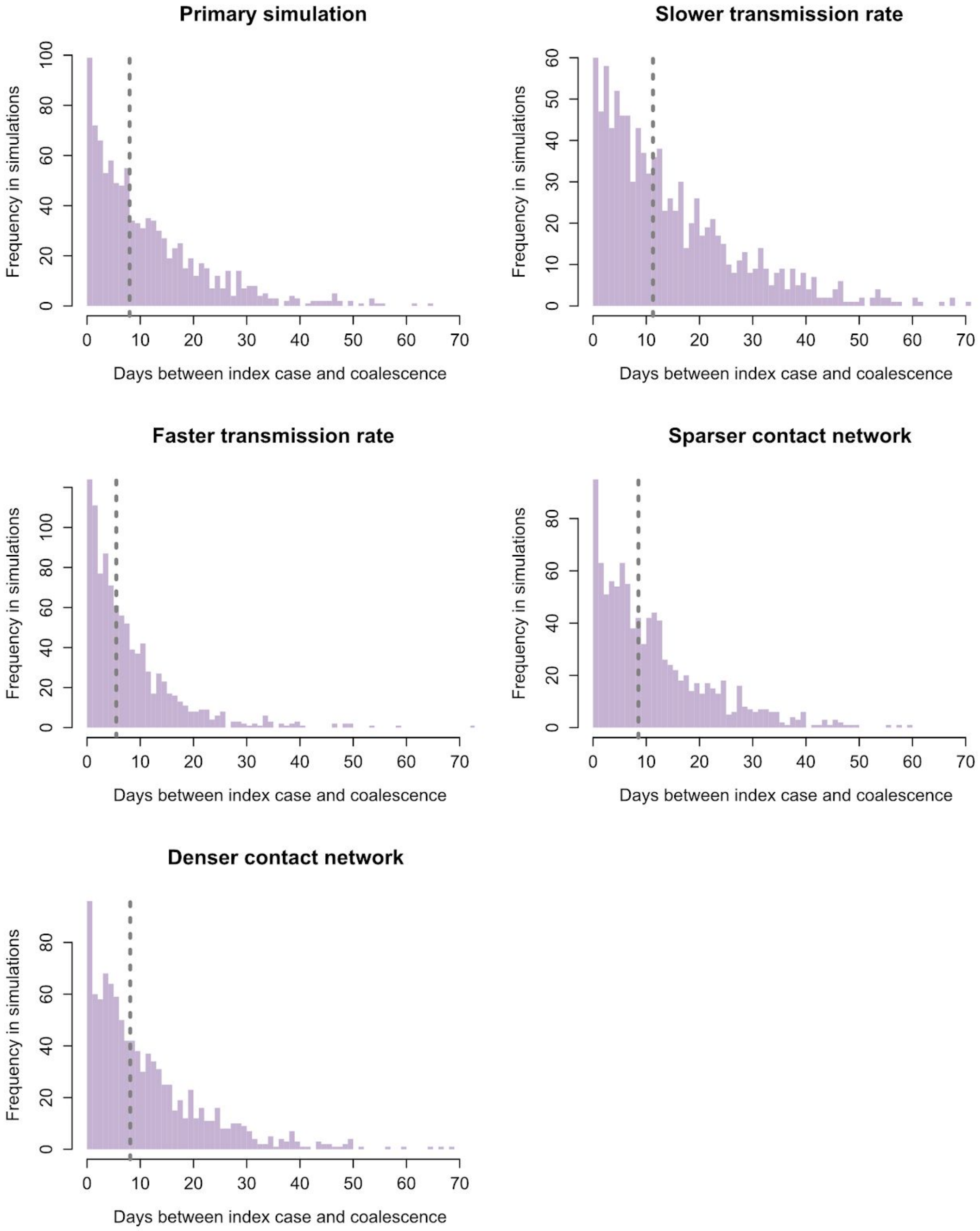

Fig. S4. FAVITES robustness analyses. Number of days between the index case the coalescence when simulated transmission rate is slower and faster and when the contact network has fewer (mean=10) and more (mean=26) contacts.

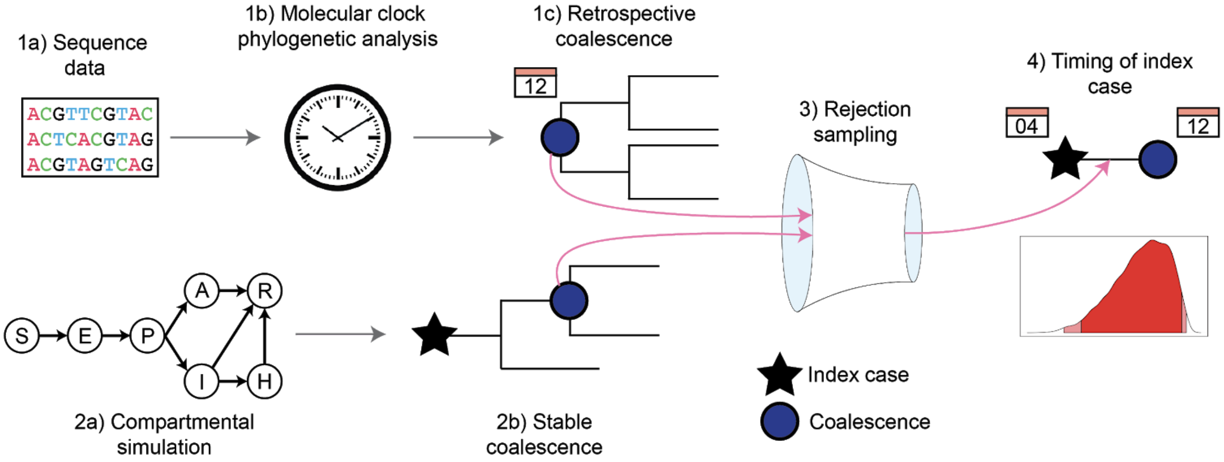

Fig. S5. Combined simulation and phylogenetic workflows to estimate the timing of the Hubei index case in a one-phase epidemic. (1a) Using sequence and epidemiological data, (1b) BEAST performs a phylodynamic molecular clock analysis to (1c) determine the tMRCA. (2a) FAVITES simulates the epidemic in Hubei using a SAPHIRE compartmental model (22) and (2b) estimates a prior distribution for the time from index case to the stable coalescence. The results of (1) and (2) are combined via rejection sampling (3; Fig. S6) to (4) determine the timing of the index case and its posterior distribution.

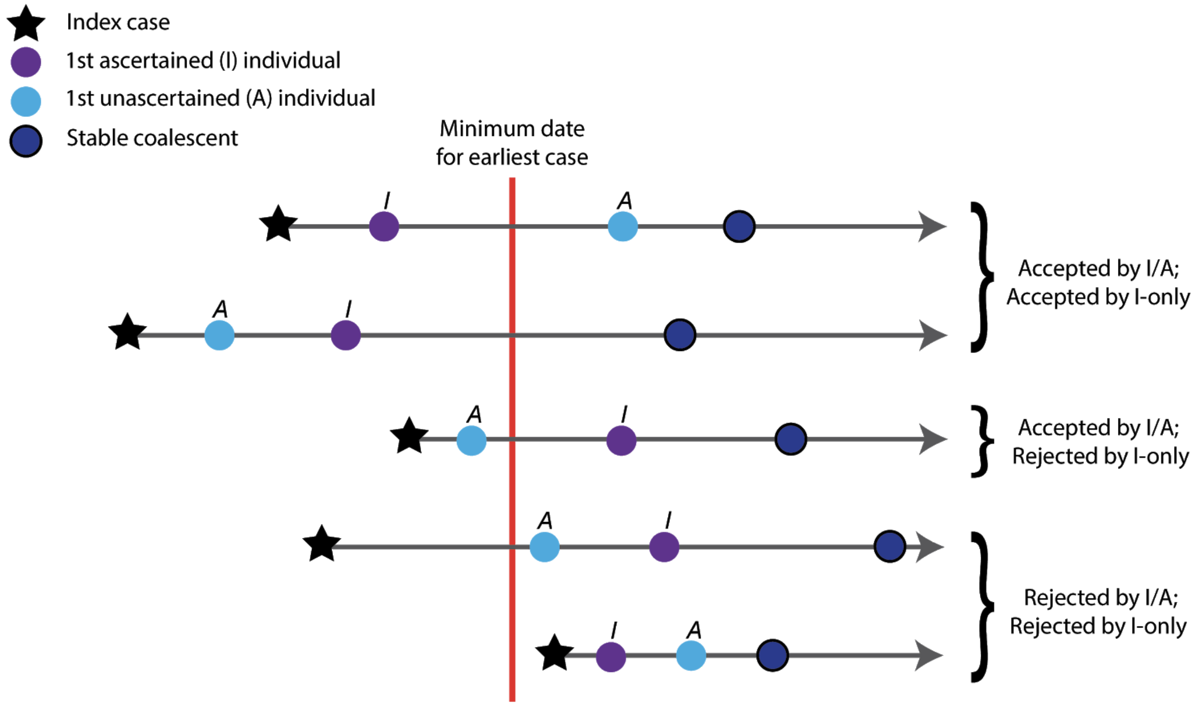
Index case
1st ascertained (I) individual
1st unascertained (A) individual
Stable coalescent
Minimum date for earliest case
I
A
A
I
A
I
A
I
I
A
Accepted by I/A; Accepted by I-only
Accepted by I/A; Rejected by I-only
Rejected by I/A; Rejected by I-only

Fig. S6. Illustration of the interplay rejection sampling scheme and the SAPHIRE compartmental model. The top two examples would be accepted by both I/A and I-only rejection sampling, because the first ascertained (I) case came before the minimum date for the earliest case In contrast, the middle example would only be accepted by I/A sampling, because the first unascertained (A) case came before the minimum date for the earliest case, but the first ascertained (I) case did not. The bottom two examples would be always be rejected, because both the first ascertained and unascertained cases came after the minimum date for the earliest case Note that coalescence can happen any time after the index case is first infected, and there is no requirement for coalescence to occur after the first ascertained and unascertained individuals. The index case begins in the exposed (E) compartment and can be the individual that first transitions into the A or I compartments; though other subsequently infected individuals can progress to the A or I compartments before index case. This methodology applies to both one-phase and two-phase epidemic simulations, where ascertained and unascertained cases can come from either Phase 1 or Phase 2 in the latter.

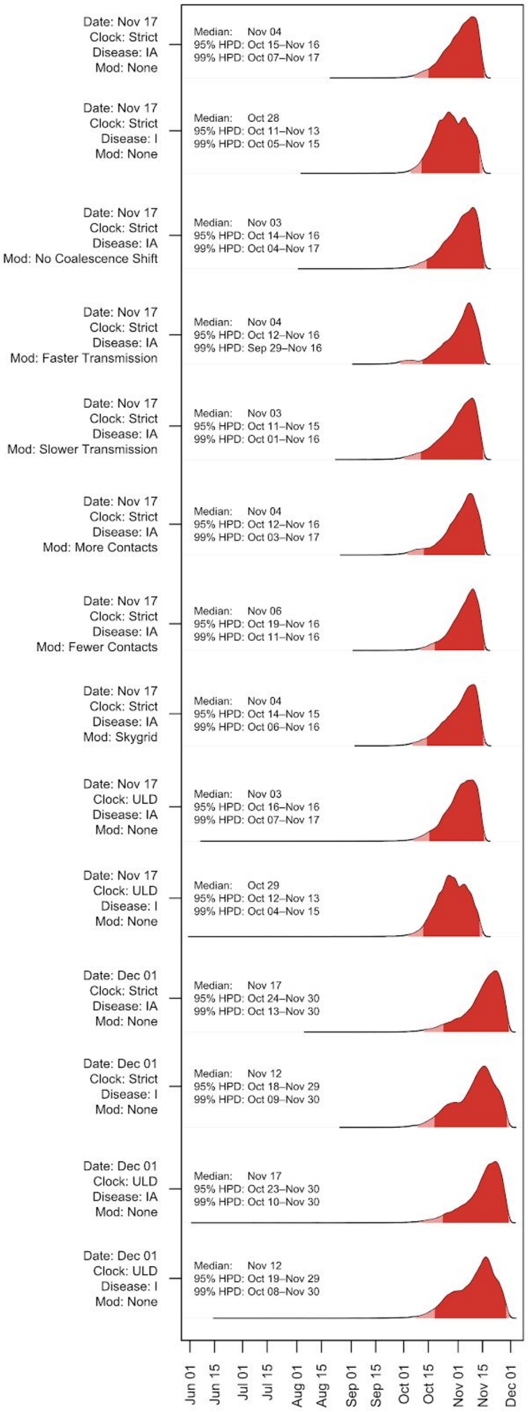

Fig. S7. Robustness analysis for timing of SARS-CoV-2 index case in Hubei province. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red. Primary analysis is shown on top. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. 'Mod' denotes if any robustness modifications were explored.

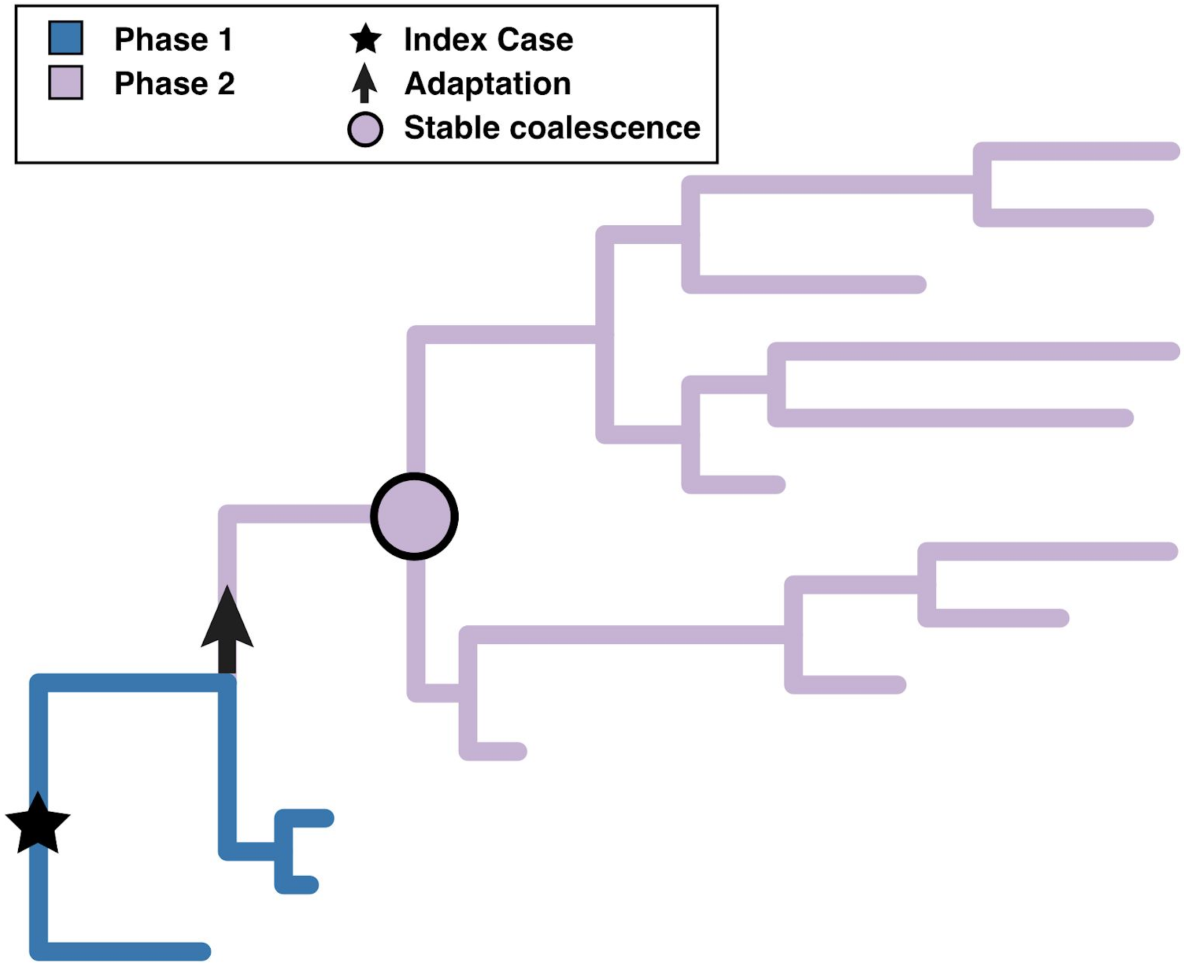

Fig. S8. Example phylogeny of the two-phase epidemic simulations. The index case occurs during Phase 1, characterized by a less-fit viral variant (shown in blue). At some point before Phase 1 goes extinct, an adapted variant arises (arrow), giving rise to Phase 2, characterized by the variant observed in Hubei province (shown in light purple). The stable coalescence (circle) occurs during Phase 2.

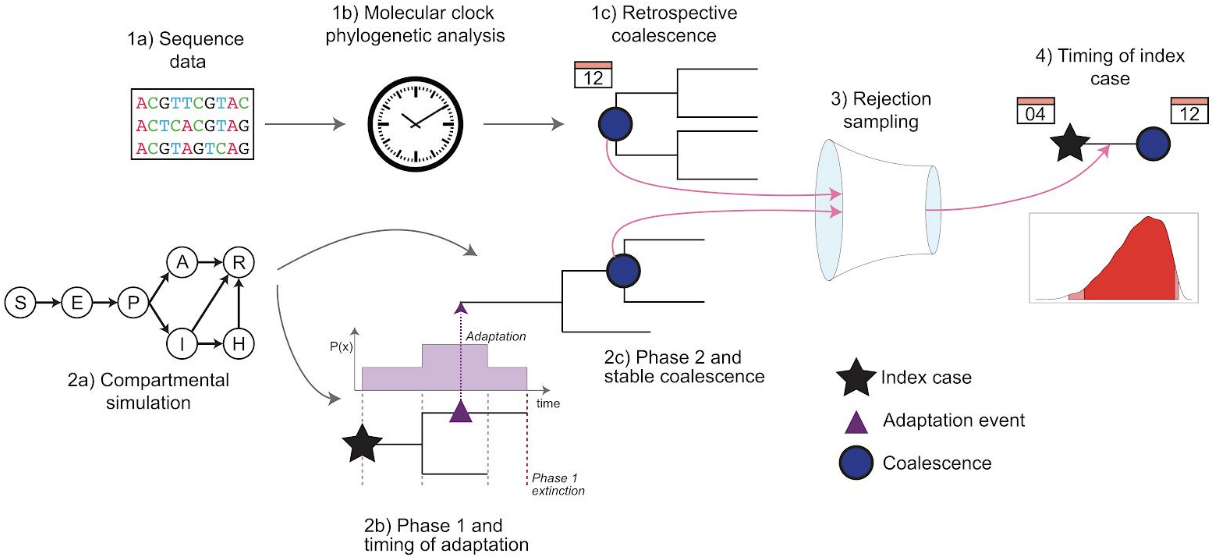

Fig. S9. Combined simulation and phylogenetic workflows to estimate the timing of the Hubei index case in a two-phase epidemic. (1a) Using sequence and epidemiological data, (1b) BEAST performs a phylodynamic molecular clock analysis to (1c) determine the tMRCA. (2a) FAVITES simulates the epidemic in Hubei using a SAPHIRE compartmental model, (2b) estimates a prior distribution for the length of Phase 1 of the epidemic, and (2c) estimates a prior distribution for the time from adaptation to the stable coalescence for Phase 2 of the epidemic. The timing of adaptation (2b) is calculated based on the number of circulating lineages at a given time in Phase 1 (see Methods). The results of (1) and (2) are combined via rejection sampling (3; Fig. S6) to (4) determine the timing of the index case and its posterior distribution.

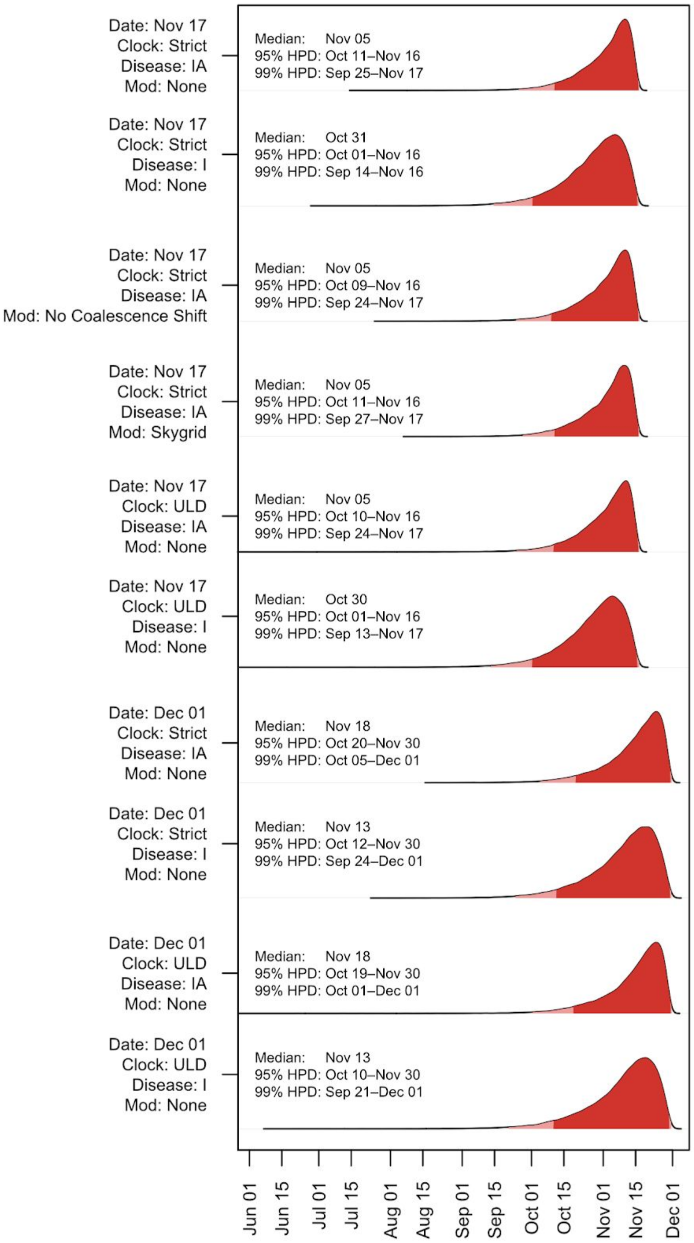

Fig. S10. Robustness analysis for timing of SARS-CoV-2 index case in Hubei province in the two-phase epidemic simulations. Primary two-phase analysis (from Fig. 3H) is shown on top. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. 'Mod' denotes if any robustness modifications were explored. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red.

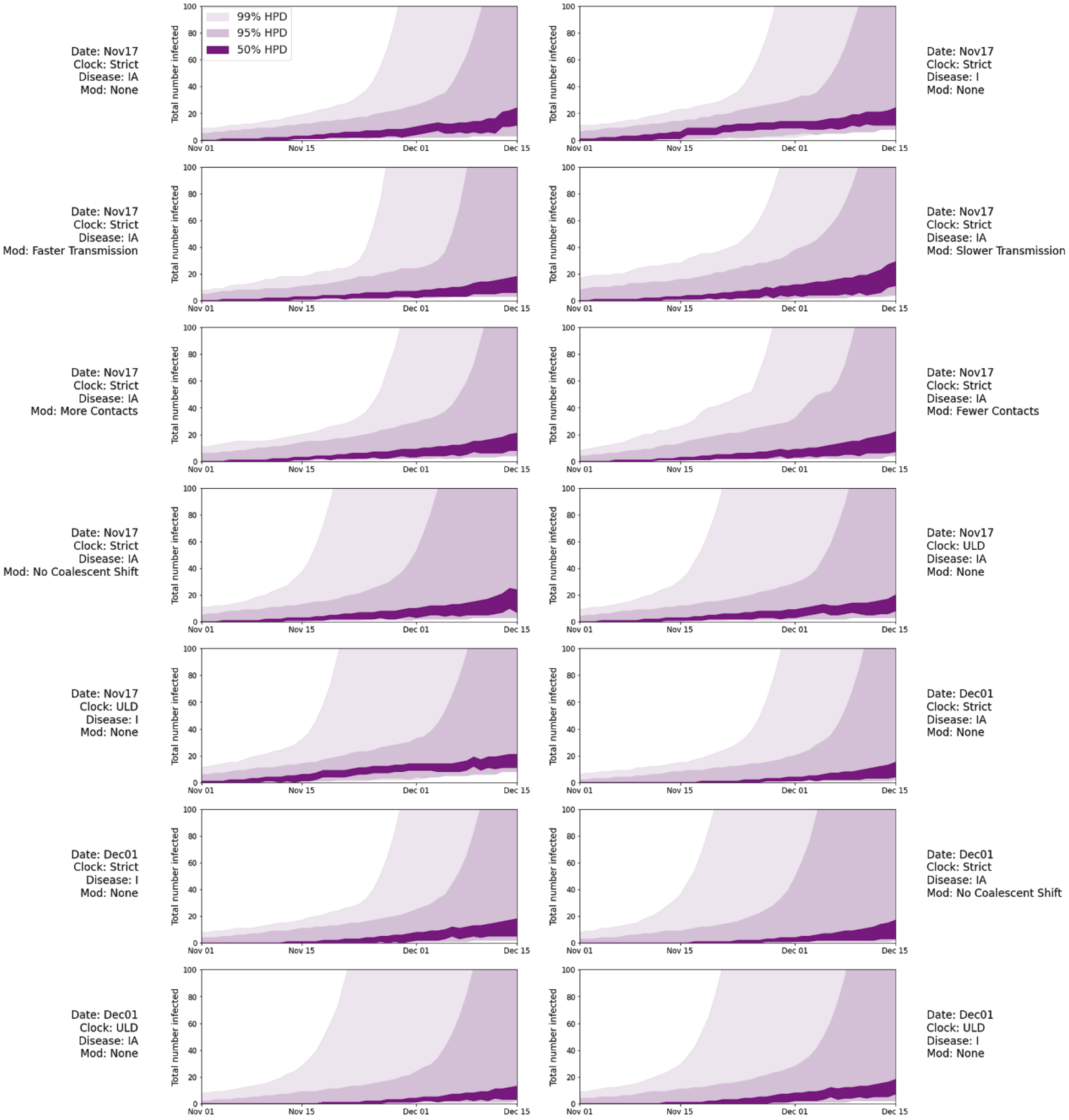

Fig. S11. Robustness analysis for the number of people infected SARS-CoV-2 based on the SAPHIRE model in late 2019. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling.

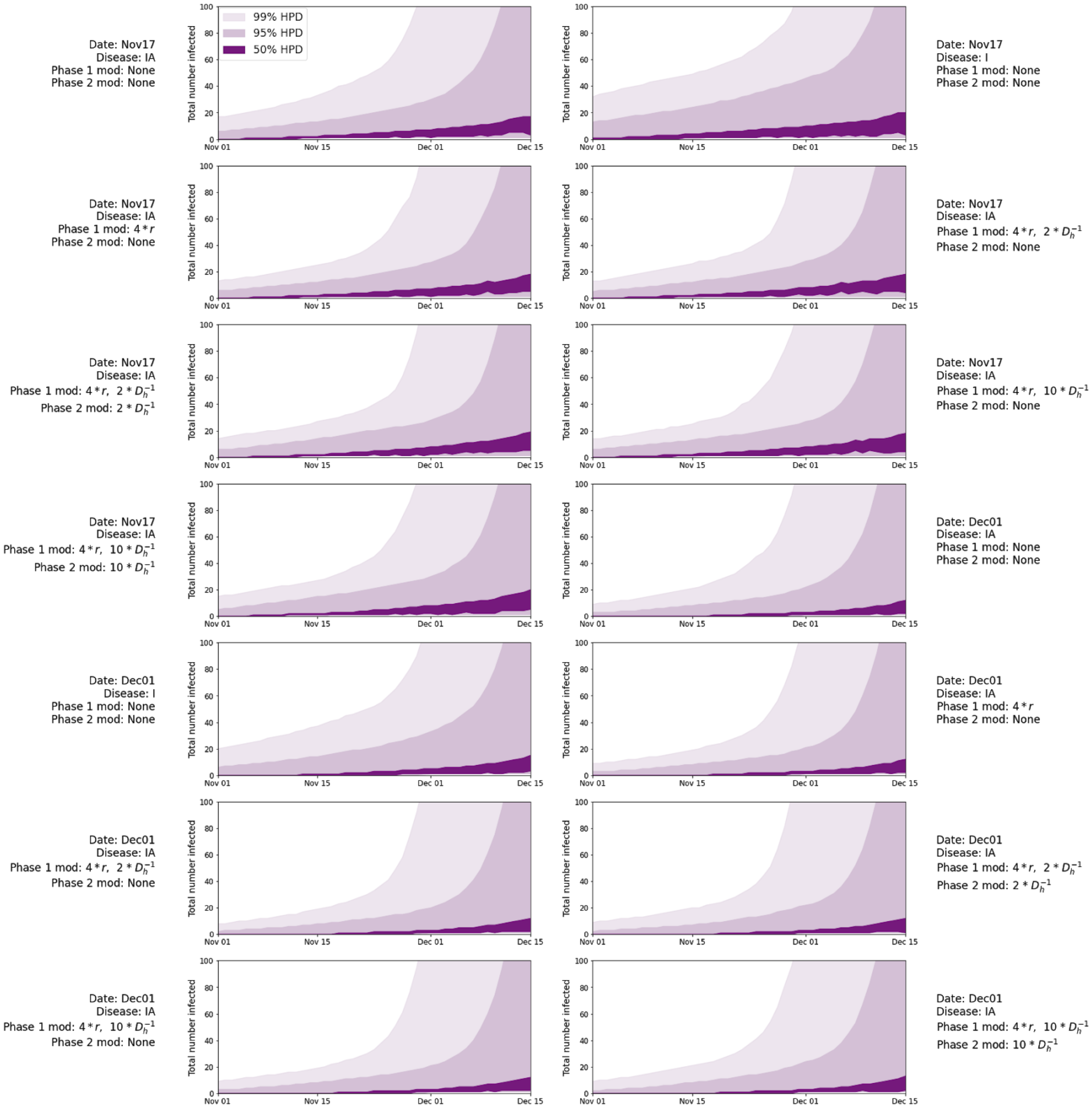

Fig. S12. Robustness analysis for the number of people infected SARS-CoV-2 based on the SAPHIRE model in late 2019 using two-phase simulations. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. D_h is the average time to hospitalization and r is the ascertainment rate, with the same values as in the primary analysis (**Table S1**).

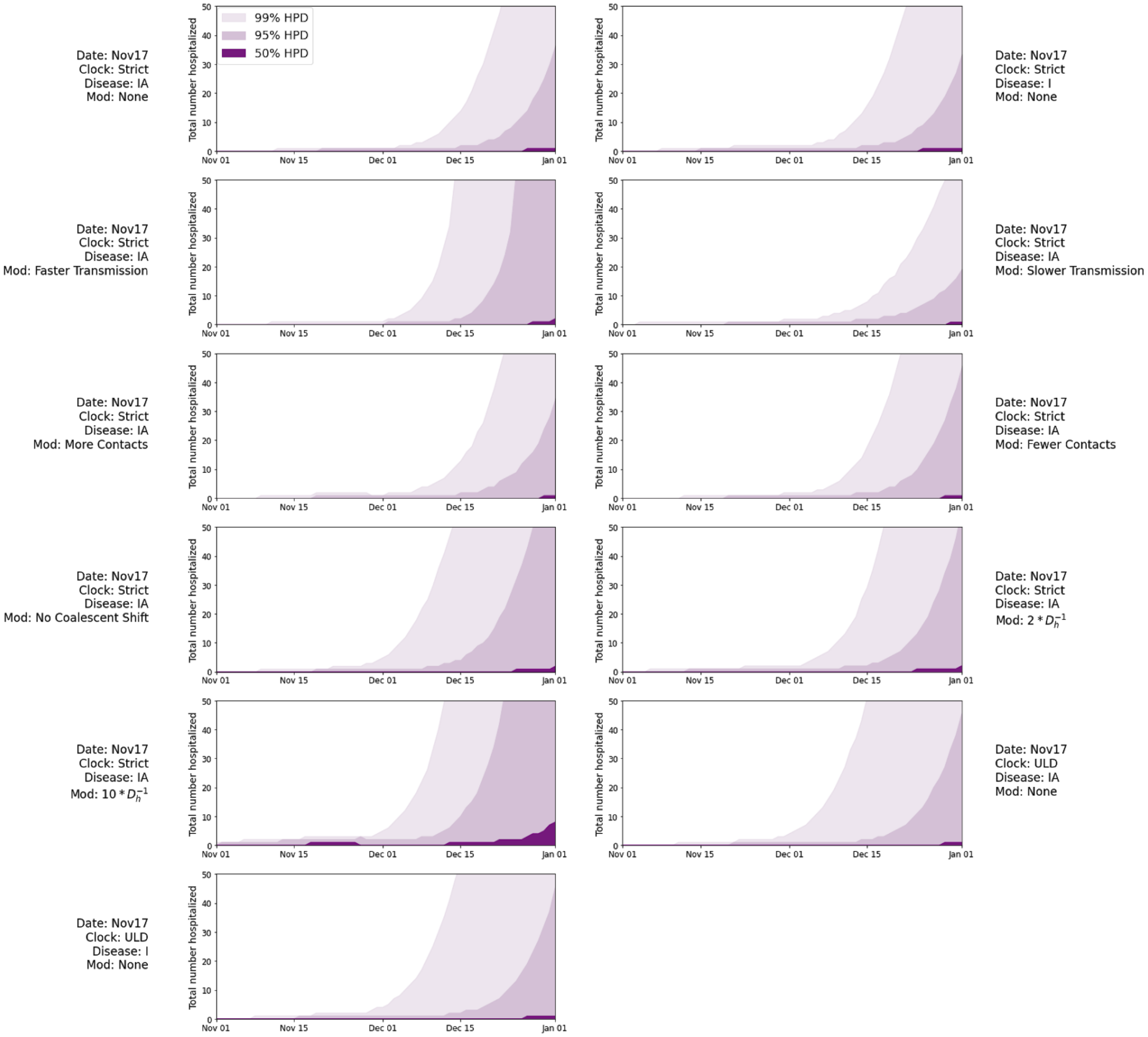

Fig. S13. Number of people hospitalized with SARS-CoV-2 based on the SAPHIRE model in late 2019. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling.

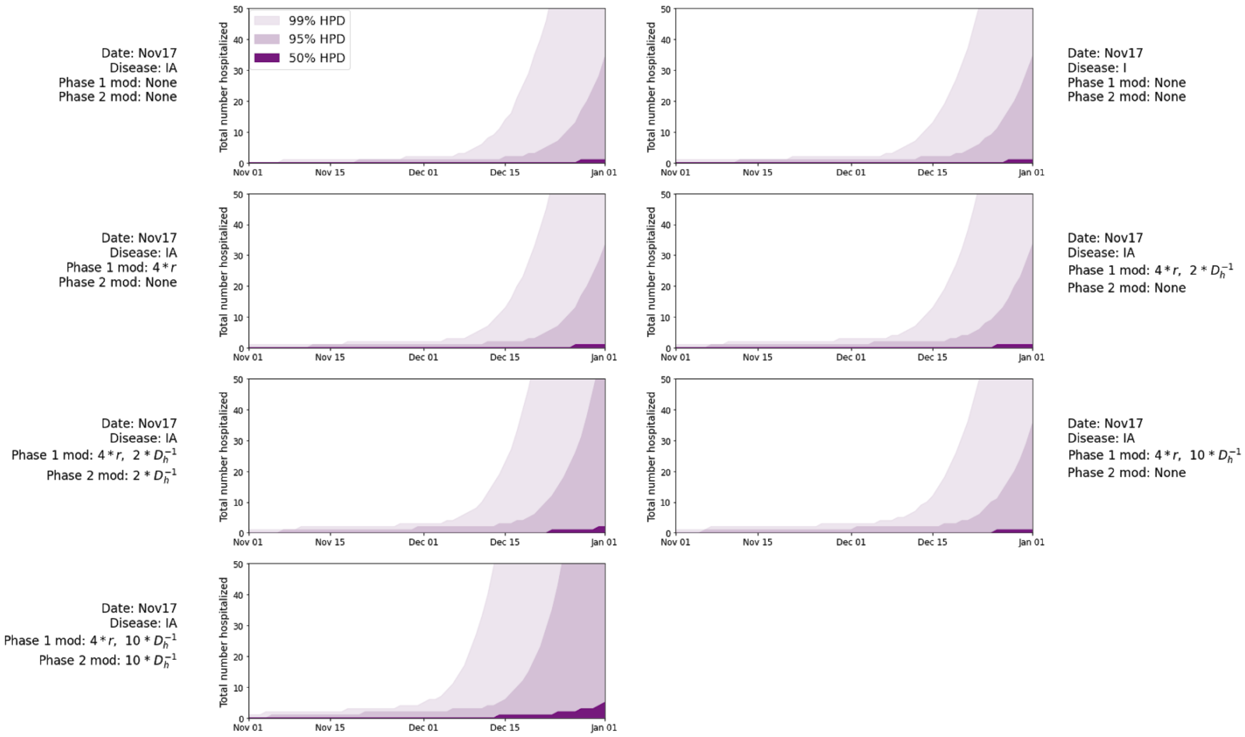

Fig. S14. Number of people hospitalized with SARS-CoV-2 based on the SAPHIRE model in late 2019 using two-phase simulations. Innermost shading is 50% HPD, middle shading is 95% HPD, and outer shading is 99% HPD. Dates indicates the minimum bound for rejection sampling (Nov 17 or Dec 01). Clock indicates whether the tMRCA was inferred using a strict or relaxed (ULD) clock. Disease denotes which stage of infection in the SAPHIRE model (IA, Ascertained or Unascertained; or I, Ascertained only) must have been reached by the given date for rejection sampling. D_h is the average time to hospitalization and r is the ascertainment rate, with the same values as in the primary analysis (**Table S1**).

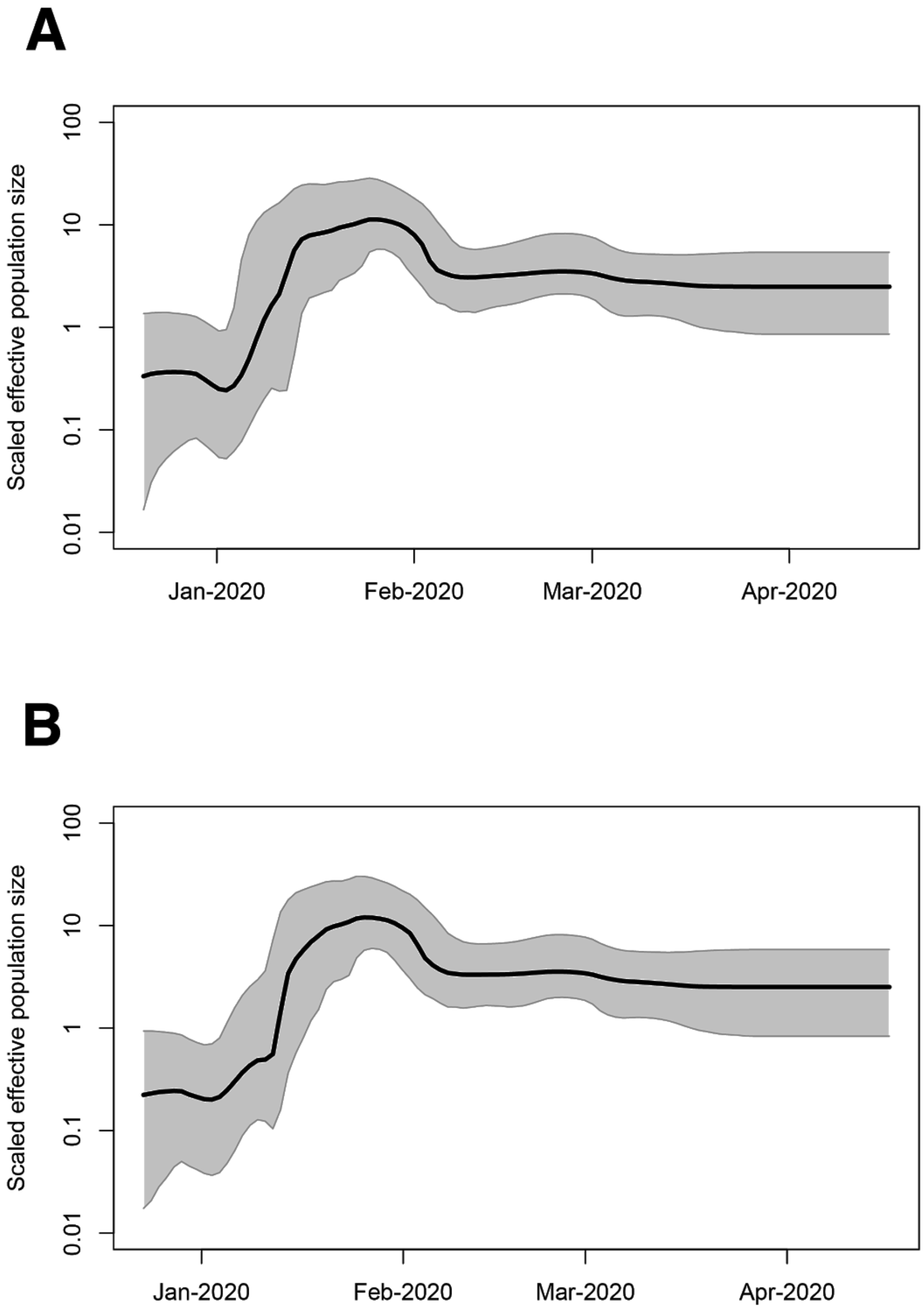

Fig. S15. Bayesian skyline plot reconstruction. (A) Strict molecular clock analysis and (B) ULD relaxed molecular clock. Solid line is the median estimate. Shaded area is the 95% HPD.

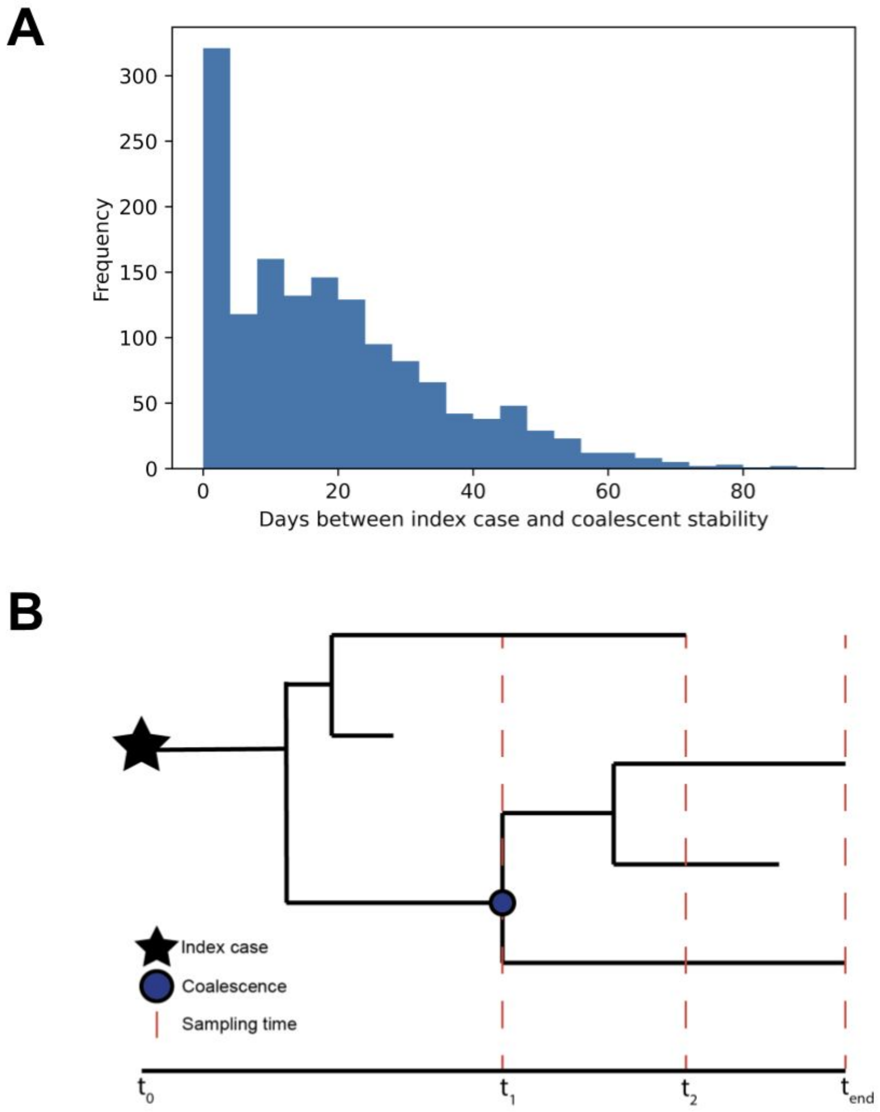

Fig. S16. Stable coalesce in forward compartmental simulations. (A) Days between index case infection and time stable coalescence is achieved in primary simulations. (B) Distinction between the date of index case infection (at t_0), the tMRCA of surviving lineages (at t_1), and the time at which stable coalescence is achieved when the last basal lineage goes extinct (at t_2).

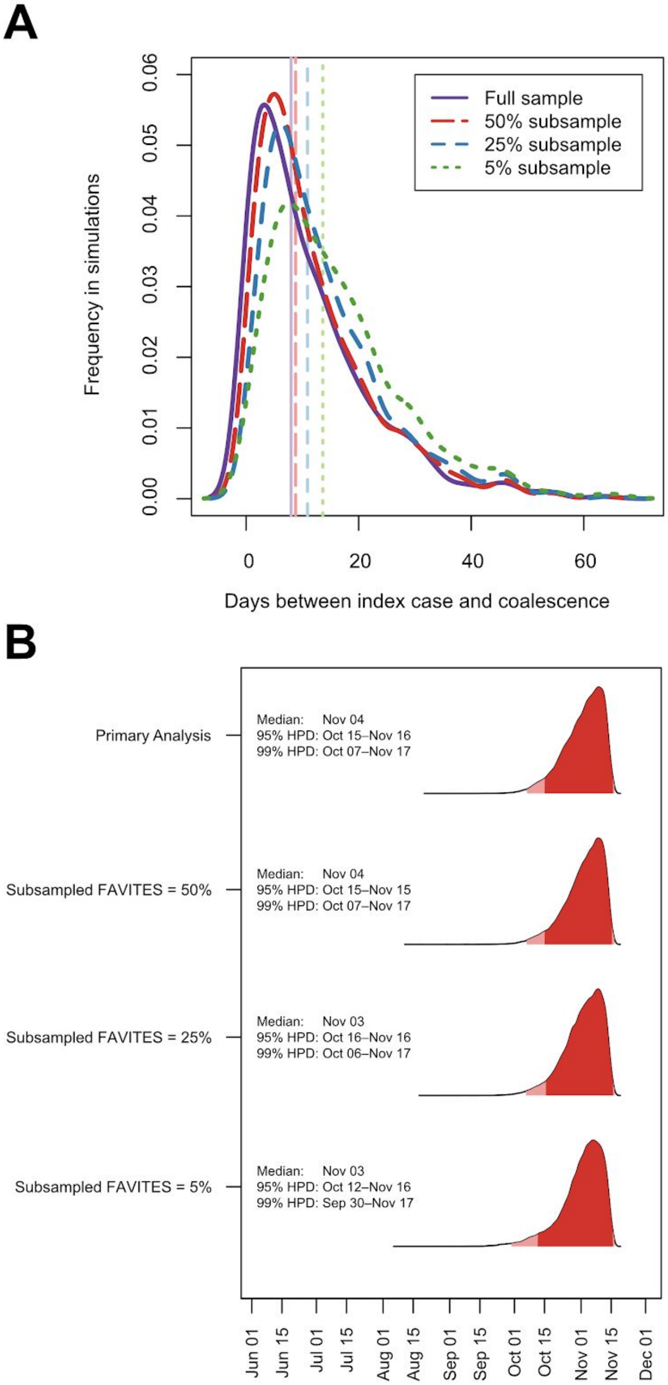

Fig. S17. Impact of subsampling taxa from FAVITES epidemic simulations. (A) Days between index case and stable coalescent with varying amounts of subsampling. (B) Timing of the index case in Hubei province upon subsampling.The primary analysis (from Fig. 3C) is shown on top. Inference is conditioned on an ascertained or unascertained case (I or A in SAPHIRE model) by 17 November using a strict clock. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red.

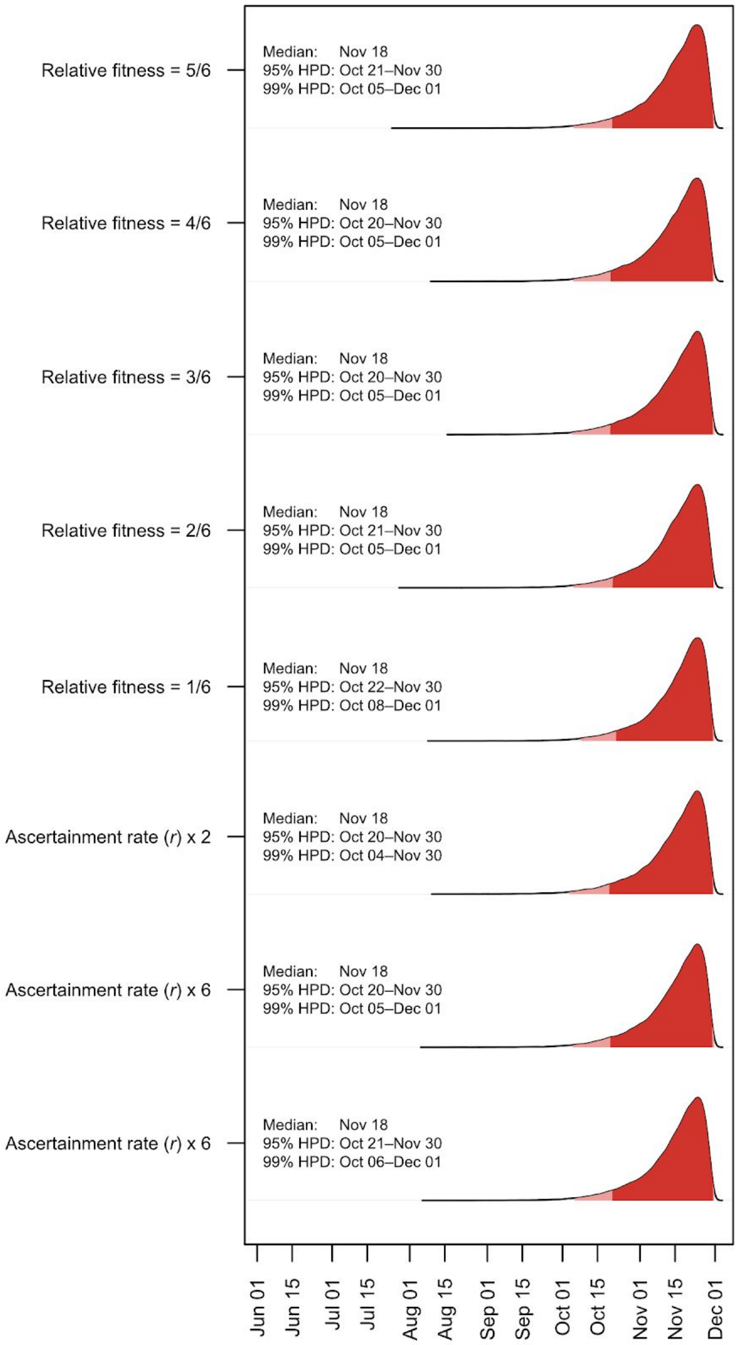

Fig. S18. Impact of relative fitness and virulence of Phase 1 variant on the timing of the index case in Hubei province during two-phase epidemic simulations. The primary two-phase analysis described in the main text (Fig. 3F-H) used a relative fitness of 3/6 (or 0.5). Virulence is modulated by the ascertainment rate (*r* in the SAPHIRE model). Inference is conditioned on an ascertained or unascertained case (I or A in SAPHIRE model) by 17 November using a strict clock. The 95% HPD is shown in dark red, and the 99% HPD is shown in light red.

Data S1.
GISAID acknowledgements table.

References and Notes

1. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan; China Novel Coronavirus Investigating and Research Team, A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020). doi:10.1056/NEJMoa2001017 Medline

2. Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau, J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Liu, W. Tu, C. Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z. Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. Wu, G. F. Gao, B. J. Cowling, B. Yang, G. M. Leung, Z. Feng, Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020). doi:10.1056/NEJMoa2001316 Medline

3. World Health Organization, "Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)"; www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19).

4. H. Fei, X. Yinyin, C. Hui, W. Ni, D. Xin, C. Wei, L. Tao, H. Shitong, S. Miaomiao, C. Mingting, S. Keshavjee, Z. Yanlin, D. P. Chin, L. Jianjun, The impact of the COVID-19 epidemic on tuberculosis control in China. *Lancet Reg. Health West. Pac.* **3**, 100032 (2020). doi:10.1016/j.lanwpc.2020.100032

5. World Health Organization, "WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020"; www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

6. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao, L. Guo, J. Xie, G. Wang, R. Jiang, Z. Gao, Q. Jin, J. Wang, B. Cao, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020). doi:10.1016/S0140-6736(20)30183-5 Medline

7. X. Zhang, Y. Tan, Y. Ling, G. Lu, F. Liu, Z. Yi, X. Jia, M. Wu, B. Shi, S. Xu, J. Chen, W. Wang, B. Chen, L. Jiang, S. Yu, J. Lu, J. Wang, M. Xu, Z. Yuan, Q. Zhang, X. Zhang, G. Zhao, S. Wang, S. Chen, H. Lu, Viral and host factors related to the clinical outcome of COVID-19. *Nature* **583**, 437–440 (2020). doi:10.1038/s41586-020-2355-0 Medline

8. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020). doi:10.1038/s41586-020-2008-3 Medline

9. R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020). doi:10.1016/S0140-6736(20)30251-8 Medline

10. J. Ma, “Coronavirus: China’s first confirmed Covid-19 case traced back to November 17.” *South China Morning Post* (2020); www.scmp.com/news/china/society/article/3074991/coronavirus-chinas-first-confirmed-covid-19-case-traced-back.

11. A. Rambaut, “Phylodynamic Analysis | 176 genomes | 6 Mar 2020.” *Virological* (2020); https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356.

12. S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey, G. Baele, Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, veaa061 (2020). doi:10.1093/ve/veaa061 Medline

13. K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, R. F. Garry, The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020). doi:10.1038/s41591-020-0820-9 Medline

14. L. du Plessis, O. Pybus, “Further musings on the tMRCA.” *Virological* (2020); https://virological.org/t/further-musings-on-the-tmrca/340.

15. N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pépin, D. Posada, M. Peeters, O. G. Pybus, P. Lemey, HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61 (2014). doi:10.1126/science.1256739 Medline

16. M. Worobey, M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J.-J. Muyembe, J.-M. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. P. Gilbert, S. M. Wolinsky, Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664 (2008). doi:10.1038/nature07390 Medline

17. B. F. Keele, F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, S. Loul, E. M. Ngole, Y. Bienvenue, E. Delaporte, J. F. Y. Brookfield, P. M. Sharp, G. M. Shaw, M. Peeters, B. H. Hahn, Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**, 523–526 (2006). doi:10.1126/science.1126531 Medline

18. T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, M.-L. Huang, A. Nalla, G. Pepper, A. Reinhardt, H. Xie, L. Shrestha, T. N. Nguyen, A. Adler, E. Brandstetter, S. Cho, D. Giroux, P. D. Han, K. Fay, C. D. Frazar, M. Ilcisin, K. Lacombe, J. Lee, A. Kiavand, M. Richardson, T. R. Sibley, M. Truong, C. R. Wolf, D. A. Nickerson, M. J. Rieder, J. A. Englund, J. Hadfield, E. B. Hodcroft, J. Huddleston, L. H. Moncla, N. F. Müller, R. A. Neher, X. Deng, W. Gu, S. Federman, C. Chiu, J. S. Duchin, R. Gautom, G. Melly, B. Hiatt, P. Dykema, S. Lindquist, K. Queen, Y. Tao, A. Uehara, S. Tong, D. MacCannell, G. L. Armstrong, G. S. Baird, H. Y. Chu, J. Shendure, K. R. Jerome; Seattle Flu Study Investigators, Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020). doi:10.1126/science.abc0523 Medline

19. Q. Bi, Y. Wu, S. Mei, C. Ye, X. Zou, Z. Zhang, X. Liu, L. Wei, S. A. Truelove, T. Zhang, W. Gao, C. Cheng, X. Tang, X. Wu, Y. Wu, B. Sun, S. Huang, Y. Sun, J. Zhang, T. Ma, J. Lessler, T. Feng, Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020). doi:10.1016/S1473-3099(20)30287-5 Medline

20. M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, A. Rambaut, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018). doi:10.1093/ve/vey016 Medline

21. L. Pipes, H. Wang, J. P. Huelsenbeck, R. Nielsen, Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *Mol. Biol. Evol.* 10.1093/molbev/msaa316 (2020). doi:10.1093/molbev/msaa316 Medline

22. X. Hao, S. Cheng, D. Wu, T. Wu, X. Lin, C. Wang, Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* **584**, 420–424 (2020). doi:10.1038/s41586-020-2554-8 Medline

23. N. Moshiri, M. Ragonnet-Cronin, J. O. Wertheim, S. Mirarab, FAVITES: Simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics* **35**, 1852–1861 (2019). doi:10.1093/bioinformatics/bty921 Medline

24. R. Laxminarayan, B. Wahl, S. R. Dudala, K. Gopal, C. Mohan B, S. Neelima, K. S. Jawahar Reddy, J. Radhakrishnan, J. A. Lewnard, Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science* **370**, 691–697 (2020). doi:10.1126/science.abd7672 Medline

25. A. Endo, S. Abbott, A. J. Kucharski, S. Funk; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020). doi:10.12688/wellcomeopenres.15842.3 Medline

26. X.-K. Xu, X. F. Liu, Y. Wu, S. T. Ali, Z. Du, P. Bosetti, E. H. Y. Lau, B. J. Cowling, L. Wang, Reconstruction of transmission pairs for novel coronavirus disease 2019 (COVID-19) in mainland China: Estimation of superspreading events, serial interval, and hazard of infection. *Clin. Infect. Dis.* **71**, 3163–3167 (2020). doi:10.1093/cid/ciaa790 Medline

27. D. C. Adam, P. Wu, J. Y. Wong, E. H. Y. Lau, T. K. Tsang, S. Cauchemez, G. M. Leung, B. J. Cowling, Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.* **26**, 1714–1719 (2020). doi:10.1038/s41591-020-1092-0 Medline

28. 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). doi:10.1016/j.bsheal.2019.10.004 Medline

29. M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020). doi:10.1126/science.abc8169 Medline

30. L. du Plessis, J. T. McCrone, A. E. Zarebski, V. Hill, C. Ruis, B. Gutierrez, J. Raghwani, J. Ashworth, R. Colquhoun, T. R. Connor, N. R. Faria, B. Jackson, N. J. Loman, Á. O’Toole, S. M. Nicholls, K. V. Parag, E. Scher, T. I. Vasylyeva, E. M. Volz, A. Watts, I. I. Bogoch, K. Khan, D. M. Aanensen, M. U. G. Kraemer, A. Rambaut, O. G. Pybus; COVID-19 Genomics UK (COG-UK) Consortium, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021). doi:10.1126/science.abf2946 Medline

31. A. Deslandes, V. Berti, Y. Tandjaoui-Lambotte, C. Alloui, E. Carbonnelle, J. R. Zahar, S. Brichler, Y. Cohen, SARS-CoV-2 was already spreading in France in late December 2019. *Int. J. Antimicrob. Agents* **55**, 106006 (2020). doi:10.1016/j.ijantimicag.2020.106006 Medline

32. X. Deng, W. Gu, S. Federman, L. du Plessis, O. G. Pybus, N. R. Faria, C. Wang, G. Yu, B. Bushnell, C.-Y. Pan, H. Guevara, A. Sotomayor-Gonzalez, K. Zorn, A. Gopez, V. Servellita, E. Hsu, S. Miller, T. Bedford, A. L. Greninger, P. Roychoudhury, L. M. Starita, M. Famulare, H. Y. Chu, J. Shendure, K. R. Jerome, C. Anderson, K. Gangavarapu, M. Zeller, E. Spencer, K. G. Andersen, D. MacCannell, C. R. Paden, Y. Li, J. Zhang, S. Tong, G. Armstrong, S. Morrow, M. Willis, B. T. Matyas, S. Mase, O. Kasirye, M. Park, G. Masinde, C. Chan, A. T. Yu, S. J. Chai, E. Villarino, B. Bonin, D. A. Wadford, C. Y. Chiu, Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582–587 (2020). doi:10.1126/science.abb9263 Medline

33. M. A. Jorden, S. L. Rudman, E. Villarino, S. Hoferka, M. T. Patel, K. Bemis, C. R. Simmons, M. Jespersen, J. Iberg Johnson, E. Mytty, K. D. Arends, J. J. Henderson, R. W. Mathes, C. X. Weng, J. Duchin, J. Lenahan, N. Close, T. Bedford, M. Boeckh, H. Y. Chu, J. A. Englund, M. Famulare, D. A. Nickerson, M. J. Rieder, J. Shendure, L. M. Starita; CDC COVID-19 Response Team, Evidence for limited early spread of COVID-19 within the United States, January–February 2020. *Morb. Mortal. Wkly. Rep.* **69**, 680–684 (2020). doi:10.15585/mmwr.mm6922e1 Medline

34. G. Chavarria-Miró, E. Anfruns-Estrada, S. Guix, M. Paraira, B. Galofré, G. Sánchez, R. M. Pintó, A. Bosch, Sentinel surveillance of SARS-CoV-2 in wastewater anticipates the occurrence of COVID-19 cases. medRxiv (2020); https://medrxiv.org/lookup/doi/10.1101/2020.06.13.20129627.

35. G. Fongaro, P. H. Stoco, D. S. Marques Souza, E. C. Grisard, M. E. Magri, P. Rogovski, M. A. Schörner, F. Hartmann Barazzetti, A. P. Christoff, L. F. V. de Oliveira, M. L. Bazzo, G. Wagner, M. Hernández, D. Rodriguez-Lázaro, SARS-CoV-2 in human sewage in Santa Catalina, Brazil, November 2019. medRxiv (2020); www.medrxiv.org/content/10.1101/2020.06.26.20140731v1.

36. G. La Rosa, P. Mancini, G. Bonanno Ferraro, C. Veneri, M. Iaconelli, L. Bonadonna, L. Lucentini, E. Suffredini, SARS-CoV-2 has been circulating in northern Italy since December 2019: Evidence from environmental monitoring. *Sci. Total Environ.* **750**, 141711 (2021). doi:10.1016/j.scitotenv.2020.141711 Medline

37. A. Amendola, S. Bianchi, M. Gori, D. Colzani, M. Canuti, E. Borghi, M. C. Raviglione, G. V. Zuccotti, E. Tanzi, Evidence of SARS-CoV-2 RNA in an oropharyngeal swab specimen, Milan, Italy, early December 2019. *Emerg. Infect. Dis.* **27**, 648–650 (2021). doi:10.3201/eid2702.204632 Medline

38. E. O. Nsoesie, B. Rader, Y. L. Barnoon, L. Goodwin, J. Brownstein, Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019. (2020); available at https://dash.harvard.edu/handle/1/42669767.

39. J. Peccia, A. Zulli, D. E. Brackney, N. D. Grubaugh, E. H. Kaplan, A. Casanovas-Massana, A. I. Ko, A. A. Malik, D. Wang, M. Wang, J. L. Warren, D. M. Weinberger, W. Arnold,

S. B. Omer, Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nat. Biotechnol.* **38**, 1164–1167 (2020). doi:10.1038/s41587-020-0684-z Medline

40. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N. Loman, J. T. McCrone, S. Gonçalves, D. Jorgensen, R. Myers, V. Hill, D. K. Jackson, K. Gaythorpe, N. Groves, J. Sillitoe, D. P. Kwiatkowski, The COVID-19 Genomics UK (COG-UK) consortium, S. Flaxman, O. Ratman, S. Bhatt, S. Hopkins, A. Gandy, A. Rambaut, N. M. Ferguson, Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data. medRxiv (2021), doi:10.1101/2020.12.30.20249034.

41. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori; Sheffield COVID-19 Genomics Group, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812–827.e19 (2020). doi:10.1016/j.cell.2020.06.043 Medline

42. S. Lytras, J. Hughes, W. Xia, X. Jiang, D. L. Robertson, Exploring the natural origins of SARS-CoV-2. bioRxiv (2021); www.biorxiv.org/content/10.1101/2021.01.22.427830v2.

43. M. F. Boni, P. Lemey, X. Jiang, T. T.-Y. Lam, B. W. Perry, T. A. Castoe, A. Rambaut, D. L. Robertson, Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020). doi:10.1038/s41564-020-0771-4 Medline

44. J. Pekar, J. Wertheim, Data for "Timing the SARS-CoV-2 Index Case in Hubei Province." Data Dryad (2021); https://doi.org/10.5061/dryad.4f4qrfjbm.

45. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* **22**, 30494 (2017). doi:10.2807/1560-7917.ES.2017.22.13.30494 Medline

46. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020). doi:10.1093/molbev/msaa015 Medline

47. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016). doi:10.1093/ve/vew007 Medline

48. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018). doi:10.1093/sysbio/syy032 Medline

49. M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, M. A. Suchard, Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol. Biol. Evol.* **30**, 713–724 (2013). doi:10.1093/molbev/mss265 Medline

50. A.-L. Barabási, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). doi:10.1126/science.286.5439.509 Medline

51. J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, W. J. Edmunds, Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Med.* **5**, e74 (2008). doi:10.1371/journal.pmed.0050074 Medline

52. F. D. Sahneh, A. Vajdi, H. Shakeri, F. Fan, C. Scoglio, GEMFsim: A stochastic simulator for the generalized epidemic modeling framework. *J. Comput. Sci.* **22**, 36–44 (2017). doi:10.1016/j.jocs.2017.08.014

53. O. Ratmann, E. B. Hodcroft, M. Pickles, A. Cori, M. Hall, S. Lycett, C. Colijn, B. Dearlove, X. Didelot, S. Frost, A. S. M. M. Hossain, J. B. Joy, M. Kendall, D. Kühnert, G. E. Leventhal, R. Liang, G. Plazzotta, A. F. Y. Poon, D. A. Rasmussen, T. Stadler, E. Volz, C. Weis, A. J. Leigh Brown, C. Fraser; PANGEA-HIV Consortium, Phylogenetic tools for generalized HIV-1 epidemics: Findings from the PANGEA-HIV Methods Comparison. *Mol. Biol. Evol.* **34**, 185–203 (2017). doi:10.1093/molbev/msw217 Medline

54. N. Moshiri, TreeSwift: A massively scalable Python tree package. *SoftwareX* **11**, 100436 (2020). doi:10.1016/j.softx.2020.100436