ARTICLES
https://doi.org/10.1038/s41564-020-0771-4
nature microbiology

Check for updates

Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic

Maciej F. Boni[1,8]✉, Philippe Lemey[2,8]✉, Xiaowei Jiang[3], Tommy Tsan-Yuk Lam[4], Blair W. Perry[5], Todd A. Castoe[5], Andrew Rambaut[6]✉ and David L. Robertson[7]✉

There are outstanding evolutionary questions on the recent emergence of human coronavirus SARS-CoV-2 including the role of reservoir species, the role of recombination and its time of divergence from animal viruses. We find that the sarbecoviruses—the viral subgenus containing SARS-CoV and SARS-CoV-2—undergo frequent recombination and exhibit spatially structured genetic diversity on a regional scale in China. SARS-CoV-2 itself is not a recombinant of any sarbecoviruses detected to date, and its receptor-binding motif, important for specificity to human ACE2 receptors, appears to be an ancestral trait shared with bat viruses and not one acquired recently via recombination. To employ phylogenetic dating methods, recombinant regions of a 68-genome sarbecovirus alignment were removed with three independent methods. Bayesian evolutionary rate and divergence date estimates were shown to be consistent for these three approaches and for two different prior specifications of evolutionary rates based on HCoV-OC43 and MERS-CoV. Divergence dates between SARS-CoV-2 and the bat sarbecovirus reservoir were estimated as 1948 (95% highest posterior density (HPD): 1879–1999), 1969 (95% HPD: 1930–2000) and 1982 (95% HPD: 1948–2009), indicating that the lineage giving rise to SARS-CoV-2 has been circulating unnoticed in bats for decades.

In December 2019, a cluster of pneumonia cases epidemiologically linked to an open-air live animal market in the city of Wuhan (Hubei Province), China[1,2] led local health officials to issue an epidemiological alert to the Chinese Center for Disease Control and Prevention and the World Health Organization's (WHO) China Country Office. In early January, the aetiological agent of the pneumonia cases was found to be a coronavirus[3], subsequently named SARS-CoV-2 by an International Committee on Taxonomy of Viruses (ICTV) Study Group[4] and also named hCoV-19 by Wu et al.[5]. The first available sequence data[6] placed this novel human pathogen in the *Sarbecovirus* subgenus of *Coronaviridae*[7], the same subgenus as the SARS virus that caused a global outbreak of >8,000 cases in 2002–2003. By mid-January 2020, the virus was spreading widely within Hubei province and by early March SARS-CoV-2 was declared a pandemic[8].

In outbreaks of zoonotic pathogens, identification of the infection source is crucial because this may allow health authorities to separate human populations from the wildlife or domestic animal reservoirs posing the zoonotic risk[9,10]. If stopping an outbreak in its early stages is not possible—as was the case for the COVID-19 epidemic in Hubei—identification of origins and point sources is nevertheless important for containment purposes in other provinces and prevention of future outbreaks. When the first genome sequence of SARS-CoV-2, Wuhan-Hu-1, was released on 10 January 2020 (GMT) on Virological.org by a consortium led by Zhang[6], it enabled immediate analyses of its ancestry. Across a large region of the virus genome, corresponding approximately to ORF1b, it did not cluster with any of the known bat coronaviruses indicating that recombination probably played a role in the evolutionary history of these viruses[5,7]. Subsequently a bat sarbecovirus—RaTG13, sampled from a *Rhinolophus affinis* horseshoe bat in 2013 in Yunnan Province—was reported that clusters with SARS-CoV-2 in almost all genomic regions with approximately 96% genome sequence identity[2]. Zhou et al.[2] concluded from the genetic proximity of SARS-CoV-2 to RaTG13 that a bat origin for the current COVID-19 outbreak is probable. Concurrent evidence also proposed pangolins as a potential intermediate species for SARS-CoV-2 emergence and suggested them as a potential reservoir species[11–13].

Unlike other viruses that have emerged in the past two decades, coronaviruses are highly recombinogenic[14–16]. Influenza viruses reassort[17] but they do not undergo homologous recombination within RNA segments[18,19], meaning that origins questions for influenza outbreaks can always be reduced to origins questions for each of influenza's eight RNA segments. For coronaviruses, however, recombination means that small genomic subregions can have independent origins, identifiable if sufficient sampling has been done in the animal reservoirs that support the endemic circulation, co-infection and recombination that appear to be common. Here, we analyse the evolutionary history of SARS-CoV-2 using available genomic data on sarbecoviruses. We demonstrate that the sarbecoviruses circulating in horseshoe bats have complex recombination histories as reported by others[15,20–26]. Despite the SARS-CoV-2 lineage's acquisition of residues in its Spike (S) protein's receptor-binding domain (RBD) permitting the use of human

[1]Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, PA, USA. [2]Department of Microbiology, Immunology and Transplantation, KU Leuven, Rega Institute, Leuven, Belgium. [3]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China. [4]State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China. [5]Department of Biology, University of Texas Arlington, Arlington, TX, USA. [6]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. [7]MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. [8]These authors contributed equally: Maciej F. Boni, Philippe Lemey. ✉e-mail: mfb9@psu.edu; philippe.lemey@kuleuven.be; a.rambaut@ed.ac.uk; david.l.robertson@glasgow.ac.uk

ACE2 (ref. [27]) receptors and its RBD being genetically closer to a pangolin virus than to RaTG13 (refs. [11–13,22,28])—a signal that suggests recombination—the divergence patterns in the S protein do not show evidence of recombination between the lineage leading to SARS-CoV-2 and known sarbecoviruses. Our results indicate the presence of a single lineage circulating in bats with properties that allowed it to infect human cells, as previously described for bat sarbecoviruses related to the first SARS-CoV lineage[29–31].

To gauge the length of time this lineage has circulated in bats, we estimate the time to the most recent common ancestor (TMRCA) of SARS-CoV-2 and RaTG13. We use three bioinformatic approaches to remove the effects of recombination, and we combine these approaches to identify putative non-recombinant regions that can be used for reliable phylogenetic reconstruction and dating. Collectively our analyses point to bats being the primary reservoir for the SARS-CoV-2 lineage. While it is possible that pangolins, or another hitherto undiscovered species, may have acted as an intermediate host facilitating transmission to humans, current evidence is consistent with the virus having evolved in bats resulting in bat sarbecoviruses that can replicate in the upper respiratory tract of both humans and pangolins[25,32].

Results

Recombination analysis and identification of breakpoint-free genome regions. Among the 68 sequences in the aligned sarbecovirus sequence set, 67 show evidence of mosaicism (all Dunn–Sidak-corrected $P < 4 \times 10^{-4}$ and 3SEQ[14]), indicating involvement in homologous recombination either directly with identifiable parentals or in their deeper shared evolutionary history—that is, due to shared ancestral recombination events. This is evidence for numerous recombination events occurring in the evolutionary history of the sarbecoviruses[22,33]; specifying all past events in their correct temporal order[34] is challenging and not shown here. Figure 1 (top) shows the distribution of all identified breakpoints (using 3SEQ's exhaustive triplet search) by the number of candidate recombinant sequences supporting them. The histogram allows for the identification of non-recombining regions (NRRs) by revealing regions with no breakpoints. Sorting these breakpoint-free regions (BFRs) by length results in two segments >5 kb: an ORF1a subregion spanning nucleotides (nt) 3,625–9,150 and the first half of ORF1b spanning nt 13,291–19,628 (sequence numbering given in Source Data, https://github.com/plemey/SARSCoV2origins). Eight other BFRs <500 nt were identified, and the regions were named BFR A–J in order of length. Of the nine breakpoints defining these ten BFRs, four showed phylogenetic incongruence (PI) signals with bootstrap support >80%, adopting previously published criteria on using a combination of mosaic and PI signals to show evidence of past recombination events[19]. All four of these breakpoints were also identified with the tree-based recombination detection method GARD[35].

The extent of sarbecovirus recombination history can be illustrated by five phylogenetic trees inferred from BFRs or concatenated adjacent BFRs (Fig. 1c). BFRs were concatenated if no phylogenetic incongruence signal could be identified between them. When viewing the last 7 kb of the genome, a clade of viruses from northern China appears to cluster with sequences from southern Chinese provinces but, when inspecting trees from different parts of ORF1ab, the N. China clade is phylogenetically separated from the S. China clade. Individual sequences such as RpShaanxi2011, Guangxi GX2013 and two sequences from Zhejiang Province (CoVZXC21/CoVZC45), as previously shown[22,25], have strong phylogenetic recombination signals because they fall on different evolutionary lineages (with bootstrap support >80%) depending on what region of the genome is being examined.

Despite the high frequency of recombination among bat viruses, the block-like nature of the recombination patterns across the genome permits retrieval of a clean subalignment for phylogenetic analysis. Conservatively, we combined the three BFRs >2 kb identified above into non-recombining region 1 (NRR1). Removal of five sequences that appear to be recombinants and two small subregions of BFR A was necessary to ensure that there were no phylogenetic incongruence signals among or within the three BFRs. Alternatively, combining 3SEQ-inferred breakpoints, GARD-inferred breakpoints and the necessity of PI signals for inferring recombination, we can use the 9.9-kb region spanning nucleotides 11,885–21,753 (NRR2) as a putative non-recombining region; this approach is breakpoint-conservative because it is conservative in identifying breakpoints but not conservative in identifying non-recombining regions. Using a third consensus-based approach for identifying recombinant regions in individual sequences—with six different recombination detection methods in RDP5 (ref. [36])—gives a putative recombination-free alignment that we call non-recombinant alignment 3 (NRA3) (see Methods).

All three approaches to removal of recombinant genomic segments point to a single ancestral lineage for SARS-CoV-2 and RaTG13. Two other bat viruses (CoVZXC21 and CoVZC45) from Zhejiang Province fall on this lineage as recombinants of the RaTG13/SARS-CoV-2 lineage and the clade of Hong Kong bat viruses sampled between 2005 and 2007 (Fig. 1c). Specifically, progenitors of the RaTG13/SARS-CoV-2 lineage appear to have recombined with the Hong Kong clade (with inferred breakpoints at 11.9 and 20.8 kb) to form the CoVZXC21/CoVZC45-lineage. Sibling lineages to RaTG13/SARS-CoV-2 include a pangolin sequence sampled in Guangdong Province in March 2019 and a clade of pangolin sequences from Guangxi Province sampled in 2017.

Because the SARS-CoV-2 S protein has been implicated in past recombination events or possibly convergent evolution[12], we specifically investigated several subregions of the S protein—the N-terminal domain of S1, the C-terminal domain of S1, the variable-loop region of the C-terminal domain, and S2. The variable-loop region in SARS-CoV-2 shows closer identity to the 2019 pangolin coronavirus sequence than to the RaTG13 bat virus, supported by phylogenetic inference (Fig. 2). On first examination this would suggest that that SARS-CoV-2 is a recombinant of an ancestor of Pangolin-2019 and RaTG13, as proposed by others[11,22]. However, on closer inspection, the relative divergences in the phylogenetic tree (Fig. 2, bottom) show that SARS-CoV-2 is unlikely to have acquired the variable loop from an ancestor of Pangolin-2019 because these two sequences are approximately 10–15% divergent throughout the entire S protein (excluding the N-terminal domain). It is RaTG13 that is more divergent in the variable-loop region (Extended Data Fig. 1) and thus likely to be the product of recombination, acquiring a divergent variable loop from a hitherto unsampled bat sarbecovirus[28]. This is notable because the variable-loop region contains the six key contact residues in the RBD that give SARS-CoV-2 its ACE2-binding specificity[27,37]. These residues are also in the Pangolin Guangdong 2019 sequence. The most parsimonious explanation for these shared ACE2-specific residues is that they were present in the common ancestors of SARS-CoV-2, RaTG13 and Pangolin Guangdong 2019, and were lost through recombination in the lineage leading to RaTG13. This provides compelling support for the SARS-CoV-2 lineage being the consequence of a direct or nearly-direct zoonotic jump from bats, because the key ACE2-binding residues were present in viruses circulating in bats.

Ancestry in non-recombinant regions. Using the most conservative approach to identification of a non-recombinant genomic region (NRR1), SARS-CoV-2 forms a sister lineage with RaTG13, with genetically related cousin lineages of coronavirus sampled in pangolins in Guangdong and Guangxi provinces (Fig. 3). Given that these pangolin viruses are ancestral to the progenitor of the

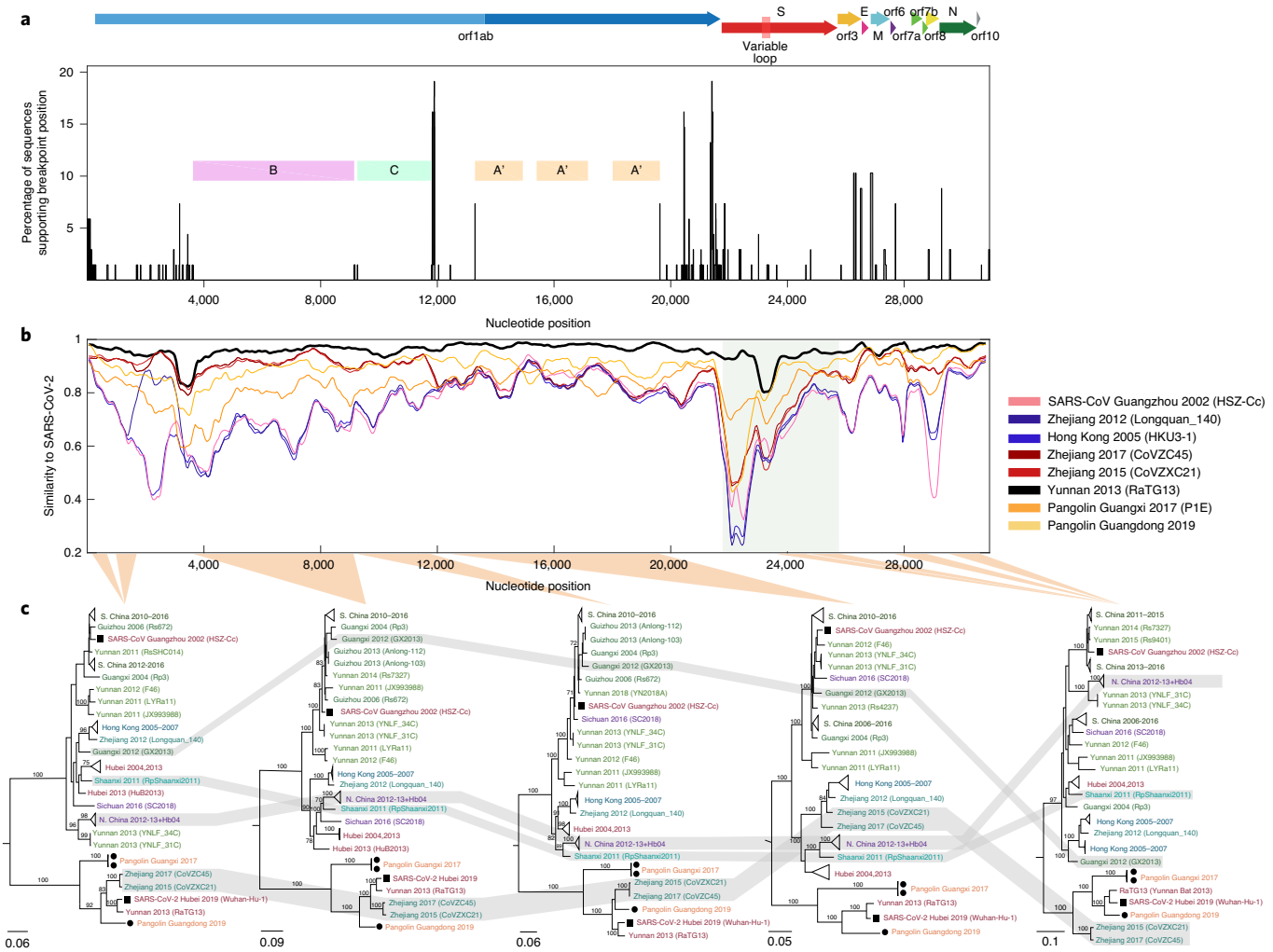


Fig. 1 | Breakpoints identified by 3SEQ. a, Breakpoints identified by 3SEQ illustrated by percentage of sequences (out of 68) that support a particular breakpoint position. Note that breakpoints can be shared between sequences if they are descendants of the same recombination events. Pink, green and orange bars show BFRs, with region A (nt 13,291–19,628) showing two trimmed segments yielding region A′ (nt 13,291–14,932, 15,405–17,162, 18,009–19,628). Regions B and C span nt 3,625–9,150 and 9,261–11,795, respectively. Concatenated region A′BC is NRR1. Open reading frames are shown above the breakpoint plot, with the variable-loop region indicated in the S protein. **b**, Similarity plot between SARS-CoV-2 and several selected sequences including RaTG13 (black), SARS-CoV (pink) and two pangolin sequences (orange). The shaded region corresponds to the S protein. **c**, Maximum likelihood phylogenetic trees rooted on a 2007 virus sampled in Kenya (BtKy72; root truncated from images), shown for five BFRs of the sarbecovirus alignment. Nucleotide positions for phylogenetic inference are 147–695, 962–1,686 (first tree), 3,625–9,150 (second tree, also BFR B), 9,261–11,795 (third tree, also BFR C), 12,443–19,638 (fourth tree) and 23,631–24,633, 24,795–25,847, 27,702–28,843 and 29,574–30,650 (fifth tree). Relevant bootstrap values are shown on branches, and grey-shaded regions show sequences exhibiting phylogenetic incongruence along the genome. S. China corresponds to Guangxi, Yunnan, Guizhou and Guangdong provinces. N. China corresponds to Jilin, Shanxi, Hebei and Henan provinces, and the N. China clade also includes one sequence sampled in Hubei Province in 2004.

RaTG13/SARS-CoV-2 lineage, it is more likely that they are also acquiring viruses from bats. While pangolins could be acting as intermediate hosts for bat viruses to get into humans—they develop severe respiratory disease[38] and commonly come into contact with people through trafficking—there is no evidence that pangolin infection is a requirement for bat viruses to cross into humans.

Phylogenies of subregions of NRR1 depict an appreciable degree of spatial structuring of the bat sarbecovirus population across different regions (Fig. 3). One geographic clade includes viruses from provinces in southern China (Guangxi, Yunnan, Guizhou and Guangdong), with its major sister clade consisting of viruses from provinces in northern China (Shanxi, Henan, Hebei and Jilin) as well as Hubei Province in central China and Shaanxi Province in northwestern China. Several of the recombinant sequences in these trees show that recombination events do occur across geographically divergent clades. The Sichuan (SC2018) virus appears to be a recombinant of northern/central and southern viruses, while the two Zhejiang viruses (CoVZXC21 and CoVZC45) appear to carry a recombinant region from southern or central China.

TMRCA for NRRs of SARS-CoV-2 lineage. To avoid artefacts due to recombination, we focused on NRR1 and NRR2 and the recombination-masked alignment NRA3 to infer time-measured evolutionary histories. Visual exploration using TempEst[39] indicates that there is no evidence for temporal signal in these datasets (Extended Data Fig. 2). This is not surprising for diverse viral populations with relatively deep evolutionary histories. In such cases, even moderate rate variation among long, deep phylogenetic branches will substantially impact expected root-to-tip divergences over a sampling time range that represents only a small fraction of

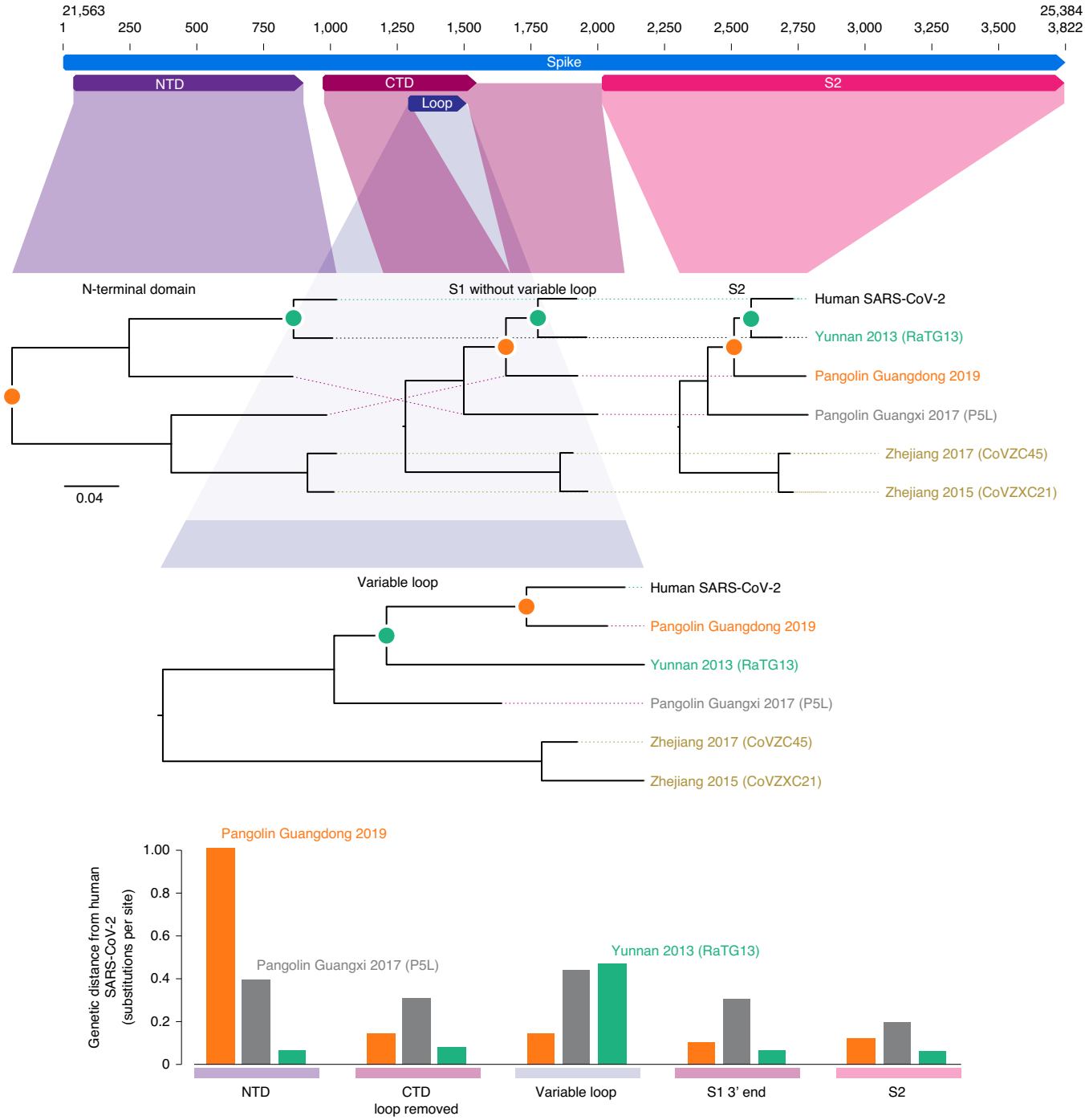


Fig. 2 | Phylogenetic relationships among SARS-CoV-2 and closely related sequences for subregions of the S protein. SARS-CoV-2 and RaTG13 are the most closely related (their most recent common ancestor nodes denoted by green circles), except in the 222-nt variable-loop region of the C-terminal domain (bar graphs at bottom). In the variable-loop region, RaTG13 diverges considerably with the TMRCA, now outside that of SARS-CoV-2 and the Pangolin Guangdong 2019 ancestor, suggesting that RaTG13 has acquired this region from a more divergent and undetected bat lineage. The genetic distances between SARS-CoV-2 and RaTG13 (bottom) demonstrate that their relationship is consistent across all regions except for the variable loop. The genetic distances between SARS-CoV-2 and Pangolin Guangdong 2019 are consistent across all regions except the N-terminal domain, implying that a recombination event between these two sequences in this region is unlikely. Uncertainty measures are shown in Extended Data Fig. 1. NTD, N-terminal domain; CTD, C-terminal domain.

the evolutionary history[40]. However, formal testing using marginal likelihood estimation[41] does provide some evidence of a temporal signal, albeit with limited log Bayes factor support of 3 (NRR1), 10 (NRR2) and 3 (NRA3); see Supplementary Table 1.

In the absence of a strong temporal signal, we sought to identify a suitable prior rate distribution to calibrate the time-measured trees by examining several coronaviruses sampled over time, including HCoV-OC43, MERS-CoV, and SARS-CoV virus genomes. These datasets were subjected to the same recombination masking approach as NRA3 and were characterized by a strong temporal signal (Fig. 4), but also by markedly different evolutionary rates. Specifically, using a formal Bayesian approach[42] (see Methods), we

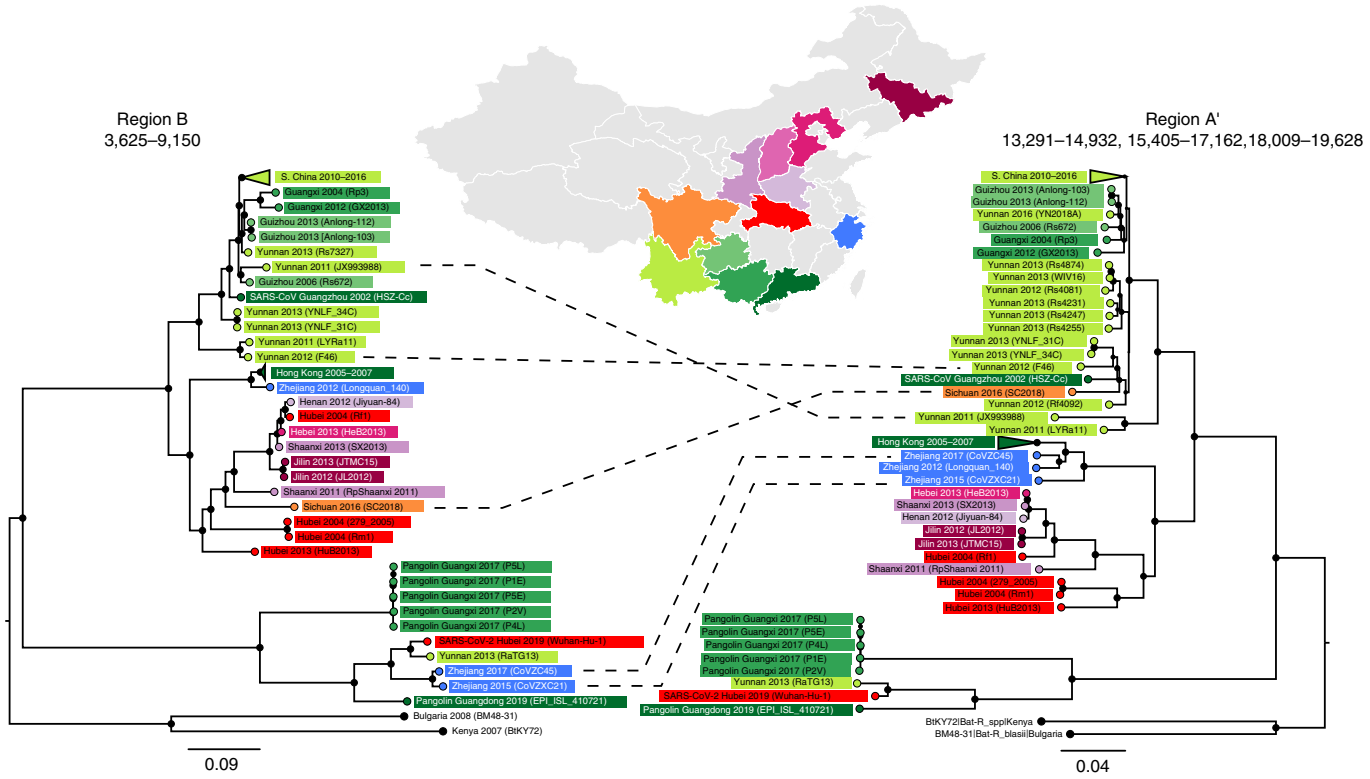


Fig. 3 | Maximum likelihood trees of the sarbecoviruses using the two longest BFRs, rooted on the Kenya/Bulgaria lineage. Region A has been shortened to A′ (5,017 nt) based on potential recombination signals within the region. Region B is 5,525 nt long. Sequences are colour-coded by province according to the map. Five example sequences with incongruent phylogenetic positions in the two trees are indicated by dashed lines.

estimate a fast evolutionary rate (0.00169 substitutions per site yr^{-1}, 95% highest posterior density (HPD) interval (0.00131,0.00205)) for SARS viruses sampled over a limited timescale (1 year), a slower rate (0.00078 (0.00063,0.00092) substitutions per site yr^{-1}) for MERS-CoV on a timescale of about 4 years and the slowest rate (0.00024 (0.00019,0.00029) substitutions per site yr^{-1}) for HCoV-OC43 over almost five decades. These differences reflect the fact that rate estimates can vary considerably with the timescale of measurement, a frequently observed phenomenon in viruses known as time-dependent evolutionary rates[41,43,44]. Over relatively shallow timescales, such differences can primarily be explained by varying selective pressure, with mildly deleterious variants being eliminated more strongly by purifying selection over longer timescales[44–46]. Consistent with this, we estimate a concomitantly decreasing non-synonymous-to-synonymous substitution rate ratio over longer evolutionary timescales: 1.41 (1.20,1.68), 0.35 (0.30,0.41) and 0.133 (0.129,0.136) for SARS, MERS-CoV and HCoV-OC43, respectively. In light of these time-dependent evolutionary rate dynamics, a slower rate is appropriate for calibration of the sarbecovirus evolutionary history. We compare both MERS-CoV- and HCoV-OC43-centred prior distributions (Extended Data Fig. 3) to examine the sensitivity of date estimates to this prior specification.

We infer time-measured evolutionary histories using a Bayesian phylogenetic approach while incorporating rate priors based on mean MERS-CoV and HCoV-OC43 rates and with standard deviations that allow for more uncertainty than the empirical estimates for both viruses (see Methods). Using both prior distributions, this results in six highly similar posterior rate estimates for NRR1, NRR2 and NRA3, centred around 0.00055 substitutions per site yr^{-1}. The fact that these estimates lie between the rates for MERS-CoV and HCoV-OC43 is consistent with the intermediate sampling time range of about 18 years (Fig. 5). The consistency of the posterior rates for the different prior means also implies that the data do contribute to the evolutionary rate estimate, despite the fact that a temporal signal was visually not apparent (Extended Data Fig. 2). Below, we report divergence time estimates based on the HCoV-OC43-centred rate prior for NRR1, NRR2 and NRA3 and summarize corresponding estimates for the MERS-CoV-centred rate priors in Extended Data Fig. 4. Divergence time estimates based on the HCoV-OC43-centred rate prior for the separate BFRs (Supplementary Table 3) show consistency in TMRCA estimates across the genome.

The divergence time estimates for SARS-CoV-2 and SARS-CoV from their respective most closely related bat lineages are reasonably consistent among the three approaches we use to eliminate the effects of recombination in the alignment. Using the most conservative approach (NRR1), the divergence time estimate for SARS-CoV-2 and RaTG13 is 1969 (95% HPD: 1930–2000), while that between SARS-CoV and its most closely related bat sequence is 1962 (95% HPD: 1932–1988); see Fig. 5. These are in general agreement with estimates using NRR2 and NRA3, which result in divergence times of 1982 (1948–2009) and 1948 (1879–1999), respectively, for SARS-CoV-2, and estimates of 1952 (1906–1989) and 1970 (1932–1996), respectively, for the divergence time of SARS-CoV from its closest known bat relative. The SARS-CoV divergence times are somewhat earlier than dates previously estimated[15] because previous estimates were obtained using a collection of SARS-CoV genomes from human and civet hosts (as well as a few closely related bat genomes), which implies that evolutionary rates were predominantly informed by the short-term SARS outbreak scale and probably biased upwards. Indeed, the rates reported by these studies are in line with the short-term SARS rates that we estimate (Fig. 4). The estimated divergence times for the pangolin virus most closely related to the SARS-CoV-2/RaTG13 lineage

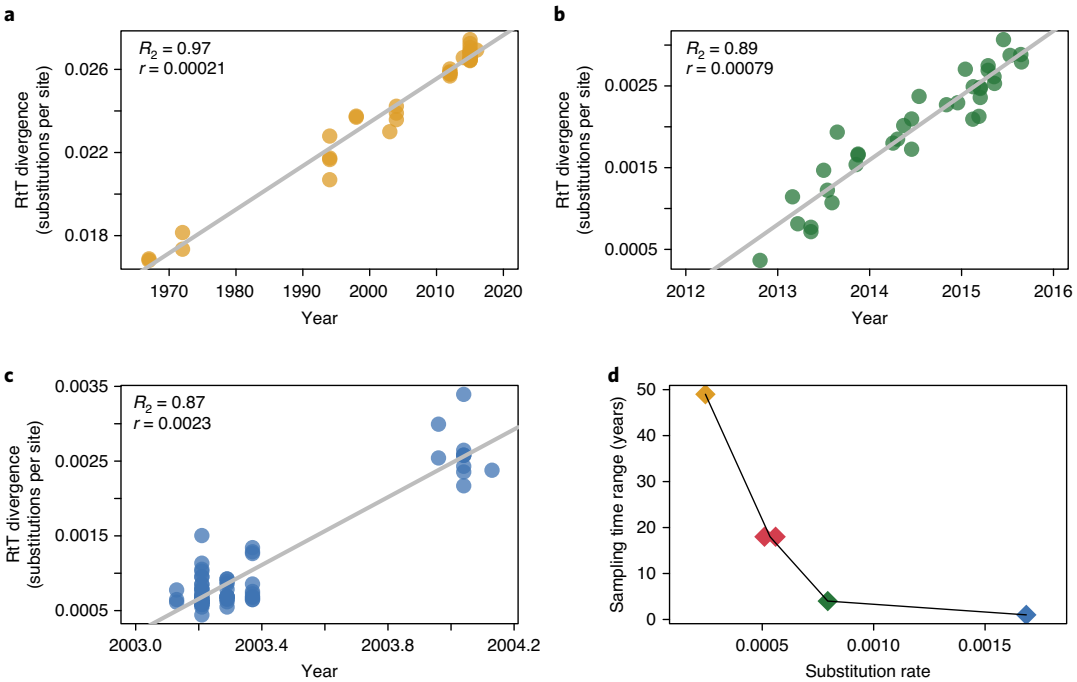


Fig. 4 | Temporal signal and mean evolutionary rate estimates for coronaviruses HCoV-OC43, MERS and SARS. a–c, Root-to-tip (RtT) divergence as a function of sampling time for the three coronavirus evolutionary histories unfolding over different timescales (HCoV-OC43 (n = 37; **a**) MERS (n = 35; **b**) and SARS (n = 69; **c**)). Decimal years are shown on the x axis for the 1.2 years of SARS sampling in **c**. **d**, Mean evolutionary rate estimates plotted against sampling time range for the same three datasets (represented by the same colour as the data points in their respective RtT divergence plots), as well as for the comparable NRA3 using the two different priors for the rate in the Bayesian inference (red points).

range from 1851 (1730–1958) to 1877 (1746–1986), indicating that these pangolin lineages were acquired from bat viruses divergent to those that gave rise to SARS-CoV-2. Current sampling of pangolins does not implicate them as an intermediate host.

Discussion

Identifying the origins of an emerging pathogen can be critical during the early stages of an outbreak, because it may allow for containment measures to be precisely targeted at a stage when the number of daily new infections is still low. Early detection via genomics was not possible during Southeast Asia's initial outbreaks of avian influenza H5N1 (1997 and 2003–2004) or the first SARS outbreak (2002–2003). By 2009, however, rapid genomic analysis had become a routine component of outbreak response. The 2009 influenza pandemic and subsequent outbreaks of MERS-CoV (2012), H7N9 avian influenza (2013), Ebola virus (2014) and Zika virus (2015) were met with rapid sequencing and genomic characterization. For the current pandemic, the 'novel pathogen identification' component of outbreak response delivered on its promise, with viral identification and rapid genomic analysis providing a genome sequence and confirmation, within weeks, that the December 2019 outbreak first detected in Wuhan, China was caused by a coronavirus[3]. Unfortunately, a response that would achieve containment was not possible. Given what was known about the origins of SARS, as well as identification of SARS-like viruses circulating in bats that had binding sites adapted to human receptors[29–31], appropriate measures should have been in place for immediate control of outbreaks of novel coronaviruses. The key to successful surveillance is knowing which viruses to look for and prioritizing those that can readily infect humans[47].

The difficulty in inferring reliable evolutionary histories for coronaviruses is that their high recombination rate[48,49] violates the assumption of standard phylogenetic approaches because different parts of the genome have different histories. To begin characterizing any ancestral relationships for SARS-CoV-2, NRRs of the genome must be identified so that reliable phylogenetic reconstruction and dating can be performed. Evolutionary rate estimation can be profoundly affected by the presence of recombination[50]. Because there is no single accepted method of inferring breakpoints and identifying clean subregions with high certainty, we implemented several approaches to identifying three classic statistical signals of recombination: mosaicism, phylogenetic incongruence and excessive homoplasy[51]. Our most conservative approach attempted to ensure that putative NRRs had no mosaic or phylogenetic incongruence signals. A second breakpoint-conservative approach was conservative with respect to breakpoint identification, but this means that it is accepting of false-negative outcomes in breakpoint inference, resulting in less certainty that a putative NRR truly contains no breakpoints. A third approach attempted to minimize the number of regions removed while also minimizing signals of mosaicism and homoplasy. The origins we present in Fig. 5 (NRR1) are conservative in the sense that NRR1 is more likely to be non-recombinant than NRR2 or NRA3. Because the estimated rates and divergence dates were highly similar in the three datasets analysed, we conclude that our estimates are robust to the method of identifying a genome's NRRs.

Due to the absence of temporal signal in the sarbecovirus datasets, we used informative prior distributions on the evolutionary rate to estimate divergence dates. Calibration of priors can be performed using other coronaviruses (SARS-CoV, MERS-CoV and HCoV-OC43), but estimated rates vary with the timescale of sample collection. In the presence of time-dependent rate variation, a widely observed phenomenon for viruses[43,44,52], slower prior rates appear more appropriate for sarbecoviruses that currently encompass a sampling time range of about 18 years. Our approach resulted in similar posterior rates using two different prior means, implying

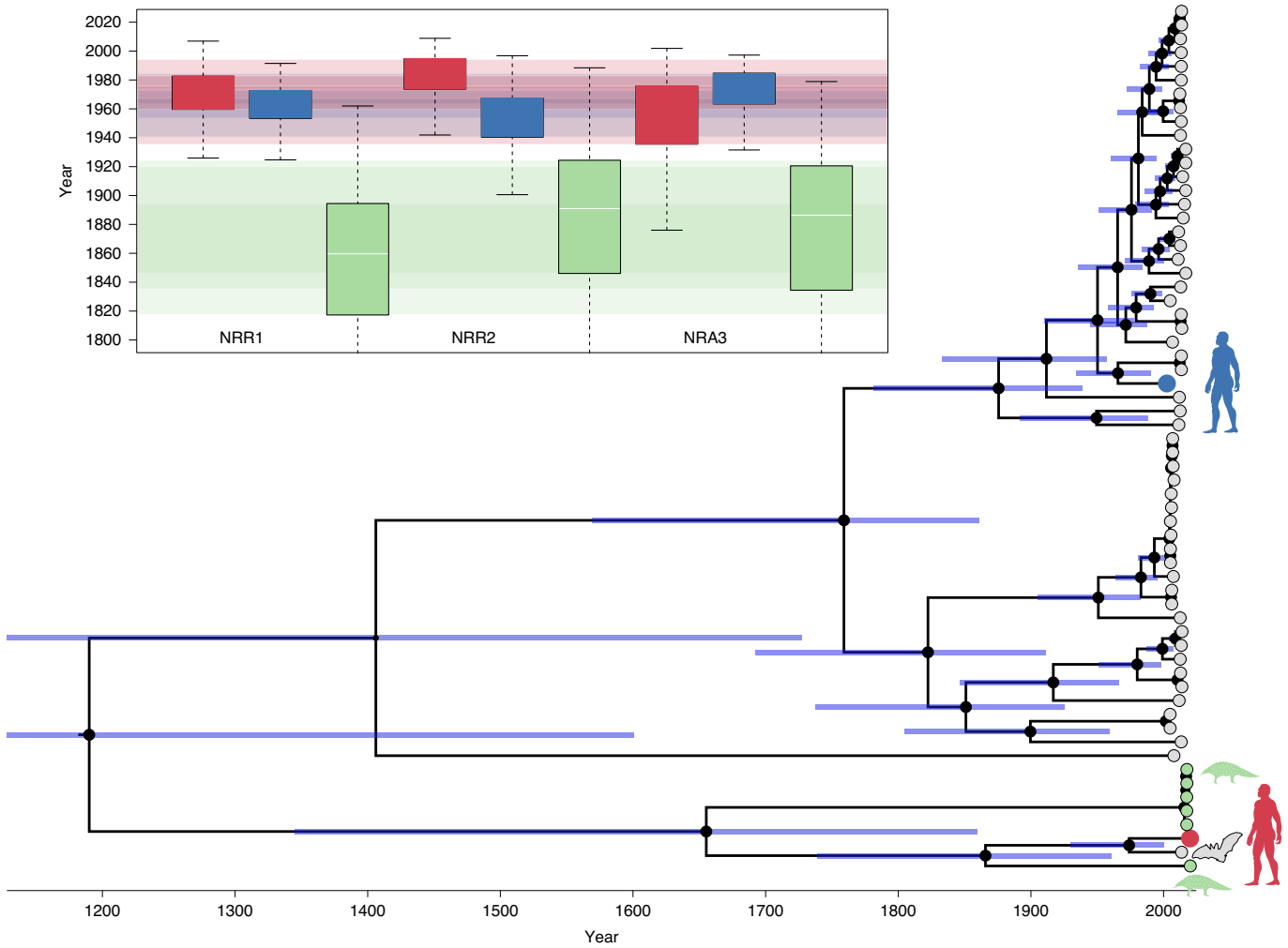


Fig. 5 | Time-measured phylogenetic estimates and divergence times for sarbecovirus lineages using an HCoV-OC43-centred rate prior. The time-calibrated phylogeny represents a maximum clade credibility tree inferred for NRR1. Grey tips correspond to bat viruses, green to pangolin, blue to SARS-CoV and red to SARS-CoV-2. The sizes of the black internal node circles are proportional to the posterior node support. 95% credible interval bars are shown for all internal node ages. The inset represents divergence time estimates based on NRR1, NRR2 and NRA3. The boxplots show divergence time estimates (posterior medians) for SARS-CoV-2 (red) and the 2002–2003 SARS-CoV virus (blue) from their most closely related bat virus. Green boxplots show the TMRCA estimate for the RaTG13/SARS-CoV-2 lineage and its most closely related pangolin lineage (Guangdong 2019). Boxplots show interquartile ranges, white lines are medians and box whiskers show the full range of posterior distribution. Transparent bands of interquartile range width and with the same colours are superimposed to highlight the overlap between estimates. In Extended Data Fig. 4 we compare these divergence time estimates to those obtained using the MERS-CoV-centred rate priors for NRR1, NRR2 and NRA3.

that the sarbecovirus data do inform the rate estimate even though a root-to-tip temporal signal was not apparent.

The relatively fast evolutionary rate means that it is most appropriate to estimate shallow nodes in the sarbecovirus evolutionary history. Accurate estimation of ages for deeper nodes would require adequate accommodation of time-dependent rate variation. While such models have recently been made available, we lack the information to calibrate the rate decline over time (for example, through internal node calibrations[44]). As a proxy, it would be possible to model the long-term purifying selection dynamics as a major source of time-dependent rates[43,44,52], but this is beyond the scope of the current study. The assumption of long-term purifying selection would imply that coronaviruses are in endemic equilibrium with their natural host species, horseshoe bats, to which they are presumably well adapted. While there is evidence of positive selection in the sarbecovirus lineage leading to RaTG13/SARS-CoV-2 (ref. [53]), this is inferred to have occurred before the divergence of RaTG13 and SARS-CoV-2 and thus should not influence our inferences.

Of importance for future spillover events is the appreciation that SARS-CoV-2 has emerged from the same horseshoe bat subgenus that harbours SARS-like coronaviruses. Another similarity between SARS-CoV and SARS-CoV-2 is their divergence time (40–70 years ago) from currently known extant bat virus lineages (Fig. 5). This long divergence period suggests there are unsampled virus lineages circulating in horseshoe bats that have zoonotic potential due to the ancestral position of the human-adapted contact residues in the SARS-CoV-2 RBD. Without better sampling, however, it is impossible to estimate whether or how many of these additional lineages exist. While there is involvement of other mammalian species—specifically pangolins for SARS-CoV-2—as a plausible conduit for transmission to humans, there is no evidence that pangolins are facilitating adaptation to humans. A hypothesis of snakes as intermediate hosts of SARS-CoV-2 was posited during the early epidemic phase[54], but we found no evidence of this[55,56]; see Extended Data Fig. 5.

With horseshoe bats currently the most plausible origin of SARS-CoV-2, it is important to consider that sarbecoviruses cir-

culate in a variety of horseshoe bat species with widely overlapping species ranges[57]. Nevertheless, the viral population is largely spatially structured according to provinces in the south and southeast on one lineage, and provinces in the centre, east and northeast on another (Fig. 3). This boundary appears to be rarely crossed. Two exceptions can be seen in the relatively close relationship of Hong Kong viruses to those from Zhejiang Province (with two of the latter, CoVZC45 and CoVZXC21, identified as recombinants) and a recombinant virus from Sichuan for which part of the genome (region B of SC2018 in Fig. 3) clusters with viruses from provinces in the centre, east and northeast of China. SARS-CoV-2 and RaTG13 are also exceptions because they were sampled from Hubei and Yunnan, respectively. The fact that they are geographically relatively distant is in agreement with their somewhat distant TMRCA, because the spatial structure suggests that migration between their locations may be uncommon. From this perspective, it may be useful to perform surveillance for more closely related viruses to SARS-CoV-2 along the gradient from Yunnan to Hubei.

It is clear from our analysis that viruses closely related to SARS-CoV-2 have been circulating in horseshoe bats for many decades. The unsampled diversity descended from the SARS-CoV-2/RaTG13 common ancestor forms a clade of bat sarbecoviruses with generalist properties—with respect to their ability to infect a range of mammalian cells—that facilitated its jump to humans and may do so again. Although the human ACE2-compatible RBD was very likely to have been present in a bat sarbecovirus lineage that ultimately led to SARS-CoV-2, this RBD sequence has hitherto been found in only a few pangolin viruses. Furthermore, the other key feature thought to be instrumental in the ability of SARS-CoV-2 to infect humans—a polybasic cleavage site insertion in the S protein—has not yet been seen in another close bat relative of the SARS-CoV-2 virus.

The existing diversity and dynamic process of recombination amongst lineages in the bat reservoir demonstrate how difficult it will be to identify viruses with potential to cause major human outbreaks before they emerge. This underscores the need for a global network of real-time human disease surveillance systems, such as that which identified the unusual cluster of pneumonia in Wuhan in December 2019, with the capacity to rapidly deploy genomic tools and functional studies for pathogen identification and characterization.

Methods

Dataset compilation. *Sarbecovirus data.* Complete genome sequence data were downloaded from GenBank and ViPR; accession numbers of all 68 sequences are available in Supplementary Table 4. Sequences were aligned by MAFFT[58] v.7.310, with a final alignment length of 30,927, and used in the analyses below.

HCoV-OC43. We compiled a dataset including 27 human coronavirus OC43 virus genomes and ten related animal virus genomes (six bovine, three white-tailed deer and one canine virus). The canine viral genome was excluded from the Bayesian phylogenetic analyses because temporal signal analyses (see below) indicated that it was an outlier.

MERS-CoV. We extracted a similar number ($n = 35$) of genomes from a MERS-CoV dataset analysed by Dudas et al.[59] using the phylogenetic diversity analyser tool[60] (v.0.5).

SARS-CoV. We compiled a set of 69 SARS-CoV genomes including 58 sampled from humans and 11 sampled from civets and raccoon dogs. This dataset comprises an updated version of that used in Hon et al.[15] and includes a cluster of genomes sampled in late 2003 and early 2004, but the evolutionary rate estimate without this cluster (0.00175 substitutions per site yr^{-1} (0.00117,0.00229)) is consistent with the complete dataset (0.00169 substitutions per site yr^{-1}, (0.00131,0.00205)).

Sarbecovirus, HCoV-OC43 and SARS-CoV data were assembled from GenBank to be as complete as possible, with sampling year as an inclusion criterion. MERS-CoV data were subsampled to match sample sizes with SARS-CoV and HCoV-OC43.

Recombination analysis. Because coronaviruses are known to be highly recombinant, we used three different approaches to identify non-recombinant regions for use in our Bayesian time-calibrated phylogenetic inference.

First, we took an approach that relies on identification of mosaic regions (via 3SEQ[14] v.1.7) that are also supported by PI signals[19]. Because 3SEQ is the most statistically powerful of the mosaic methods[61], we used it to identify the best-supported breakpoint history for each potential child (recombinant) sequence in the dataset. A single 3SEQ run on the genome alignment resulted in 67 out of 68 sequences supporting some recombination in the past, with multiple candidate breakpoint ranges listed for each putative recombinant. Next, we (1) collected all breakpoints into a single set, (2) complemented this set to generate a set of non-breakpoints, (3) grouped non-breakpoints into contiguous BFRs and (4) sorted these regions by length. A phylogenetic tree—using RAxML v8.2.8 (ref. [62,63]), the GTR + Γ model and 100 bootstrap replicates—was inferred for each BFR >500 nt.

We considered (1) the possibility that BFRs could be combined into larger non-recombinant regions and (2) the possibility of further recombination within each BFR.

We named the length-sorted BFRs as: BFR A (nt positions 13,291–19,628, length = 6,338 nt), BFR B (nt positions 3,625–9,150, length = 5,526 nt), BFR C (nt positions 9,261–11,795, length = 2,535 nt), BFR D (nt positions 27,702–28,843, length = 1,142 nt) and six further regions (E–J). Phylogenetic trees and exact breakpoints for all ten BFRs are shown in Supplementary Figs. 1–10. Regions A–C had similar phylogenetic relationships among the southern China bat viruses (Yunnan, Guangxi and Guizhou provinces), the Hong Kong viruses, northern Chinese viruses (Jilin, Shanxi, Hebei and Henan provinces, including Shaanxi), pangolin viruses and the SARS-CoV-2 lineage. Because these subclades had different phylogenetic relationships in region D (Supplementary Fig. 4), that region and shorter BFRs were not included in combined putative non-recombinant regions.

Regions A–C were further examined for mosaic signals by 3SEQ, and all showed signs of mosaicism. In region A, we removed subregion A1 (nt positions 3,872–4,716 within region A) and subregion A4 (nt 1,642–2,113) because both showed PI signals with other subregions of region A. After removal of A1 and A4, we named the new region A′. In addition, sequences NC_014470 (Bulgaria 2008), CoVZXC21, CoVZC45 and DQ412042 (Hubei-Yichang) needed to be removed to maintain a clean non-recombinant signal in A′. Region B showed no PI signals within the region, except one including sequence SC2018 (Sichuan), and thus this sequence was also removed from the set. Region C showed no PI signals within it. Combining regions A′, B and C and removing the five named sequences gives us putative NRR1, as an alignment of 63 sequences. We say that this approach is conservative because sequences and subregions generating recombination signals have been removed, and BFRs were concatenated only when no PI signals could be detected between them. The construction of NRR1 is the most conservative as it is least likely to contain any remaining recombination signals.

In our second stage, we wanted to construct non-recombinant regions where our approach to breakpoint identification was as conservative as possible. We call this approach breakpoint-conservative, but note that this has the opposite effect to the construction of NRR1 in that this approach is the most likely to allow breakpoints to remain inside putative non-recombining regions. In other words, a true breakpoint is less likely to be called as such (this is breakpoint-conservative), and thus the construction of a non-recombining region may contain true recombination breakpoints (with insufficient evidence to call them as such). In this approach, we considered a breakpoint as supported only if it had three types of statistical support: from (1) mosaic signals identified by 3SEQ, (2) PI signals identified by building trees around 3SEQ's breakpoints and (3) the GARD algorithm[35], which identifies breakpoints by identifying PI signals across proposed breakpoints. Because 3SEQ identified ten BFRs >500 nt, we used GARD's (v.2.5.0) inference on 10, 11 and 12 breakpoints. A reduced sequence set of 25 sequences chosen to capture the breadth of diversity in the sarbecoviruses (obvious recombinants not involving the SARS-CoV-2 lineage were also excluded) was used because GARD is computationally intensive. GARD identified eight breakpoints that were also within 50 nt of those identified by 3SEQ. PI signals were identified (with bootstrap support >80%) for seven of these eight breakpoints: positions 1,684, 3,046, 9,237, 11,885, 21,753, 22,773 and 24,628. Using these breakpoints, the longest putative non-recombining segment (nt 1,885–21,753) is 9.9 kb long, and we call this region NRR2.

Our third approach involved identifying breakpoints and masking minor recombinant regions (with gaps, which are treated as unobserved characters in probabilistic phylogenetic approaches). Specifically, we used a combination of six methods implemented in v.5.5 of RDP5 (ref. [36]) (RDP, GENECONV, MaxChi, Bootscan, SisScan and 3SEQ) and considered recombination signals detected by more than two methods for breakpoint identification. Except for specifying that sequences are linear, all settings were kept to their defaults. Based on the identified breakpoints in each genome, only the major non-recombinant region is kept in each genome while other regions are masked. To evaluate the performance procedure, we confirmed that the recombination masking resulted in (1) a markedly different outcome of the PHI test[64], (2) removal of well-supported (bootstrap value >95%) incompatible splits in Neighbor-Net[65] and (3) a

near-complete reduction of mosaic signal as identified by 3SEQ. If the latter still identified non-negligible recombination signal, we removed additional genomes that were identified as major contributors to the remaining signal. This produced non-recombining alignment NRA3, which included 63 of the 68 genomes.

Bayesian divergence time estimation. We focused on these three non-recombining regions/alignments for divergence time estimation; this avoids inappropriate modelling of evolutionary processes with recombination on strictly bifurcating trees, which can result in different artefacts such as homoplasies that inflate branch lengths and lead to apparently longer evolutionary divergence times. To examine temporal signal in the sequenced data, we plotted root-to-tip divergence against sampling time using TempEst[39] v.1.5.3 based on a maximum likelihood tree. The latter was reconstructed using IQTREE[66] v.2.0 under a general time-reversible (GTR) model with a discrete gamma distribution to model inter-site rate variation.

Time-measured phylogenetic reconstruction was performed using a Bayesian approach implemented in BEAST[42] v.1.10.4. When the genomic data included both coding and non-coding regions we used a single GTR + Γ substitution model; for concatenated coding genes we partitioned the alignment by codon position and specified an independent GTR + Γ model for each partition with a separate gamma model to accommodate inter-site rate variation. We used an uncorrelated relaxed clock model with log-normal distribution for all datasets, except for the low-diversity SARS data for which we specified a strict molecular clock model. For the HCoV-OC43, MERS-CoV and SARS datasets we specified flexible skygrid coalescent tree priors. In the absence of any reasonable prior knowledge on the TMRCA of the sarbecovirus datasets (which is required for grid specification in a skygrid model), we specified a simpler constant size population prior. As informative rate priors for the analysis of the sarbecovirus datasets, we used two different normal prior distributions: one with a mean of 0.00078 and s.d. = 0.00075 and one with a mean of 0.00024 and s.d. = 0.00025. These means are based on the mean rates estimated for MERS-CoV and HCoV-OC43, respectively, while the standard deviations are set ten times higher than empirical values to allow greater prior uncertainty and avoid strong bias (Extended Data Fig. 3). In our analyses of the sarbecovirus datasets, we incorporated the uncertainty of the sampling dates when exact dates were not available. To estimate non-synonymous over synonymous rate ratios for the concatenated coding genes, we used the empirical Bayes Renaissance counting'procedure[67]. Temporal signal was tested using a recently developed marginal likelihood estimation procedure[41] (Supplementary Table 1).

Posterior distributions were approximated through Markov chain Monte Carlo sampling, which were run sufficiently long to ensure effective sampling sizes >100. BEAST inferences made use of the BEAGLE v.3 library[68] for efficient likelihood computations. We used TreeAnnotator to summarize posterior tree distributions and annotated the estimated values to a maximum clade credibility tree, which was visualized using FigTree.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequence data analysed in this manuscript are available at https://github.com/plemey/SARSCoV2origins. Note that six of these sequences fall under the terms of use of the GISAID platform.

Code availability

All custom code used in the manuscript is available at https://github.com/plemey/SARSCoV2origins.

Received: 13 April 2020; Accepted: 10 July 2020;
Published online: 28 July 2020

References

1. Li, Q. et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
2. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
3. *Novel Coronavirus (2019-nCoV) Situation Report 1, 21 January 2020* (World Health Organization, 2020).
4. Gorbalenya, A. E. et al. The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).
5. Wu, Y. et al. SARS-CoV-2 is an appropriate name for the new coronavirus. A distinct name is needed for the new coronavirus. *Lancet* **395**, 949–950 (2020).
6. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
7. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
8. *Coronavirus Disease 2019 (COVID-19) Situation Report – 51* (World Health Organization, 2020).
9. Yu, H. et al. Effect of closure of live poultry markets on poultry-to-person transmission of avian influenza A H7N9 virus: an ecological study. *Lancet* **383**, 541–548 (2013).
10. Stegeman, A. et al. Avian influenza a virus (H7N7) epidemic in The Netherlands in 2003: course of the epidemic and effectiveness of control measures. *J. Infect. Dis.* **190**, 2088–2095 (2004).
11. Xiao, K. et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286–289 (2020).
12. Lam, T. T. et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285 (2020).
13. Liu, P. et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog.* **16**, e1008421 (2020).
14. Lam, H. M., Ratmann, O. & Boni, M. F. Improved algorithmic complexity for the 3SEQ recombination detection algorithm. *Mol. Biol. Evol.* **35**, 247–251 (2018).
15. Hon, C. et al. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J. Virol.* **82**, 1819–1826 (2008).
16. Forni, D., Cagliani, R., Clerici, M. & Sironi, M. Molecular evolution of human coronavirus genomes. *Trends Microbiol.* **25**, 35–48 (2017).
17. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179 (1992).
18. Boni, M. F., Zhou, Y., Taubenberger, J. K. & Holmes, E. C. Homologous recombination is very rare or absent in human influenza A virus. *J. Virol.* **82**, 4807–4811 (2008).
19. Boni, M. F., de Jong, M. D., van Doorn, H. R. & Holmes, E. C. Guidelines for identifying homologous recombination events in influenza A virus. *PLoS ONE* **5**, e10434 (2010).
20. He, B. et al. Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China. *J. Virol.* **88**, 7070–7082 (2014).
21. Hu, B. et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
22. Li, X. et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci. Adv.* **6**, eabb9153 (2020).
23. Lin, X. et al. Extensive diversity of coronaviruses in bats from China. *Virology* **507**, 1–10 (2017).
24. Wang, L. et al. Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern China. *Emerg. Microbes Infect.* **6**, e14 (2017).
25. Zhou, H. et al. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the Spike protein. *Curr. Biol.* **30**, 2196–2203 (2020).
26. Yuan, J. et al. Intraspecies diversity of SARS-like coronaviruses in *Rhinolophus sinicus* and its implications for the origin of SARS coronaviruses in humans. *J. Gen. Virol.* **91**, 1058–1062 (2010).
27. Wan, Y., Shang, J., Graham, R., Baric, R. & Li, F. Receptor recognition by the novel Coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* **94**, e00127–20 (2020).
28. Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. Preprint at https://doi.org/10.1101/2020.04.20.052019 (2020).
29. Menachery, V. D. et al. A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1514 (2015).
30. Menachery, V. D. et al. SARS-like WIV1-CoV poised for human emergence. *Proc. Natl Acad. Sci. USA* **113**, 3048–3053 (2016).
31. Ge, X. et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
32. Zhang, Y.-Z. & Holmes, E. C. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* **181**, 223–227 (2020).
33. Patino-Galindo, J. A., Filip, I., AlQuraishi, M. & Rabadan, R. Recombination and lineage-specific mutations led to the emergence of SARS-CoV-2. Preprint at https://doi.org/10.1101/2020.02.10.942748 (2020).
34. Eden, J.-S., Tanaka, M. M., Boni, M. F., Rawlinson, W. D. & White, P. A. Recombination within the pandemic norovirus GII.4 lineage. *J. Virol.* **87**, 6270–6282 (2013).
35. Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. W. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.* **23**, 1891–1901 (2006).
36. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1**, vev003 (2015).
37. Anderson, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).

38. Lie, P., Chen, W. & Chen, J.-P. Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses* **11**, 979 (2019).
39. Rambaut, A., Lam, T. T., Carvalho, L. M. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
40. Trova, S. et al. Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* **1**, vev016 (2015).
41. Duchene, S. et al. Bayesian evaluation of temporal signal in measurably evolving populations. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msaa163 (2020).
42. Suchard, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
43. Aiewsakun, P. & Katzourakis, A. Time-dependent rate phenomenon in viruses. *J. Virol.* **90**, 7184–7195 (2016).
44. Membrebe, J. V., Suchard, M. A., Rambaut, A., Baele, G. & Lemey, P. Bayesian inference of evolutionary histories under time-dependent substitution rates. *Mol. Biol. Evol.* **36**, 1793–1803 (2019).
45. Holmes, E. C. *The Evolution and Emergence of RNA Viruses* (Oxford Univ. Press, 2009).
46. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of Ebola virus: insights from the 2013–2016 epidemic. *Nature* **538**, 193–200 (2016).
47. Holmes, E. C., Rambaut, A. & Andersen, K. G. Pandemics: spend on surveillance, not prediction. *Nature* **558**, 180–182 (2018).
48. Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *J. Virol.* **84**, 3134–3146 (2010).
49. Su, S. et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.* **24**, 490–502 (2016).
50. Schierup, M. H. & Hein, J. Recombination and the molecular clock. *Mol. Biol. Evol.* **17**, 1578–1579 (1999).
51. Posada, D., Crandall, K. A. & Holmes, E. C. Recombination in evolutionary genomics. *Annu Rev. Genet.* **36**, 75–97 (2002).
52. Duchene, S., Holmes, E. C. & Ho, S. Y. W. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc. Lond. B* **281**, 20140732 (2014).
53. Maclean, O. A., Lytras, S., Singer, J. B., Weaver, S. & Sergei, L. Evidence of significant natural selection in the evolution of SARS-CoV-2 in bats, not humans. Preprint at https://doi.org/10.1101/2020.05.28.122366 (2020).
54. Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *J. Med Virol.* **92**, 433–440 (2020).
55. Anderson, K. G. nCoV-2019 codon usage and reservoir (not snakes v2). *Virological.org* http://virological.org/t/ncov-2019-codon-usage-and-reservoir-not-snakes-v2/339 (2020).
56. Robertson, D. nCoV's relationship to bat coronaviruses & recombination signals (no snakes) – no evidence the 2019-nCoV lineage is recombinant. *Virological.org* http://virological.org/t/ncovs-relationship-to-bat-corona-viruses-recombination-signals-no-snakes-no-evidence-the-2019-ncov-lineage-is-recombinant/331 (2020).
57. Wong, A. C. P., Li, X., Lau, S. K. P. & Woo, P. C. Y. Global epidemiology of bat coronaviruses. *Viruses* **11**, 174 (2019).
58. Katoh, K., Asimenos, G. & Toh, H. in *Bioinformatics for DNA Sequence Analysis* (ed. Press, H.) 39–64 (Springer, 2009).
59. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. MERS-CoV spillover at the camel–human interface. *eLife* **7**, e31257 (2018).
60. Chernomor, O. et al. Split diversity in constrained conservation prioritization using integer linear programming. *Methods Ecol. Evol.* **6**, 83–91 (2015).
61. Boni, M. F., Posada, D. & Feldman, M. W. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176**, 1035–1047 (2007).
62. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
63. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
64. Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681 (2006).
65. Bryant, D. & Moulton, V. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**, 255–265 (2004).
66. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2014).
67. Lemey, P., Minin, V. N., Bielejec, F., Pond, S. L. K. & Suchard, M. A. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics* **28**, 3248–3256 (2012).
68. Yres, D. L. et al. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Softw. Syst. Evol.* **68**, 1052–1061 (2019).

Acknowledgements

We thank all authors who have kindly deposited and shared genome data on GISAID. We thank T. Bedford for providing M.F.B. with an alignment on which an initial recombination analysis was done. We thank A. Chan and A. Irving for helpful comments on the manuscript. The research leading to these results received funding (to A.R. and P.L.) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS). D.L.R. is funded by the MRC (no. MC_UU_1201412). The Artic Network receives funding from the Wellcome Trust through project no. 206298/Z/17/Z. P.L. acknowledges support by the Research Foundation—Flanders ('Fonds voor Wetenschappelijk Onderzoek—Vlaanderen' (nos. G066215N, G0D5117N and G0B9317N)) and by the European Union's Horizon 2020 project MOOD (no. 874850). T.L. is funded by The National Natural Science Foundation of China Excellent Young Scientists Fund (Hong Kong and Macau; no. 31922087). We thank originating laboratories at South China Agricultural University (Y. Shen, L. Xiao and W. Chen; no. EPI_ISL_410721) and Beijing Institute of Microbiology and Epidemiology (W.-C. Cao, T.T.-Y.L., N. Jia, Y.-W. Zhang, J.-F. Jiang and B.-G. Jiang, nos. EPI_ISL_410538, EPI_ISL_410539, EPI_ISL_410540, EPI_ISL_410541 and EPI_ISL_410542) for the use of sequence data via the GISAID platform.

Author contributions

All authors contributed to analyses and interpretations. D.L.R. and X.J. performed recombination and phylogenetic analysis and annotated virus names with geographical and sampling dates. A.R. performed S recombination analysis. B.W.P. and T.A.C. performed codon usage analysis. M.F.B. performed recombination analysis for non-recombining regions 1 and 2, breakpoint analysis and phylogenetic inference on recombinant segments. P.L. performed recombination analysis for non-recombining alignment 3, calibration of rate of evolution and phylogenetic reconstruction and dating. T.T.-Y.L. collected SARS-CoV data and assisted in analyses of SARS-CoV and SARS-CoV-2 data. M.F.B., P.L. and D.L.R. wrote the first draft of the manuscript, and all authors contributed to manuscript editing.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-020-0771-4.

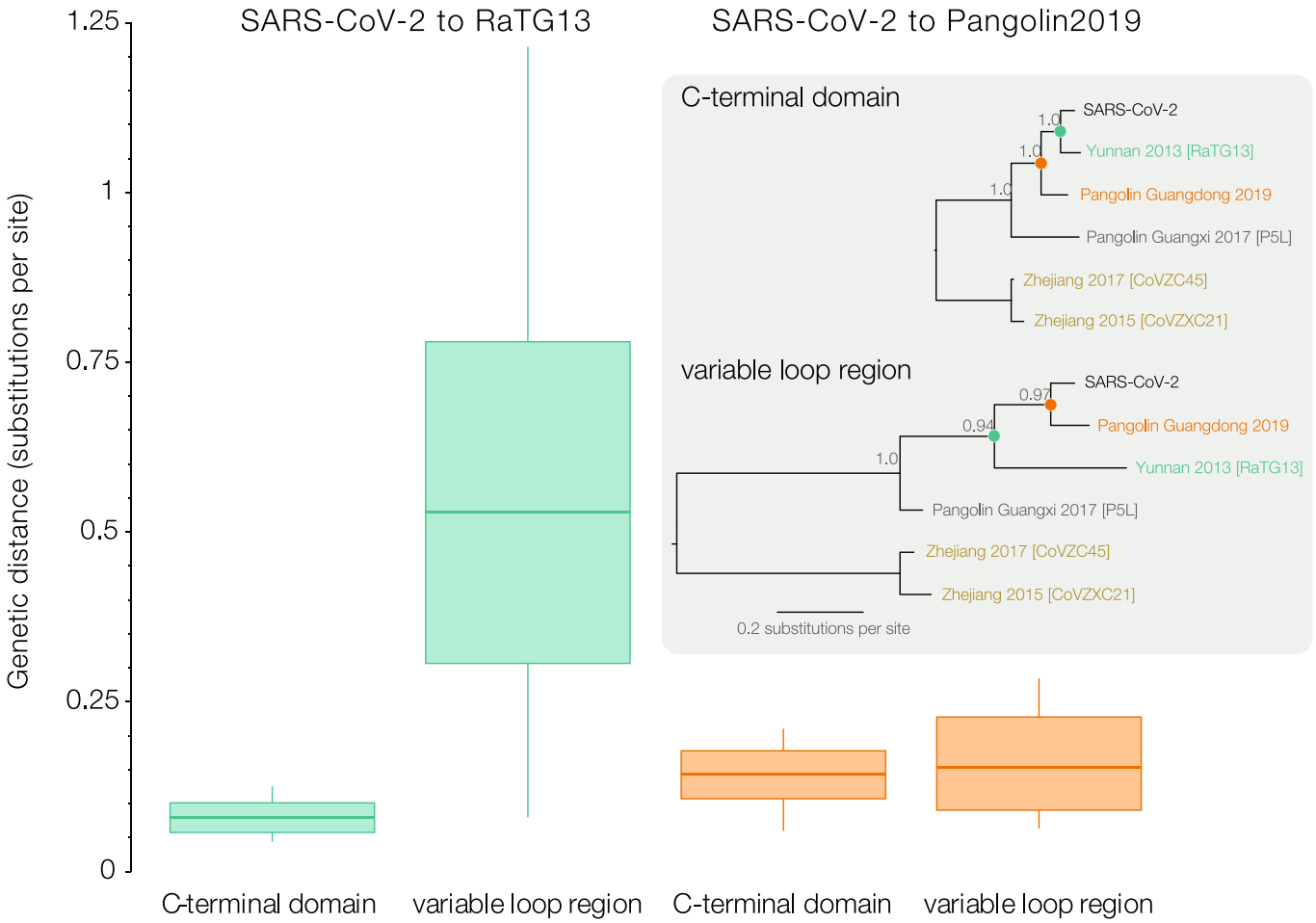
Supplementary information is available for this paper at https://doi.org/10.1038/s41564-020-0771-4.

Correspondence and requests for materials should be addressed to M.F.B., P.L., A.R. or D.L.R.

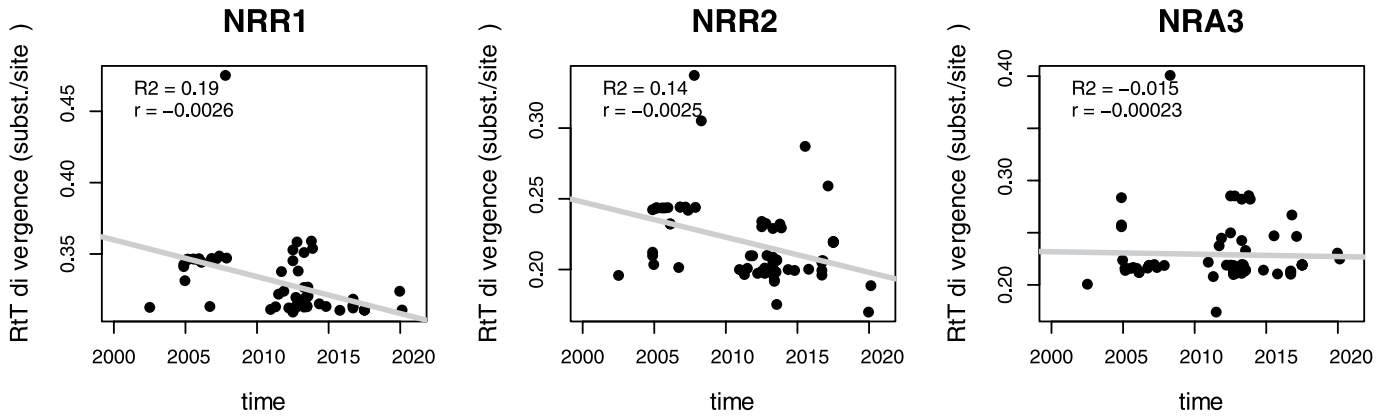
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

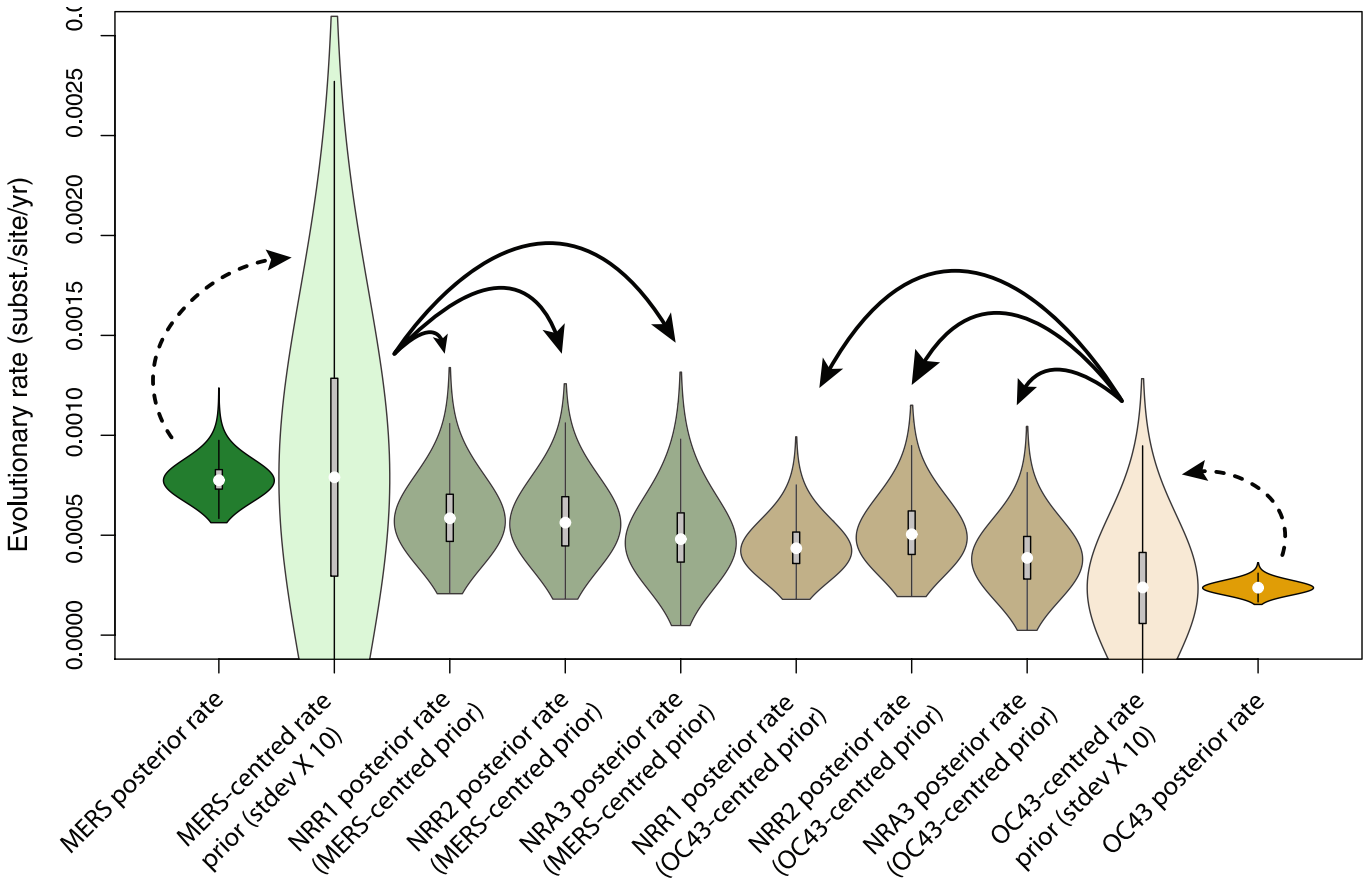
© The Author(s), under exclusive licence to Springer Nature Limited 2020



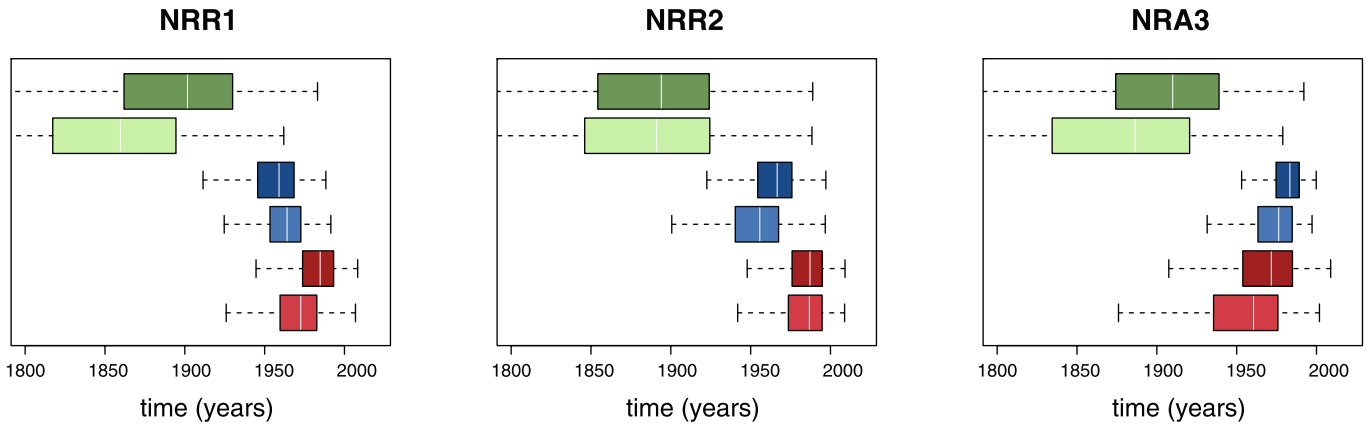
Extended Data Fig. 1 | Phylogenetic relationships in the C-terminal domain (CTD). Posterior means (horizontal bars) of patristic distances between SARS-CoV-2 and its closest bat and pangolin sequences, for the spike protein's variable loop region and CTD region excluding the variable loop. Boxes show 95% HPD credible intervals. Means and 95% HPD intervals are 0.080 [0.058–0.101] and 0.530 [0.304–0.780] for the patristic distances between SARS-CoV-2 and RaTG13 (green) and 0.143 [0.109–0.180] and 0.154 [0.093–0.231] for the patristic distances between SARS-CoV-2 and Pangolin 2019 (orange). Gray inset shows majority rule consensus trees with mean posterior branch lengths for the two regions, with posterior probabilities on the key nodes showing the relationships among SARS-CoV-2, RaTG13, and Pangolin 2019.



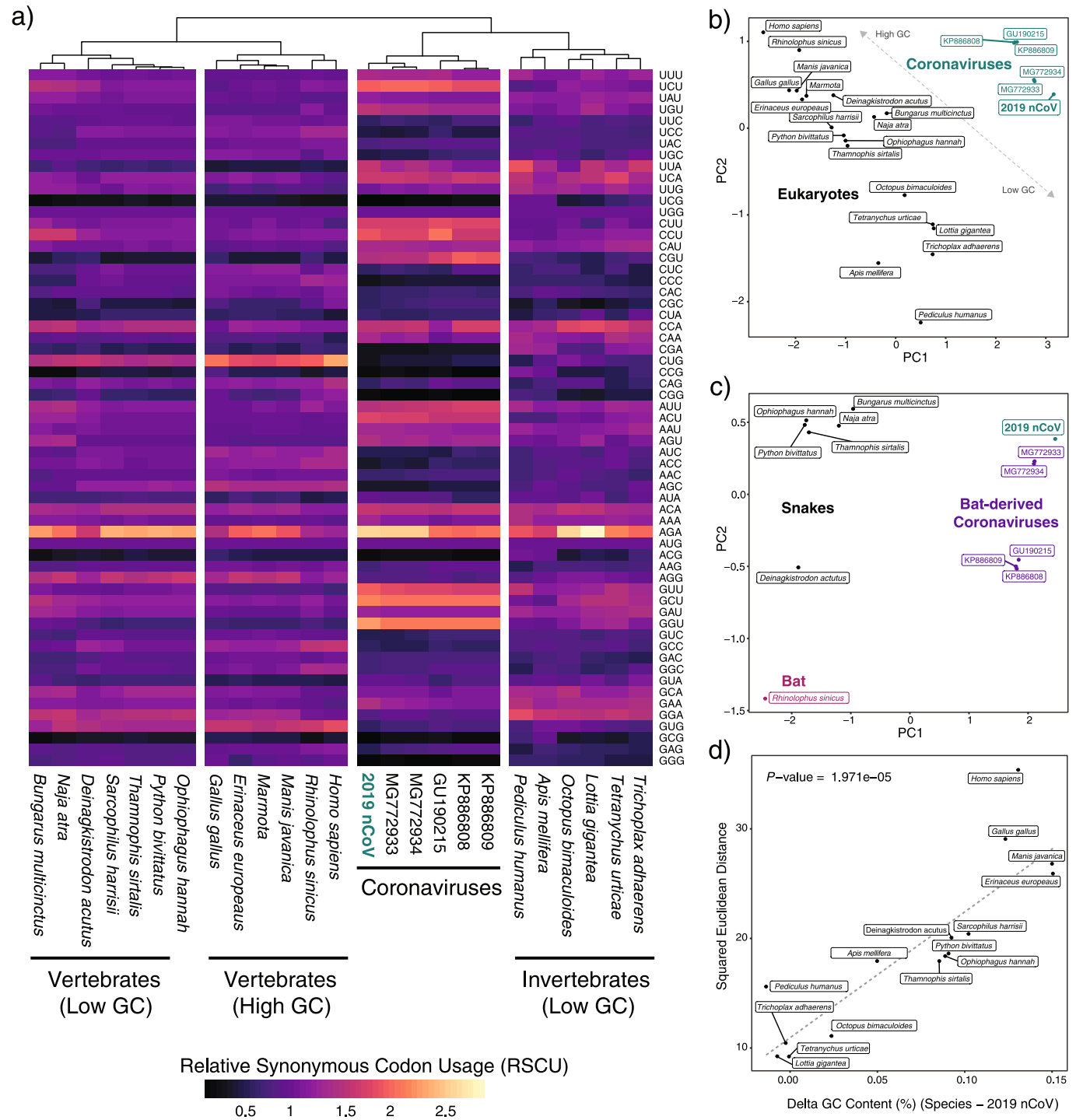
Extended Data Fig. 2 | Lack of root-to-tip temporal signal in SARS-CoV-2. Root-to-tip divergence as a function of sampling time for non-recombinant regions **NRR1** and **NRR2** and recombination-masked alignment set **NRA3**. The plots are based on maximum likelihood tree reconstructions with a root position that maximises the residual mean squared for the regression of root-to-tip divergence and sampling time.



Extended Data Fig. 3 | Priors and posteriors for evolutionary rate of SARS-CoV-2. Posterior rate distributions for MERS-CoV (far left) and HCoV-OC43 (far right) using BEAST on $n = 27$ sequences spread over 4 years (MERS-CoV) and $n = 27$ sequences spread over 49 years (HCoV-OC43). As illustrated by the dashed arrows, these two posteriors motivate our specification of prior distributions with standard deviations inflated 10-fold (light color). These rate priors are subsequently used in the Bayesian inference of posterior rates for **NRR1, NRR2**, and **NRA3** as indicated by the solid arrows.



Extended Data Fig. 4 | TMRCAs for SARS-CoV and SARS-CoV-2. Divergence time estimates based on the three regions/alignments where the effects of recombination have been removed. The red and blue boxplots represent the divergence time estimates for SARS-CoV-2 (red) and the 2002-2003 SARS-CoV (blue) from their most closely related bat virus, with the light- and dark-colored versions based on the HCoV-OC43 and MERS-CoV centered priors, respectively. Green boxplots show the TMRCA estimate for the RaTG13/SARS-CoV-2 lineage and its most closely related pangolin lineage (Guangdong 2019), with the light and dark coloured version based on the HCoV-OC43 and MERS-CoV centred priors, respectively. TMRCA estimates for SARS-CoV-2 and SARS-CoV from their respective most closely related bat lineages are reasonably consistent for the different data sets and different rate priors in our analyses. Posterior means with 95% HPDs are shown in Supplementary Information Table 2.



Extended Data Fig. 5 | Comparisons of GC content across taxa. Conducting analogous analyses of codon usage bias as Ji et al. (2020) with additional (and higher quality) snake coding sequence data and several miscellaneous eukaryotes with low genomic GC content failed to find any meaningful clustering of the SARS-CoV-2 with snake genomes (**a**). Instead, similarity in codon usage metrics between the SARS-CoV-2 and eukaryotes analyzed was correlated with coding sequence GC content of the eukaryote, with more similar codon usage being identified in eukaryotes with low GC content similar to that of the coronavirus (**b**).

Corresponding author(s): Maciej Boni, Philippe Lemey, Andrew Rambaut, David Robertson

Last updated by author(s): July 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All in Methods section of paper.
Data analysis	Software used: RAxML (v8.2.8), 3SEQ (v1.7), RDP5 (v5), GARD (v2.5.0), BEAST (v1.10.4), BEAGLE (v3), MAFFT (v7.310), Neighbor-Nets (SplitsTree v4.15.1), PHI-Test (SplitsTree v4.15.1), IQTREE (v2.0), TempEst (v1.5.3), Tree Annotator (v1.10.4), FigTree(v1.4.2), Phylogenetic Diversity Analyzer Tool (v0.5), BBMap (v38.75), pheatmap (v1.0.12). All custom code available at https://github.com/plemey/SARSCoV2origins

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- All sequence data are available in GenBank and GISAID, with accession numbers listed in Supplementary Table S4.

All sequence data analyzed in this manuscript are available at https://github.com/plemey/SARSCoV2origins. Note that one of the sequences requires GISAID permissions from original authors, but will be made available either individually or publicly depending on original depositor's request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Sarbecovirus genomes were downloaded from publicly available sources. Non-recombinant regions were identified. Evolutionary rate calibration for coronaviruses was performed using publicly available data sets for SARS, MERS, and OC43. An evolutionary rate was inferred for three putative non-recombinant regions of the sarbecoviruses. Time to the most common ancestor between SARS-CoV-2 and its closest bat-virus relative were computed.
Research sample	All available genomes (n=68) for sarbecoviruses -- including the SARS-CoV-2 genome; 27 genomes of human coronavirus OC43; 35 genomes of MERC-CoV; 69 genomes of SARS-CoV.
Sampling strategy	No sample size calculation was done as it was possible to use all samples (all available sequences) in the analysis.
Data collection	Publicly available sequence data were downloaded from GenBank and GISAID.
Timing and spatial scale	Sarbecovirus genomes were collected over an 18-year period (2002-2020), OC43 sequences from 1968 to 2016, MERS sequences from 2012 to 2015, and SARS-CoV sequences from 2002-2004.
Data exclusions	Sequences without 'sampling year' were excluded. This exclusion criterion was not formally pre-established, but this exclusion is both common and necessary when estimating evolutionary rates from sequence data. It goes without saying that sequences without dates must be excluded.
Reproducibility	N/A;
Randomization	N/A
Blinding	N/A

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging