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

Citation for published version:

Digital Object Identifier (DOI):
10.1038/s41564-020-0771-4

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Nature Microbiology

Publisher Rights Statement:
This is the accepted version of the following article: Boni, M.F., Lemey, P., Jiang, X. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol (2020). https://doi.org/10.1038/s41564-020-0771-4, which has been published in final form at https://doi.org/10.1038/s41564-020-0771-4.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic

Maciej F Boni¹*, Philippe Lemey²*, Xiaowei Jiang³, Tommy Tsan-Yuk Lam⁴, Blair Perry⁵, Todd Castoe⁵, Andrew Rambaut⁶* and David L Robertson⁷*

*joint first authors

¹Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, University Park, PA, USA.
²Department of Microbiology, Immunology and Transplantation, KU Leuven, Leuven, Belgium.
³Department of Biological Sciences, Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, Jiangsu, China.
⁴State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China.
⁵Department of Biology, University of Texas Arlington, Arlington, TX, USA.
⁶Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
⁷MRC-University of Glasgow Centre for Virus Research (CVR), Glasgow, UK.

Lead contact: Maciej F Boni, mfb9@psu.edu

*Corresponding authors: mfb9@psu.edu, philippe.lemey@kuleuven.be, a.rambaut@ed.ac.uk, david.l.robertson@glasgow.ac.uk

Word Count (Introduction, Results, Discussion): 3956

Date: July 6, 2020
Abstract (189 words)

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, not one acquired recently via recombination. In order 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% 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.
Introduction.

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 etiological 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 \textit{Sarbecovirus} subgenus of \textit{Coronaviridae}\(^7\), the same sub-genus as the SARS virus that caused a global outbreak of more than 8000 cases in 2002-2003. By mid-January, the virus was spreading widely inside 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, as this may allow health authorities to separate human populations from the wild life 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 January 10, 2020 (UCT) on Virological.org by a consortium led by Yong-Zhen Zhang\(^6\) it enabled immediate analyses of its ancestry. Across a large region of the virus’s genome, corresponding approximately to ORF1b, it did not cluster with any of the known bat coronaviruses indicating that recombination likely played a role in the evolutionary history of these viruses\(^5,7\). Subsequently, a bat sarbecovirus – RaTG13 sampled from a \textit{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 sub-regions can have independent origins, identifiable if sufficient sampling has been done in the animal reservoirs that support the endemic circulation, coinfection, 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\(^5,20-26\). Despite the SARS-CoV-2 lineage’s acquiring residues in its Spike (S) protein’s receptor-binding domain (RBD) permitting the use of human ACE2\(^27\) receptors and its RBD being genetically closer to a pangolin virus than to RaTG13\(^11-13,22,28\) – 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 the bat sarbecoviruses related to the first SARS-CoV lineage\textsuperscript{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 as yet 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\textsuperscript{25,32}.

Results

Recombination analysis and identification of breakpoint-free genome regions

Of the 68 sequences in the aligned sarbecovirus sequence set, 67 show evidence of mosaicism (all Dunn-Sidak corrected \( p < 4 \times 10^{-4}, \text{3SEQ}^{14} \)), indicating involvement in homologous recombination either directly with identifiable parentals or in their deeper shared evolutionary history, i.e., due to shared ancestral recombination events. This is evidence for numerous recombination events occurring in the evolutionary history of the sarbecoviruses\textsuperscript{22,33}; specifying all past events in their correct temporal order\textsuperscript{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 by revealing regions with no breakpoints. Sorting these breakpoint-free regions (BFRs) by length results in two segments longer than 5kb: an ORF1a sub-region spanning nucleotides 3625-9150 and the first half of ORF1b spanning nucleotides 13291-19628 (sequence numbering in Source Data https://github.com/plemey/SARSCoV2origins). Eight other BFRs longer than 500nt were identified, and the regions were named BFR A to BFR J in order of length. Of the nine breakpoints defining these ten BFRs, four showed phylogenetic incongruence (PI) signals with bootstrap support higher than 80%, adopting previously published criteria on using a combination of mosaic and PI-signals to show evidence of past recombination events\textsuperscript{39}. All four of these breakpoints were also identified with the tree-based recombination detection method GARD\textsuperscript{35}.

The extent of the sarbecoviruses’ recombination history can be illustrated by five phylogenetic trees inferred from BFRs or concatenated adjacent BFRs (Figure 1, bottom). BFRs were concatenated if no phylogenetic incongruence signal could be identified between them. When viewing the last 7kb 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\textsuperscript{22,25}, show strong
phylogenetic recombination signals, as they fall on different evolutionary lineages (with bootstrap support >80%) depending 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 sub-alignment for phylogenetic analysis. Conservatively, we combined the three >2kb BFRs identified above into non-recombining region 1 (NRR1). Removing five sequences that appear to be recombinants and two small sub-regions of BFR A was necessary in order 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 9.9kb region spanning nucleotides 11885-21753 (NRR2) as a putative non-recombining region; this approach is breakpoint-conservative as 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 – gives a putative recombination-free alignment that we call non-recombinant alignment 3 (NRA3) (see Methods).

All three approaches to removing 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 (Figure 1, bottom). Specifically, progenitors of the RaTG13/SARS-CoV-2 lineage appear to have recombined with the Hong Kong clade (with inferred breakpoints at 11.9kb and 20.8kb) 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.

As the SARS-CoV-2 S protein has been implicated in past recombination events or possibly convergent evolution12, we specifically investigated several sub-regions 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 (Figure 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 others11,22. However, on closer inspection, the relative divergences in the phylogenetic tree (Figure 2, bottom) show that SARS-CoV-2 is unlikely to have acquired the variable loop from an ancestor of Pangolin-2019 as these two sequences are approximately 10% to 15% divergent throughout the entire S protein (excluding the NTD). It is RaTG13 that is more divergent in the variable loop region (Extended Data Figure 1) and thus likely to be the product of recombination, acquiring a divergent variable loop from an as yet unsampled bat sarbecovirus28. This is notable because the variable loop region contains the six key contact residues in the receptor binding domain (RBD) that give SARS-CoV-2 its ACE2 binding specificity27,37. 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 identifying 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 (Figure 3). Given that these pangolin viruses are ancestral to the progenitor of the 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 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 sub-regions of NRR1 depict an appreciable degree of spatial structuring of the bat sarbecoviruses population across different regions (Figure 3). One geographic clade includes viruses from provinces in southern China (Guangxi, Yunnan, Guizhou, Guangdong), with its major sister clade consisting of viruses from provinces in northern China (Shanxi, Henan, Hebei, 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 viruses and southern viruses, while the two Zhejiang viruses (CoVZXC21, CoVZC45) appear to carry a recombinant region from southern or central China.

TMRCA for non-recombinant regions of SARS-CoV-2 lineage

To avoid artefacts due to recombination, we focused on the non-recombining regions NRR1, NRR2, and the recombination-masked alignment NRA3 for inferring time-measured evolutionary histories. Visual exploration using TempEst indicates there is no evidence for temporal signal in these data sets (Extended Data Figure 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 significantly impact expected root-to-tip divergences over a sampling time range that represents only a small fraction of the evolutionary history. However, formal testing using marginal likelihood estimation 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 Information 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 through time, including HCoV-OC43, MERS-CoV, and SARS-CoV virus genomes. These data sets were subjected to the same
recombination masking approach as **NRA3** and were characterized by a strong temporal signal (Figure 4), but also by markedly different evolutionary rates. Specifically, using a formal Bayesian approach\(^2\) (see **Methods**), we estimate a fast evolutionary rate (0.00169 subst/site/yr, 95% highest posterior density (HPD) interval [0.00131,0.00205]) for SARS viruses sampled over a limited time scale (one year), a slower rate (0.00078 [0.00063,0.00092] subst/site/yr) for MERS-CoV on a time scale of about four years, and the slowest rate (0.00024 [0.00019,0.00029] subst/site/yr) for HCoV-OC43 over almost five decades. These differences reflect the fact that rate estimates can vary considerably with the time scale of measurement, a frequently observed phenomenon in viruses known as time-dependent evolutionary rates\(^41,43,44\). Over relatively shallow time-scales, such differences can primarily be explained by varying selective pressure with mildly deleterious variants being eliminated more strongly by purifying selection over longer time scales\(^44–46\). Consistent with this, we estimate a concomitantly decreasing nonsynonymous-to-synonymous substitution rate ratio (dN/dS) over longer evolutionary time-scales: 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 calibrating the sarbecovirus evolutionary history. We compare both MERS-CoV and HCoV-OC43 centered prior distributions (Extended Data Figure 3) in order to examine the sensitivity of the date estimates to this prior specification.

We infer time-measured evolutionary histories using a Bayesian phylogenetic approach while incorporating rate priors based on the 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**, centered around 0.00055 subst/site/yr. The fact that these estimates lie between the MERS-CoV and HCoV-OC43 rates is consistent with the intermediate sampling time range of about 18 years (Figure 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 Figure 2). Below, we report divergence time estimates based on the HCoV-OC43-centered rate prior for **NRR1**, **NRR2**, and **NRA3**, and summarize corresponding estimates for the MERS-CoV-centred rate priors in Extended Data Figure 4. Divergence time estimates based on the HCoV-OC43-centered 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 the divergence time estimate between SARS-CoV and its most closely related bat sequence is 1962 (95% HPD: 1932-1988); see Figure 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 previously estimated dates\(^15\) as...
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 the evolutionary rates were predominantly informed by the short-term SARS outbreak scale and likely biased upwards. Indeed, the rates reported by these studies are in line with the short-term SARS rates we estimate (Figure 4). The estimated divergence times for the pangolin virus most closely related to the SARS-CoV-2/RaTG13 lineage 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 as 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 SE Asia’s initial outbreaks of avian influenza H5N1 (1997, 2003-2004) or the first SARS outbreak (2002-2003). By 2009, however, rapid genomic analysis was 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 in Wuhan was caused by a coronavirus. 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, 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.

The difficulty in inferring reliable evolutionary histories for coronaviruses is that their high recombination rate violates the assumption of standard phylogenetic approaches because different parts of the genome have different histories. In order to begin characterizing any ancestral relationships for SARS-CoV-2, non-recombinant regions 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. As there is no single accepted method of inferring breakpoints and identifying clean sub-regions with high certainty, we implemented several approaches to identifying three classic statistical signals of recombination: mosaicism, phylogenetic incongruence, and excessive homoplasy. Our most conservative approach attempted to ensure that putative non-recombining regions had no mosaic or phylogenetic incongruence signals. A second breakpoint-conservative approach was conservative with respect to breakpoint identification, but this means it is accepting of false negative outcomes in breakpoint inference, resulting in less certainty that a putative non-recombining region truly contains no breakpoints.
third approach attempted to minimize the number of regions removed while also minimizing signals of mosaicism and homoplasy. The origins we present in Figure 5 (NRR1) are conservative in the sense that NRR1 is more likely to be non-recombinant than NRR2 or NRA3. As the estimated rates and divergence dates were highly similar in the three data sets analyzed, we conclude that our estimates are robust to the method of identifying a genome’s non-recombinant regions.

Due to the absence of temporal signal in the sarbecovirus data sets, 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 time scale of sample collection. In the presence of time-dependent rate variation, a widely observed phenomenon for viruses, 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 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. Accurately estimating ages for deeper nodes would require adequately accommodating time-dependent rate variation. While such models have recently been made available, we lack the information to calibrate the rate decline through time (e.g. through internal node calibrations). As a proxy, it would be possible to model the long-term purifying selection dynamics as a major source of time-dependent rates, but this is beyond the scope of the current study. The assumption of long-term purifying selection would imply that coronaviruses are at 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, this is inferred to have occurred before the divergence of RaTG13 and SARS-CoV-2 so 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 (Figure 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 SARS-CoV-2’s receptor-binding domain. 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 pangolins are facilitating adaptation to humans. A hypothesis of snakes as intermediate hosts of SARS-CoV-2 was posited during the early epidemic phase, but we found no evidence of this, see Extended Data Figure 5.

With horseshoe bats currently the most plausible origin of SARS-CoV-2, it is important to consider that sarbecoviruses circulate in a variety of horseshoe bat species with widely overlapping species ranges. Yet, the viral population is largely spatially structured according to provinces in the south and southeast on
one lineage, and provinces in the center, east, and northeast on another (Figure 3). This boundary appears to be rarely crossed. Two exceptions can be seen in the relatively close relationship of Hong Kong viruses to viruses 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 Figure 3) clusters with viruses from provinces in the center, east, and northeast of China. SARS-CoV-2 and RaTG13 are also exceptions as 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 receptor binding domain was very likely to have been present in a bat sarbecovirus lineage that ultimately lead to SARS-CoV-2, this RBD sequence has thus far only been found in a few pangolin viruses. Furthermore, the other key feature thought to be instrumental to SARS-CoV-2’s ability to infect humans – a polybasic cleavage site insertion in the Spike 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 significant human outbreaks before they emerge. This underscores the need for a global network of real-time human disease surveillance systems, like that which identified the unusual cluster of pneumonia in Wuhan in December 2019, with capacity to rapidly deploy genomic tools and functional studies for pathogen identification and characterization.
Methods

Data set compilation

Sarbecovirus data. Complete genome sequence data were downloaded from GenBank and ViPR; accession numbers of all 68 sequences available in Supplementary Information Table 4. Sequences were aligned by MAFFT58 v7.310, with a final alignment length of 30927, and used in the analyses below.

HCoV-OC43. We compiled a data set including 27 human coronavirus OC43 virus genomes and 10 related animal virus genomes (6 bovine, 3 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 data set analysed by Dudas et al59 using the phylogenetic diversity analyser tool60 (v0.5)

SARS-CoV. We compiled a set of 69 SARS-CoV genomes including 58 genomes sampled from human and 11 sampled from civets and raccoon dogs. This data set comprises an updated version of the one used in Hon et al15 and includes a cluster of genomes sampled late 2003 and early 2004, but the evolutionary rate estimate without this cluster (0.00175 subst/site/yr, [0.00117,0.00229]) is consistent with the complete data set (0.00169 subst/site/yr, [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

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

First, we took an approach that relies on identifying mosaic regions (via 3SEQ14 v1.7) that are also supported by phylogenetic incongruence (PI) signals19. As 3SEQ is the most statistically powerful of the mosaic methods61, we used 3SEQ to identify the best supported breakpoint history for each potential child (recombinant) sequence in the data set. 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. Then, we (a) collected all breakpoints into a single set, (b) complemented this set to generate a set of non-breakpoints, (c) grouped the non-breakpoints into contiguous breakpoint-free regions (BFRs), and (d) sorted these regions by length. A phylogenetic tree – using RAxML v8.2.862,63, GTR+\(\Gamma\) model, 100 bootstrap replicates – was inferred for each BFR longer than 500nt.

We considered both (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 13291-19628, length=6338nt), BFR B (nt positions 3625-9150, length=5526nt), BFR C (nt positions 9261-11795, length=2535nt), BFR D (nt positions 27702-28843, length=1142nt), and six more regions E through J. Phylogenetic trees and exact
breakpoints for all ten BFRs are shown in Supplementary Figures 1-10. Regions A, B, and C had similar phylogenetic relationships among the southern China bat viruses (Yunnan, Guangxi, Guizhou provinces), the Hong Kong viruses, northern Chinese viruses (Jilin, Shanxi, Hebei, Henan provinces; incl. Shaanxi), pangolin viruses, and the SARS-CoV-2 lineage. As these subclades had different phylogenetic relationships in region D (Supplementary Figure 4), region D and shorter BFRs were not further included into a combined putative non-recombinant regions.

Regions A, B, and C were further examined for mosaic signals by 3SEQ; all showed signs of mosaicism. In region A, we removed sub-region A1 (nt positions 3872-4716 within region A) and sub-region A4 (nt 1642-2113) as both of these showed PI-signals with other sub-regions 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), 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 non-recombining region 1, or NRR1, as an alignment of 63 sequences. We say that this approach is conservative because sequences and sub-regions 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.

Second, 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 as the construction of NRR1, in that this approach is the most likely to allow breakpoints to remain inside putative non-recombing 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 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, (3) the GARD algorithm which identifies breakpoints by identifying PI-signals across proposed breakpoints. Since 3SEQ identified 10 BFRs longer than 500nt, we used GARD’s (v2.5.0) 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 as GARD is computationally intensive. GARD identified eight breakpoints that were also within 50nt of breakpoints identified by 3SEQ. PI-signals were identified (with bootstrap support >80%) for seven of these eight breakpoints: positions 1684, 3046, 9237, 11885, 21753, 22773, and 24628. Using these breakpoints, the longest putative non-recombining segment (nt 11885-21753) is 9.9kb 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 version 5.5 of RDP5 (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 (i) a significantly different outcome of the PHI-test, (ii) a removal of well-supported (bootstrap value >95%) incompatible splits in Neighbor-Net, and (iii) 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 the 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 inappropriately modelling evolutionary processes with recombination on strictly bifurcating trees, which can result in different artefacts such as homoplasies that inflate branch lengths and result in apparently longer evolutionary divergence times. In order to examine temporal signal in the sequenced data, we plotted root-to-tip divergence against sampling time using TempEst v1.5.3 based on a maximum likelihood tree. The latter was reconstructed using IQTREE v2.0 under a General-Time Reversible (GTR) model with a discrete gamma distribution to model among-site rate variation.

Time-measured phylogenetic reconstruction was performed using a Bayesian approach implemented in BEAST v1.10.4. When the genomic data included both coding and noncoding regions we used a single GTR+Γ substitution model; for concatenated coding genes we partitioned the alignment by codon positions and specified an independent GTR+Γ model for each partition with a separate gamma model to accommodate among-site rate variation. We used an uncorrelated relaxed clock model with a lognormal distribution for all data sets, except for the low-diversity SARS data for which we specified a strict molecular clock model. For the HCoV-OC43, MERS-CoV, and SARS data sets we specified flexible skygrid coalescent tree priors. In the absence of any reasonable prior knowledge on the TMRCA of the sarbecovirus data sets (which is required for the grid specification in a skygrid model), we specified a simpler constant size population prior. As informative rate priors for the analysis of the sarbecovirus data sets, we used two different normal prior distributions: one with a mean of 0.00078 and a standard deviation of 0.00075 and one with a mean of 0.00024 and standard deviation of 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 larger than the empirical standard deviations to allow more prior uncertainty and avoid strong bias (Extended Data Figure 3). In our analyses of the sarbecovirus data sets, we incorporated the uncertainty of the sampling dates when exact dates were not available. To estimate nonsynonymous over synonymous rate ratios for the concatenated coding genes, we used the empirical Bayes ‘Renaissance counting’
Temporal signal was tested using a recently developed marginal likelihood estimation procedure\(^1\) (Supplementary Information 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\(^3\) library for efficient likelihood computations. We used TreeAnnotator to summarize the posterior tree distributions and annotated the estimated to a maximum clade credibility tree, which was visualized using FigTree.

Data Availability Statement
All sequence data analyzed 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 Statement
All custom code used in the manuscript is available at https://github.com/plemey/SARSCoV2origins.

Correspondence
Correspondence should be addressed to Maciej F Boni, Center for Infectious Disease Dynamics, Department of Biology, Pennsylvania State University, mfb9@psu.edu.

Acknowledgements
We would like to thank all the authors who have kindly deposited and shared genome data on GISAID. Thanks to Trevor Bedford for providing MFB an alignment on which an initial recombination analysis was done. Thanks to Alina Chan and Aaron Irving for helpful comments on the manuscript. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS). DLR is funded by the MRC (MC_UU_1201412). The Artic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z. PL acknowledges support by the Research Foundation -- Flanders ('Fonds voor Wetenschappelijk Onderzoek -- Vlaanderen', G066215N, G0D5117N and G0B9317N). TL is funded by The National Natural Science Foundation of China (NSFC) Excellent Young Scientists Fund (Hong Kong and Macau) (31922087). Thanks to originating laboratories at South China Agricultural University (Yongyi Shen, Lihua Xiao, Wu Chen; re EPI_ISL_410721) and Beijing Institute of Microbiology and Epidemiology (Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang; re EPI_ISL_410538, EPI_ISL_410539, EPI_ISL_410540, EPI_ISL_410541, EPI_ISL_410542) for the use of sequence data via the GISAID platform.

Author Contributions
All authors contributed to analyses and interpretations. DLR and XJ performed recombination and phylogenetic analysis, and annotated the virus names with geographical and sampling dates. AR performed the S recombination analysis. BP and TC performed the codon-usage analysis. MFB performed the recombination analysis for non-recombining regions 1 and 2, breakpoint analysis, and phylogenetic inference on recombinant segments. PL performed recombination analysis for non-recombining alignment 3, calibration of rate of evolution, and phylogenetic reconstruction and dating. TT-YL collected SARS-CoV data, and assisted in analyses of SARS-CoV and SARS-CoV-2 data. MFB, PL, and DLR wrote the first draft of the manuscript, and all authors contributed to manuscript editing.

Competing Interests
The authors declare that they have no competing interests.

References

15. Hon, C. et al. Evidence of the Recombinant Origin of a Bat Severe Acute Respiratory Syndrome (SARS)-Like Coronavirus...


**Figure Captions**

**Figure 1.** **Top:** Breakpoints identified by 3SEQ shown by the 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. The pink, green, and orange bars show breakpoint-free regions (BFRs), with Region A (nt 13291-19628) showing two trimmed segments to yield Region A’ (nt 13291-14932, 15405-17162, 18009-19628). Region B spans nucleotides 3625-9150, and region C spans 9261-11795. Concatenated region A’BC is non-recombining region 1 (NRR1). Open reading frames are shown above the breakpoint plot, with the variable loop region indicated in the S protein. **Center:** Similarity plot between SARS-CoV-2 and several selected sequences including RaTG13 (black), SARS-CoV (pink), and two pangolin sequences (orange). Shaded region corresponds to the S protein. **Bottom:** Maximum likelihood phylogenetic trees rooted on a 2007 virus sampled in Kenya (BtKy72; root truncated from images), shown for five breakpoint-free regions of the sarbecovirus alignment. Nucleotide positions for phylogenetic inference are 147-695, 962-1686 (first tree), 3625-9150 (second tree, also BFR B), 9261-11795 (third tree, also BFR C), 12443-19638 (fourth tree), and 23631-24633, 24795-25847, 27702-28843, 29574-30650 (fifth tree). Relevant bootstrap values are shown on branches, and gray-shaded regions show sequences that exhibit 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.

**Figure 2.** Maximum likelihood trees of the sarbecoviruses using the two longest breakpoint-free regions (BFRs), rooted on the Kenya/Bulgaria lineage. Region A has been shortened to A’ (5017nt) based on potential recombination signals within the region. Region B is 5525nt long. Sequences are color-coded by province according to the map. Five example sequences with incongruent phylogenetic positions in the two trees are indicated by dashed lines.

**Figure 3.** Phylogenetic relationships among SARS-CoV-2 and closely related sequences for sub-regions of the S protein. SARS-CoV-2 and RaTG13 are the most closely related (their most recent common ancestor node denoted by a green circle), except in the 222nt 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 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 panel) demonstrate 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 NTD, implying that a recombination event between these two sequences in this region is unlikely. Uncertainty measures shown in Extended Data Figure 1.
**Figure 4.** Temporal signal and mean evolutionary rate estimates for HCoV-OC43 ($n=37$), MERS ($n=35$), and SARS ($n=69$) coronaviruses. (A-C) Root-to-tip divergence as a function of sampling time for the three coronavirus evolutionary histories unfolding over different time scales. (D) Mean evolutionary rate estimates plotted against sampling time range for the same three data sets (represented by the same color as the data points in their respective root-to-tip divergence plots), as well as for the comparable recombination-masked sarbecovirus alignment (NRA3) using the two different priors for the rate in the Bayesian inference (red points).

**Figure 5.** Time-measured phylogenetic estimates and divergence times for sarbecovirus lineages using an HCoV-OC43-centered 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 the SARS-CoV virus, and red to SARS-CoV-2. The size of the black internal node circles are proportional to the posterior node support. 95% credible intervals 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 boxplot) and the 2002-2003 SARS-CoV virus (blue boxplot) 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). Box plots show interquartile ranges, white lines are medians, and box whiskers show the full range of the posterior distribution. Transparent bands of the interquartile range width and with the same colours are superimposed to highlight the overlap between the estimates. In Extended Data Figure 4, we compare these divergence time estimates to those obtained using the MERS-CoV-centred rate priors for NRR1, NRR2, and NRA3.
NRR1

$R^2 = 0.19$

$r = -0.0026$

NRR2

$R^2 = 0.14$

$r = -0.0025$

NRA3

$R^2 = -0.015$

$r = -0.00023$
Evolutionary rate (subst./site/yr)

MERS posterior rate
NRR1 posterior rate (MERS-centred prior)
NRR2 posterior rate (MERS-centred prior)
NRA3 posterior rate (OC43-centred prior)
OC43 posterior rate

MERS-centred rate
prior (stddev X 10)
OC43 posterior rate
OC43-centred rate
prior (stddev X 10)