Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination

Spyros Lytras, Joseph Hughes, Darren Martin, Phillip Swanepoel, Arné de Klerk, Rentia Lourens, Sergei L. Kosakovsky Pond, Wei Xia, Xiaowei Jiang, and David L. Robertson

1MRC-University of Glasgow Centre for Virus Research, Glasgow, United Kingdom
2Computational Biology Division, Department of Integrative Biomedical Sciences, University of Cape Town, South Africa
3Division of Neurosurgery, Department of Surgery, Neuroscience Institute, University of Cape Town, South Africa
4Department of Biology, Institute for Genomics and Evolutionary Medicine, Temple University, USA
5National School of Agricultural Institution and Development, South China Agricultural University, Guangzhou, China
6Department of Biological Sciences, Xi’an Jiaotong-Liverpool University (XJTLU), Suzhou, China

*Corresponding author: E-mail: david.l.robertson@glasgow.ac.uk.

Accepted: January 26, 2022

Abstract

The lack of an identifiable intermediate host species for the proximal animal ancestor of SARS-CoV-2, and the large geographical distance between Wuhan and where the closest evolutionary related coronaviruses circulating in horseshoe bats (members of the Sarbecovirus subgenus) have been identified, is fueling speculation on the natural origins of SARS-CoV-2. We performed a comprehensive phylogenetic study on SARS-CoV-2 and all the related bat and pangolin sarbecoviruses sampled so far. Determining the likely recombination events reveals a highly reticulate evolutionary history within this group of coronaviruses. Distribution of the inferred recombination events is nonrandom with evidence that Spike, the main target for humoral immunity, is beside a recombination hotspot likely driving antigenic shift events in the ancestry of bat sarbecoviruses. Coupled with the geographic ranges of their hosts and the sampling locations, across southern China, and into Southeast Asia, we confirm that horseshoe bats, Rhinolophus, are the likely reservoir species for the SARS-CoV-2 progenitor. By tracing the recombinant sequence patterns, we conclude that there has been relatively recent geographic movement and cocirculation of these viruses’ ancestors, extending across their bat host ranges in China and Southeast Asia over the last 100 years. We confirm that a direct proximal ancestor to SARS-CoV-2 has not yet been sampled, since the closest known relatives collected in Yunnan shared a common ancestor with SARS-CoV-2 approximately 40 years ago. Our analysis highlights the need for dramatically more wildlife sampling to: 1) pinpoint the exact origins of SARS-CoV-2’s animal progenitor, 2) the intermediate species that facilitated transmission from bats to humans (if there is one), and 3) survey the extent of the diversity in the related sarbecoviruses’ phylogeny that present high risk for future spillovers.

Keywords: SARS-CoV-2, Sarbecoviruses, bats, origin, COVID-19, host range, coronaviruses, recombination, Rhinolophus, pangolins.

Significance

The origin of SARS-CoV-2 can be unambiguously traced to horseshoe bats, genus Rhinolophus. SARS-related coronaviruses, like SARS-CoV-2, are dispersed over a large geographical area across southern China and Southeast Asia. They have undergone extensive recombination throughout their evolutionary history indicating frequent transmission among their Rhinolophus host species. Breakpoint patterns are consistent with recombination hotspots in the coronavirus genome, particularly upstream of the spike open reading frame with a coldspot in S1. Accounting for these recombination patterns is important when inferring relatedness to SARS-CoV-2.
Introduction

Two years since the emergence of SARS-CoV-2, the origins of this new pandemic human coronavirus remain uncertain. First detected in association with an unusual respiratory disease outbreak in December 2019 in Wuhan city, Hubei province, China (Li, Guan, et al. 2020) no definitive progenitor of animal origin has been identified. The first reports of the initial outbreak were linked to the Huanan animal and seafood market (WHO 2021; Worobey 2021) and, while there are some cases with no identifiable association to this location, this is not so surprising given that so many cases are either mild or asymptomatic (Lytras et al. 2021), and it is possible multiple spillover events at animal markets in Wuhan were involved (Holmes et al. 2021; WHO 2021). Since the 2020 coronavirus pandemic began, both metagenomic and focused sequencing efforts have uncovered a number of viruses related to SARS-CoV-2, retrieved from locations in China and Southeast Asia (Hu et al. 2018; Zhou, Chen, et al. 2020; Zhou, Yang, et al. 2020; Delaune et al. 2021; Li et al. 2021; Wacharapluesadee et al. 2021; Zhou et al. 2021). Several of these sarbecoviruses are recombinants necessitating careful analysis as the presence of mosaic genomes violates the assumption of a single evolutionary history, key to reliable phylogenetic inference from mutation patterns in molecular data.

SARS-CoV-2, responsible for COVID-19, and SARS-CoV, the causative agent of the SARS outbreak in 2002–2003, are both members of the species Severe acute respiratory syndrome-related coronavirus (SARSr-CoV) that forms the sole member of the Sarbecovirus subgenus of Betacoronaviruses (Gorbunova et al. 2020)—a group of viruses which have been primarily found in horseshoe bats (family Rhinolophidae). Coronaviruses are known to recombine with one another during mixed infections (Graham and Baric 2010; Boni et al. 2020). Here, we comprehensively characterize the recombinant nature of the SARS-CoV-2-like coronaviruses sampled so far, focusing specifically on the phylogenetic clade of sarbecoviruses that SARS-CoV-2 is a member of; hereafter referred to as the “nCoV” clade (fig. 1A) (MacLean et al. 2021). To maintain the focus on this clade from which SARS-CoV-2 emerged, we broadly refer to all other Sarbecovirus subclades as ‘non-nCoV’. We present evidence of recombination and several hotspot locations where inferred recombination breakpoints are overrepresented. By comparing the phylogenies inferred for putatively nonrecombinant regions of the genome (i.e., best estimates of SARS-CoV-2 and related sarbecoviruses true evolutionary history) with the viruses’ sampling locations and their host’s geographic range locations, we provide a detailed understanding of the recent evolutionary histories of SARS-CoV-2’s closest known relatives including relative divergence times.

Results and Discussion

Hotspots of Recombination

For a whole-genome alignment of the set of known complete genomes from 78 members of the Sarbecovirus subgenus (including a single representative of SARS-CoV and SARS-CoV-2; supplementary table S1, Supplementary Material online), we performed an initial recombination breakpoint analysis with RDP5 (see Materials and Methods) and identified 160 unique recombination events in all the bat and pangolin-derived virus genomes. To infer a reliable phylogeny of the sarbecoviruses, we removed all regions with evidence for a recombination history from the genome alignment. This reconstructed nonrecombinant phylogeny (fig. 1A) includes a total of 19 nonhuman viruses that comprise the nCoV clade that SARS-CoV-2 is a member of, a sister lineage to the non-nCoV clade SARS-CoV is part of, first emerged from in 2002.

Using the set of breakpoints inferred by RDP5, we tested for significant clustering of recombination events at specific regions of the genome, suggestive of recombination hot- or coldspots. Two permutation-based recombination breakpoint clustering tests were performed: 1) a “breakpoint distribution test” (BDT) that explicitly accounts for the underlying uncertainties in the positions of identified breakpoint positions (Heath et al. 2006) and 2) a “recombinant region test” (RRT) that focuses on point estimates of recombination breakpoint pairs that define recombination events and explicitly accounts for region-to-region variations in the detectability of recombination events (Simon-Loriere et al. 2009). Both tests provided support for the presence of several recombination hotspots: seven in the BDT and nine in the RRT analysis, assuming close locations are giving rise to the same peak (fig. 1B and C), and recombination refractory regions in the NTD and RBD domains of the Spike gene and within open-reading frame (ORF)8 (fig. 1C).

It is possible that all genomic regions where these breakpoint clusters are detected have elevated recombination rates, linked to the molecular mechanisms likely responsible for recombination (Sola et al. 2015). However, simulations of recombination patterns—in genomes with similar degrees of diversity and numbers of detectable recombination events to the genomes analyzed here—indicate that within such a data set we might expect to find, on average, two to three such clusters even in the absence of any recombination hotspots (see Materials and Methods; supplementary table S2, Supplementary Material online). Therefore, none of the identified breakpoint clusters can be definitively attributed to underlying variations in recombination rates at the genome sites where the clusters are identified. Nonetheless, the distribution of recombination breakpoints is clearly nonuniform across the Sarbecovirus genomes, and this nonuniformity is consistent with the presence of recombination hotspots. To independently validate the results of this analysis, we also performed a simple permutation test for clustering in the recombination
breakpoints inferred by the Genetic Algorithm for Recombination Detection (GARD) analysis (see below, supplementary fig. S2, Supplementary Material online). Even though this test would not identify potential hotspots in proximal genomic locations (due to the nature of the GARD method which is expected to identify focused recombination hotspots as a single recombination breakpoint), it confirms the recombination hotspots within the Spike ORF (alignment positions 24174–24648, supplementary fig. S2, Supplementary Material online—consistent with the BDT results, fig. 1C) and at the start of the N ORF (alignment positions 29388–29862, supplementary fig. S2, Supplementary Material online, consistent with both RRT and BDT results, fig. 1C).

Interestingly the pattern of potential hotspots near the Spike ORF has also been noted in previous research (Bobay et al. 2020). Although selective pressure underlying recombinant regions cannot be assessed in this analysis, antigenic selection—for immune escape—and/or selection associated with switches in host receptor specificity and efficiency—that is, antigenic shift—are two probable candidate drivers of the observed recombination patterns, consistent with the known immunodominance of the Spike NTD and RBD regions (Walls et al. 2020). It is clearly important to account for these complex recombination patterns when examining the evolutionary history of these pathogens, since multiple evolutionary histories can be inferred from the single whole-genome alignment. As SARS-CoV-2 continues circulating in humans and mutations increase its sequence diversity, identifying SARS-CoV-2 recombination events will become easier and increasingly more important to monitor (Jackson et al. 2021).

Recombination Patterns between SARS-CoV-2 Relatives

To reconstruct a reliable phylogeny for a set of viruses, sufficient information needs to be present in the underlying sequence alignment. Thus, even though a whole-genome alignment can be split into shorter subalignments with the aim of getting rid of all independent recombination events, it is unlikely that all subalignments can produce reliable phylogenies. To overcome this trade-off, we performed a secondary, more conservative, recombination analysis using GARD and identified the locations of 21 recombination breakpoints that strongly impact the inferred phylogenetic relationships of the analyzed sequences when mosaic patterns are ignored.
estimates (Boni et al. 2020; MacLean et al. 2021; Wang et al. 2021). Two more bat samples from Yunnan, RShSTT182 and RacCS203 (Wacharapluesadee et al. 2021). These viruses are, after the China sampled CoVs mentioned above, the next closest relatives to SARS-CoV-2 with common ancestor age estimates (using RBP region 5) around 1907 (95% HPD: 1873–1938) and 1883 (95% HPD: 1841–1921), respectively (fig. 4A). Similar to the other nCoV viruses, the recombination analysis uncovers more intricate phylogenetic relations for some parts of the genome. Notably, RShSTT182 and RShSTT200, despite being sampled in Cambodia, cluster with RatGI3 for RBP regions 8 and 9 (fig. 2A and supplementary fig. S1, Supplementary Material online), while in RBP region 4 of the genome RacCS203, from Thailand, clusters together with SARS-CoV-2 within the Yunnan clade (fig. 2A). This indicates that cocirculation and recombination between these viruses in the last few centuries is responsible for the observed patterns in their inferred evolutionary history, despite the current geographic range of at least 2,500km. This wide distribution of related viruses, including shared recombination breakpoints, highlights an important feature of bat species: Their frequently overlapping/sympatric ranges will provide ample opportunities for transmissions of viral variants from one bat species (or subspecies) to another.

Consistent with the Spike S1 recombination hotspots revealed in the initial analysis (fig. 1B and C), closest relatives of SARS-CoV-2 presented here have non-nCoV derived recombination sequences at the start of the Spike gene (fig. 2B). Despite one collected from Yunnan, China and the other from Thailand, viruses RmYN02 and RacCS203 share a closely related non-nCoV sequence in RBP regions 15 and 16 (encompassing the Spike NTD and RBD, respectively; fig. 2B).
**Fig. 2.**—Nonrecombinant topologies of SARS-CoV-2 relatives. Zoomed in regions of selected RBP region maximum likelihood phylogenies (A). Branches within the nCoV clade are colored in red and outside the nCoV clade in green. Genome schematics of close SARS-CoV-2 relatives with recombinant Spike regions (B). RBP regions 15 and 16 are highlighted and the non-nCoV subclades of the maximum likelihood phylogenies containing the relevant viruses are presented. The coloring of nonrecombinant segments indicates patristic distance to SARS-CoV-2 (see fig. 3 legend). Nodes with bootstrap confidence values below 80% have been collapsed.
having a distinct RBD compared with that of SARS-CoV-2. On the other hand, viruses RpYN06, PrC31, CoVZC45, and CoVZXC21 cluster within the nCoV clade for region 15 but, similar to the RmYN02 and RacCS203, form a distinct cluster in the non-nCoV clade for region 16 (fig. 2B; Wells et al. 2021). We speculate that some of the apparent patterns of recombination-mediated exchange between nCoV and non-nCoV viruses can be partly explained by sequential recombination, that is, “overprinting” of recombination events involving different ancestral parental viruses. This will occur when an nCoV virus acquires a non-nCoV genomic sequence through ancestral recombination but its progenitors cocirculating with other nCoV viruses incur subsequent recombination events that overlap portions of the original non-nCoV recombinant sequence, producing the more complex “patchy” patterns we see in the currently sampled viruses. Note, overprinting of recombination regions will result in reduced confidence in the breakpoints at deeper nodes in the phylogeny.

The finding that Sunda (also known as Malayan) pangolins, Manis javanica, nonnative to China, are the other mammal species from which nCoV sarbecoviruses have been sampled in Guangxi and Guangdong provinces in South China (Lam et al. 2020; Xiao et al. 2020), indicates these animals are likely being infected in this part of the country (fig. 3B). Pangolins are one of the most frequently trafficked animals with multiple smuggling routes leading to southern China (Xu et al. 2016). The most common routes involve moving the animals from Southeast Asia (Myanmar, Malaysia, Laos, Indonesia, Vietnam) to Guangxi, Guangdong, and Yunnan. The most likely scenario that is consistent with both the reported respiratory distress that the sampled pangolins exhibited (Liu et al. 2019; Xiao et al. 2020) and the lack of confirmed CoV infections among Sunda pangolins in Malaysia (Lee et al. 2020), is that the viruses obtained from these animals infected them (presumably from bat sources) after they were trafficked into southern China. Still, serological data of trafficked Sunda pangolins could suggest potential circulation of sarbecoviruses in

---

**Fig. 3.**—Recombination analysis and geographic distribution of Sarbecoviruses. Maximum clade credibility (MCC) dated phylogeny of RBP region 5 of 78 Sarbecoviruses (A). All tips are annotated with the geographic region the viruses have been sampled in and notable viruses are annotated with genome schematics separated into the 22 inferred RBP regions, each colored based on phylogenetic distance from SARS-CoV-2 (see scale and Materials and Methods). RBP region 21 has been removed from the schematic due to limited phylogenetic information in the alignment. The GX cluster annotated with an asterisk contains the five pangolin coronaviruses collected in Guangxi. Map of East Asia with geographic regions (provinces within China, countries outside China) colored based on Sarbecoviruses sampling (B): blue for regions with only non-nCoV clade samples, pink for regions where nCoV viruses have been sampled. Shading in the nCoV regions corresponds to phylogenetic distance from SARS-CoV-2 (see scale). Notable nCoV viruses and pangolin trafficking routes (adapted from Xu et al. [2016]) are annotated onto the map.
**FIG. 4.**—Molecular dating and *Rhinolophus* host geographic distributions. Tip-dated Bayesian phylogeny of RBP region 5 showing the nine closest relatives to SARS-CoV-2 (A). Tree nodes have been adjusted to the mean age estimates and posterior distributions are shown for each node with mean age estimate and 95% HPD confidence intervals presented to their left. Tips are annotated with the host species they were sampled in, bat silhouette colors correspond to panel (B). Geographic ranges of *Rhinolophus* species the SARS-CoV-2 closest relatives have been sampled in (B). Maps are restricted to East Asia and separated into province-level within China and country-level outside China.
the animals' wild populations (Wacharapluesadee et al. 2021).

Although the recombination patterns inferred in the pangolin-derived virus genomes seem to be less complex than those of the bat nCoV genomes, the Guangdong Pangolin-CoV has a Spike receptor binding domain that is most similar to that of SARS-CoV-2. This finding was highlighted by Li, Giorgi, et al. (2020) and attributed to recombination between the SARS-CoV-2 and Pangolin-CoV proximal ancestors. However, based on the nucleotide divergence between the two viruses in this short Spike segment, the most likely explanation is recombination in an ancestor of RaTG13, making it more divergent than Pangolin-CoV compared with SARS-CoV-2 (Boni et al. 2020) (reflected in region 17, fig. 2A). The susceptibility of pangolins to an apparently new human coronavirus is not surprising given the well-documented generalist nature of SARS-CoV-2 (Conceicao et al. 2020), which has been found to readily transmit to multiple mammals with similar ACE2 receptors, most notably, on mink farms (Oude Munnink et al. 2021).

Overlapping Horseshoe Bat Ranges
Considering that almost all sarbecoviruses have been sampled in related horseshoe bat host species, with ranges that span different regions where nCoV clade viruses have been collected (fig. 4B), these bat populations should be prioritized for sampling. For example, the least horseshoe bat species, Rhinolophus pusillus, is sufficiently dispersed across China to account for the geographical spread of: 1) bat sarbecovirus recombinants in the West and East of China, 2) infected imported pangolins in the South, 3) bat sarbecovirus recombinant links to southwest of China, and 4) SARS-CoV-2 emergence toward Hubei in Central China (fig. 3B). Strikingly, the ranges of multiple species including Rhinolophus affinis, Rhinolophus sinicus, and R. pusillus overlap all the regions in China where nCoV members have been collected (fig. 4B). Other species known to harbor nCoV viruses have more restricted ranges such as Rhinolophus malayanus found predominantly in the western part of China and countries to the Southwest of China (Myanmar, Thailand, Cambodia, Laos, Viet Nam, and Peninsular Malaysia) (Piraccini 2016; Bates et al. 2019). On the contrary, the greater horseshoe bat species, Rhinolophus ferrumequinum, is not known to harbor any nCoV viruses and is absent from large parts of South Central China (fig. 4B).

The wide geographic ranges of R. pusillus and R. affinis and the fact that two of the closest known relatives of SARS-CoV-2, RpYN06, and RaTG13, have been sampled in these species flags them as prime suspects for the source of the SARS-CoV-2’s progenitor in China. Additionally, these two bat species are found in shared roosts with R. sinicus and R. ferrumequinum in Yunnan and with R. sinicus in Guangxi (Luo et al. 2013), providing opportunities for host switches, coinfestions, and thus recombination between the sarbecoviruses that these bat species carry. Rhinolophus pusillus and R. affinis also link more regions of China with bat species such as R. shameli, R. malayanus, and R. acuminatus which are only found in Southeast Asia and southwest of China (fig. 4B). Latinne et al. (2020) published a large-scale sampling expedition of coronaviruses across bats in China. Despite there only being short RdRp sequence fragments available, the phylogeny for the novel viruses revealed a cluster of seven identical sarbecovirus sequences sampled from R. affinis within the nCoV clade (supplementary fig. S3, Supplementary Material online). Still, the fact that viruses in the Yunnan clade (consisting of RmYN02, RpYN06, and PrC31) were sampled from three different Rhinolophus species supports the hypothesis that these viruses readily infect multiple different horseshoe bat species with overlapping geographical ranges.

Based on the analysis of the sarbecovirus and host data presented here, we propose that to locate the SARS-CoV-2 progenitor sampling should focus on the ranges of horseshoe bat host populations known to harbor nCoV viruses. Specifically, samples should be collected in roosting environments spread across China with care taken both to avoid a further spillover (or reverse zoonosis) and to protect the bat populations (Luo et al. 2013). Sampling strategies will also need to consider the distinct subspecies of Rhinolophus as the delineators of genetically meaningful host populations for coronaviruses, for example, there are two R. affinis subspecies on mainland China: himalayanus and macrurus (Mao et al. 2010). Future sampling should also encompass a range of indigenous mammals other than bats that we now know can be infected by these coronaviruses. Although highly endangered, Chinese pangolins, given their susceptibility to infection and their geographical range across southern China (Challender et al. 2019), could be one of the possible candidates for the “missing” intermediate host of the SARS-CoV-2 proximal ancestor (WHO 2021).

Conclusions
The currently available data, although sparse, illustrate a complex reticulate evolutionary history involving the lineage of sarbecoviruses SARS-CoV-2 emerged from. This history is governed by cocirculation of related coronaviruses, over at least the last 100 years, across the bat populations in southern China, and into Southeast Asia with multiple recombination events imprinted on the genomes of these viruses. Considering the high frequency of recombination, it is expected that selection could preferentially favor exchanges of specific genomic regions, in line with our detection of hotspots near the Spike gene (fig. 18 and C). The functional implications of selective Spike recombination has recently been corroborated by multiple independent studies, suggesting this might be a mechanism for antigenic shift utilized by the sarbecoviruses or, more broadly, by all coronavirus groups.
et al. 2021; Xiao et al. 2021) could have facilitated transmission in the other nCoV sarbecoviruses (figs. 1A and 3A).

The evidence of recombination events between members of the Sarbecovirus subgenus sampled in different geographical regions and from different bat hosts, indicates recent extensive movement of the viruses between different regions and species (and subspecies) as a result of the continued contacts between different bat populations that carry them. Although the closest known relatives of SARS-CoV-2 were sampled in Yunnan, the location of the proximal viral population SARS-CoV-2 emerged from remains unknown. The recombination patterns detected within the nCoV genomes imply the existence of one or several primary reservoir hosts with a geographical range spanning Thailand from the Southwest and Zhejiang to the East, a distribution that is consistent with specific Chinese horseshoe bats acting as the primary reservoir hosts. Our observations are further confirmed by a recent report of more bat coronaviruses very closely related to SARS-CoV-2 sampled from R. pusillus and R. malayanus in Laos (Temmam et al. 2022). Both the sampling location and host species are consistent with expectations based on our analysis, essentially filling in the geographic gap between previous nCoV sampling locations. The recombination patterns reported in these newly discovered genomes are also consistent with the extensive recombination reported here (Temmam et al. 2022). Having presented evidence in support of R. affinis and R. pusillus’s potential significance as the reservoir species, we would be remiss not to note that at least 20 different Rhinolophus species are distributed across China (four of which are endemic to China), many of which have not yet been found hosting nCoVs. The generalist nature of Sarbecoviruses also means multiple wild or farmed animals (e.g., American mink [Neovison vison] both farmed for fur and used as a food source) (WHO 2021; Xia et al. 2021; Xiao et al. 2021) could have facilitated transmission of SARS-CoV-2 from bats to humans.

Although SARS-like antibodies detected in people from rural communities in China (Wang et al. 2018; Li et al. 2019) indicates an intermediate animal species is potentially not required for transmission to humans, it does seem that emergence in a populated area is required for significant outbreaks to occur. The association of both SARS-CoV and SARS-CoV-2 with animal markets suggests animal trafficking and selling is a key part of this transmission to humans. Human-mediated animal movement increases contact with sarbecovirus infected animals (whether they are susceptible species that have been trapped or farmed in rural locations; Xia et al. 2021) and subsequently introduces them into city markets (Lytras et al. 2021; WHO 2021; Worobey 2021). Urgent questions relating to the prevention of another emergence are: where in China or Southeast Asia is the SARS-CoV-2 progenitor located (our analysis shows this is not necessarily Yunnan), which bat or other animal species are harboring nCoV sarbecoviruses and linked to this what is the risk of future spillover events? There is undoubtedly a virus highly related to SARS-CoV-2 still present somewhere in the wild. To maximize the probability that future sampling efforts will uncover this host species or subspecies we need a wide and systematic sampling strategy of horseshoe bats.

**Materials and Methods**

**Genome Alignment**

The whole-genome sequences of the 78 Sarbecovirus members used in this analysis (supplementary table S1, Supplementary Material online) were aligned and the ORF of the major protein-coding genes were defined based on SARS-CoV-2 annotation (Wu et al. 2020). Codon-level alignments of the ORFs were created using MAFFT v7.453 (Katoh and Standley 2013) and PAL2NAL (Suyama et al. 2006). The intergenic regions were also aligned separately using MAFFT and all alignments were pieced together into the final whole-genome alignment and visually inspected in Bioedit (Hall 1999).

**Genome-Specific Recombination Analysis**

We first performed an analysis for detecting unique recombination events specific to individual genome sequences using the RDP (Martin and Rybicki 2000), GENECONV (Padidam et al. 1999), BOOTSCAN (Martin et al. 2005), MAXCHI (Smith 1992), CHIMAERA (Posada and Crandall 2001), SISCAN (Gibbs et al. 2000), and 3SEQ (Bonj et al. 2007) methods implemented in the program RDP5 (Martin et al. 2021). Default settings were used throughout except: 1) only potential recombination events detected by three or more of the above methods, coupled with phylogenetic evidence of recombination were considered significant and 2) sequences were treated as linear. We required evidence from three or more of the recombination signal detection methods because none of three methods query the same recombination signals and all have varying power to detect recombination in data sets with different degrees of sequence diversity (Posada and Crandall 2001; Posada 2002). The recombinant sequence identification, recombination breakpoint verification, and shared recombination event verification steps used are outlined in Martin et al. (2017). The approximate breakpoint positions and recombinant sequence(s) inferred for every potential recombination event, were manually checked and adjusted where necessary using the phylogenetic and recombination signal analysis features.
The SANTA-SIM settings used were: population size 5,000, inoculum = all, mutation rate = 3.5 \times 10^{-5}, rate bias matrix = (0.42, 2.49, 0.29, 1.73, 0.23, 4.73, 6.99, 9.20, 0.60, 1.02, 2.56, 0.88), dual infection probability = 0.1, background recombination probability = 0.06, and generation number = 5,000. Simulated recombination events had a maximum of two breakpoints: a setting that required the use of a slightly modified version of SANTA-SIM that can be obtained from https://github.com/phillipswanepoel/santa-sim/tree/Recomb_and_align. Whereas one of the four data sets had no simulated recombination hotspots, the other three each had a single 100-nt-long hotspot between alignment positions 6000 and 6100 wherein recombination frequencies were 4\times, 8\times, or 16\times higher than the background level.

All data sets were analyzed for recombination by RDP5 without any supervision, and RRT and BDT plots were produced for each data set (all with the same program settings used to analyze the actual sarbecovirus data set).

The true positive rate of the BDT was estimated as the proportion of 200-nt windows centered on nucleotides between positions 6000 and 6100, that is, within the simulated hotspot, that contained a number of breakpoints greater than the upper bound of the 99% confidence interval of the breakpoint clustering distribution expected under random recombination (e.g., indicated by the light gray areas of the plots in fig. 1C). Since a 200-nt sliding window was used for both breakpoint clustering tests, all windows overlapping with the hotspot (positions 5801 to 6299) were ignored when determining the BDT and RRT false positive rates. The false positive rate of BDT was calculated as the proportion (across all 100 simulated alignments of each of the four data sets) of the examined 200-nt windows centered on nucleotides outside region 5801–6299 that contained a number of breakpoints greater than the upper bound of the 99% confidence interval of the breakpoint clustering distribution expected under random recombination.

Similarly, the true positive rate of the RRT was estimated as the proportion, across all 100 simulated alignments in a data set, of 200-nt windows centered on nucleotides between positions 6000 and 6100, that is, within the simulated hotspot, that had associated breakpoint clustering permutation \( P \) values <0.01 (e.g., indicated by the upper bound of the light gray area of the plot in fig. 1B). The RRT false positive rate was calculated as the proportion, across all 100 simulated alignments in a data set, of the examined 200-nt windows centered on nucleotides outside region 5801–6339 that had associated permutation \( P \) values <0.01.

The true and false positive rates for BDT and RRT with respect to identifying the presence of the simulated recombination hotspots are indicated in supplementary table S2, Supplementary Material online. Note that, due to the nature of the simulations, it was not guaranteed that even with perfect recombination detection power and accuracy: 1) the recombination hotspot regions would contain any detectable excess of recombination breakpoints, and 2) the “normal” genome regions would contain no breakpoint clusters. What these simulations capture is the power of the two clustering tests to indirectly infer the locations of actual recombination hotspot regions that, due to chance during the simulation process, might not even contain any detectable recombination breakpoints. Nevertheless, as expected, the hotspot detection power of both BDT and RRT increases substantially with the intensity of the simulated recombination hotspots: from \( \sim 10\% \) for both tests with a 4\times increase in

Recombination Hotspot Analysis

The distribution of 236 unambiguously detected breakpoint positions defining 160 unique recombination events based on the RDP5 analysis described above were analyzed for evidence of recombination hotspots and coldspots using the permutation-based RRT (Simon-Loriere et al. 2009) and BDT (Heath et al. 2006). The RRT accounts for site-to-site variations in the detectability of individual recombination events and examines the distribution of point estimates of pairs of breakpoint locations bounding each of the unique recombination events detected by RDP5. Rather than using point estimates of recombination breakpoint locations, the BDT accounts for uncertainties in the estimation of individual breakpoint locations as determined from the state transition likelihoods yielded by the hidden Markov model-based recombination breakpoint detection method, BURT (described in the RDP5 program manual at http://web.cbio.uct.ac.za/~darren/rdp.html).

To verify whether the recombination breakpoint clusters detected with these tests were consistent with the presence of recombination hotspots, we simulated recombination with SANTA-SIM (Jariani et al. 2019). Four data sets of 100 \times 10,000 long sequences that had: 1) approximately the same degree of genetic diversity as the analyzed sarbecovirus alignment and 2) approximately the same numbers of detectable recombination events and recombination breakpoints per nucleotide as those detected in the analyzed sarbecovirus alignment. The SANTA-SIM settings used were: population size = 4,500, inoculum = all, mutation rate = 3.5 \times 10^{-5}, rate bias matrix = (0.42, 2.49, 0.29, 1.73, 0.23, 4.73, 6.99, 9.20, 0.60, 1.02, 2.56, 0.88), dual infection probability = 0.1, background recombination probability = 0.06, and generation number = 5,000. Simulated recombination events had a maximum of two breakpoints: a setting that required the use of a slightly modified version of SANTA-SIM that can be obtained from https://github.com/phillipswanepoel/santa-sim/tree/Recomb_and_align. Whereas one of the four data sets had no simulated recombination hotspots, the other three each had a single 100-nt-long hotspot between alignment positions 6000 and 6100 wherein recombination frequencies were 4\times, 8\times, or 16\times higher than the background level.

All data sets were analyzed for recombination by RDP5 without any supervision, and RRT and BDT plots were produced for each data set (all with the same program settings used to analyze the actual sarbecovirus data set).

The true positive rate of the BDT was estimated as the proportion of 200-nt windows centered on nucleotides between positions 6000 and 6100, that is, within the simulated hotspot, that contained a number of breakpoints greater than the upper bound of the 99% confidence interval of the breakpoint clustering distribution expected under random recombination (e.g., indicated by the light gray areas of the plots in fig. 1C). Since a 200-nt sliding window was used for both breakpoint clustering tests, all windows overlapping with the hotspot (positions 5801 to 6299) were ignored when determining the BDT and RRT false positive rates. The false positive rate of BDT was calculated as the proportion (across all 100 simulated alignments of each of the four data sets) of the examined 200-nt windows centered on nucleotides outside region 5801–6299 that contained a number of breakpoints greater than the upper bound of the 99% confidence interval of the breakpoint clustering distribution expected under random recombination.

Similarly, the true positive rate of the RRT was estimated as the proportion, across all 100 simulated alignments in a data set, of 200-nt windows centered on nucleotides between positions 6000 and 6100, that is, within the simulated hotspot, that had associated breakpoint clustering permutation \( P \) values <0.01 (e.g., indicated by the upper bound of the light gray area of the plot in fig. 1B). The RRT false positive rate was calculated as the proportion, across all 100 simulated alignments in a data set, of the examined 200-nt windows centered on nucleotides outside region 5801–6339 that had associated permutation \( P \) values <0.01.

The true and false positive rates for BDT and RRT with respect to identifying the presence of the simulated recombination hotspots are indicated in supplementary table S2, Supplementary Material online. Note that, due to the nature of the simulations, it was not guaranteed that even with perfect recombination detection power and accuracy: 1) the recombination hotspot regions would contain any detectable excess of recombination breakpoints, and 2) the “normal” genome regions would contain no breakpoint clusters. What these simulations capture is the power of the two clustering tests to indirectly infer the locations of actual recombination hotspot regions that, due to chance during the simulation process, might not even contain any detectable recombination breakpoints. Nevertheless, as expected, the hotspot detection power of both BDT and RRT increases substantially with the intensity of the simulated recombination hotspots: from \( \sim 10\% \) for both tests with a 4\times increase in
simulated breakpoint probabilities within the 100-nt hotspot region to ~60% for a 16× increase in breakpoint probabilities within the hotspot region. It is also noteworthy that the false positive rates for both tests are likely between 1.5 and 2× higher than the expected rate of 0.01 (which is expected given that the windows containing breakpoint clusters exceeding the 99% confidence interval were used to identify breakpoint hotspots). This false positive rate may not seem very high but, for a long alignment such as that examined for the sarbecoviruses that can be broken into many non-overlapping 200-nt windows, it indicates that for such an alignment we might expect to find on average two to three significant clusters of breakpoints that are in fact not associated with any elevation in the underlying recombination rate.

Whole-Genome Alignment Recombination Analysis

Next, we sought to conservatively examine the entire genome alignment for the subset of recombination breakpoints that had the largest impacts on the inferred evolutionary relationships between the analyzed sarbecoviruses using the GARD method (Kosakovsky et al. 2006) implemented in Hyphy v2.5.29 (Kosakovsky Pond et al. 2020). Model goodness of fit was evaluated using the small sample Akaike Inference Criterion (c-AIC) (Akaike 1998). To improve computational efficiency and statistical efficiency (as GARD requires more statistical evidence of recombination for larger phylogenies, and the minimal length of detectable nonrecombinant fragments increases with the number of sequences) and focus on the closest relatives of SARS-CoV-2, 22 of the 78 viruses that are closest to SARS-CoV-2 or had preliminary evidence of clustering near detected interclade recombinants were included in the GARD analysis (supplementary table S1, Supplementary Material online). Only breakpoints present in more than 2/3 of the 64 GARD consecutive step-up models were retained to produce a final set of 21 likely breakpoints (positions corresponding to the SARS-CoV-2 reference genome Wuhan-Hu-1 in order: 1680, 3093, 3649, 4973, 8208, 11445, 12622, 14401, 15954, 16923, 19965, 20518, 21198, 21411, 22460, 23396, 24144, 24843, 26323, 27388, 27685). Based on these the whole-genome alignment was split into 22 RBP regions. The position of each region on the alignment and relative to the SARS-CoV-2 genome as well as their length is presented in supplementary table S3, Supplementary Material online.

We further used the GARD recombination analysis to validate the RDP5 recombination hotspot analysis. We performed a permutation test of breakpoint clustering by fixing the number of all inferred breakpoints (64) and the location of the 13,550 variable sites in the alignment. Then defined a sliding window so that each window would have an average of one breakpoint in it (alignment length/64) producing 474 windows. N = 10,000 replicates were drawn where 64 variable sites were randomly chosen from one of the breakpoints. For each sliding window, we tabulated the distribution of randomly drawn breakpoints in the window. Two hotspots and 17 coldspot windows were identified, presented in supplementary figure S2, Supplementary Material online. This analysis is not expected to produce results identical to the RDP5-based hotspot analysis, since the GARD method does not distinguish between potential breakpoints in very near genomic proximity, so this post hoc test is unlikely to identify clustering of unique breakpoints that are very close to one another (in contrast to the RDP5 approach).

Phylogenetic Reconstruction

The phylogeny of each RBP alignment region based on the GARD analysis and the nonrecombinant whole-genome based on the RDP5 analysis were reconstructed using iqtree version 1.6.12 (Nguyen et al. 2015) under a general time reversible (GTR) substitution model assuming invariable sites and a four-category Γ distribution. Tree node confidence was determined using 10,000 ultrafast bootstrap replicates.

Based on the nonrecombinant whole-genome phylogeny, 20 viruses form a monophyletic nCoV clade (fig. 1A). To illustrate the distance of each virus from SARS-CoV-2 for each GARD determined genomic region, we defined the nCoV clade on each phylogeny as the subset of the aforementioned 20 nCoV viruses forming a monophyly with SARS-CoV-2 in each phylogeny. The rest of the viruses were classified as members of the non-nCoV clade for each RBP region. We then used an arbitrary tip distance scale normalized between all phylogenies so distances are comparable between regions. For each maximum likelihood tree, the patristic distance between each tip and SARS-CoV-2 is calculated using ETE 3 (Huerta-Cepas et al. 2016) as d1 for members of the nCoV clade and d2 for members of the non-nCoV clade. The distances are then normalized so that for nCoV clade members the most distant tip from SARS-CoV-2 is 1.1 for non-nCoV clade members the closest non-nCoV virus to SARS-CoV-2 and −1.1 the most distant, as follows:

\[
d'_1 = 1.1 - \frac{d_1}{d_{1,max}} (1 : nCoV)
\]

\[
d'_2 = -0.1 - \frac{d_2 - d_{2,min}}{d_{2,max} - d_{2,min}} (2 : non-nCoV).
\]

With \(d'_1\) and \(d'_2\) being the normalized values for each clade, variables denoted with “min” being the smallest distance and variables denoted with “max” being the largest distance in each given set.

Phylogenies were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and ETE 3 (Huerta-Cepas et al. 2016).
Molecular Dating

To provide temporal information to the phylogenetic history of the viruses, we performed a Bayesian phylogenetic analysis on RBP region 5, using BEAST v1.10.4 (Suchard et al. 2018). This region was selected due to its length, being one of the two longest nonrecombinant regions in the analysis (3,238 bp), and because all 20 nCoV viruses form a monophyle in the respective tree. Based on the observation of an increased evolutionary rate specific to the deepest branch of the nCoV clade reported in MacLean et al. (2021) (MacLean et al. 2021), we adopted the same approach of fitting a separate local clock model to that branch from the rest of the phylogeny. A normal rate distribution with mean $5 \times 10^{-4}$ and SD $2 \times 10^{-4}$ was used as an informative prior on all other branches. The lineage containing the BtKY72 and BM48-31 bat viruses was constrained as the outgroup to maintain overall topology. Codon positions were partitioned and a GTR $+$ I substitution model was specified independently for each partition. The maximum likelihood phylogeny reconstructed previously for RBP region 5 was used as a starting tree (rooted at the BtKY72 and BM48-31 clade). A constant size coalescent model was used for the tree prior and a lognormal prior with a mean of 6 and SD of 0.5 was specified on the population size. Two independent MCMC runs were performed for 500 million states for the data set. The two chains were inspected for convergence and combined using LogCombiner (Drummond and Rambaut 2007) using a 10% burn-in for each chain. The effective sample size for all estimated parameters was above 200.

Host Range Data

All host ranges presented in figure 4B were retrieved from the IUCN Red List of Threatened Species (https://www.iucnredlist.org/) and the Mammals of China (Princeton Pocket Guide) (Hoffmann et al. 2013). Geographic visualization was performed using D3 and JavaScript in Observable (https://observablehq.com/).

Supplementary Material

Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments

We would like to thank all the authors who have kindly deposited and shared genome data on GISAID. Credit also needs to be given to the surveillance projects for generating the genome data that is available in GenBank and to the software developers for making the tools we have used freely available. A table with genome sequence acknowledgments can be found in the Supplementary Material online (supplementary table S1, Supplementary Material online).

Data Availability

The data underlying this article are available in GenBank and GISAID, see Supplementary table S1, Supplementary Material online for accession numbers. The whole-genome alignment and phylogenetic tree files associated with this work can be found in the following online repository: https://github.com/spyros-lytras/SC2_origins_rec.

Literature Cited


Associate editor: Adi Stern