bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Exploring the natural origins of SARS-CoV-2

Spyros Lytras[1], Joseph Hughes[1], Wei Xia[2], Xiaowei Jiang[3], David L Robertson[1,*]

[1]MRC-University of Glasgow Centre for Virus Research (CVR), Glasgow, UK.
[2]National School of Agricultural Institution and Development, South China Agricultural University, Guangzhou, China.
[3]Department of Biological Sciences, Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China.

[*]Correspondence: david.l.robertson@glasgow.ac.uk

Summary. The lack of an identifiable intermediate host species for the proximal animal ancestor of SARS-CoV-2 and the distance (~1500 km) from Wuhan to Yunnan province, where the closest evolutionary related coronaviruses circulating in horseshoe bats have been identified, is fueling speculation on the natural origins of SARS-CoV-2. Here we analyse SARS-CoV-2's related horseshoe bat and pangolin *Sarbecoviruses* and confirm *Rhinolophus affinis* continues to be the likely reservoir species as its host range extends across Central and Southern China. This would explain the bat *Sarbecovirus* recombinants in the West and East China, trafficked pangolin infections and bat *Sarbecovirus* recombinants linked to Southern China. Recent ecological disturbances as a result of changes in meat consumption could then explain SARS-CoV-2 transmission to humans through direct or indirect contact with the reservoir wildlife, and subsequent emergence towards Hubei in Central China. The only way, however, of finding the animal progenitor of SARS-CoV-2 as well as the whereabouts of its close relatives, very likely capable of posing a similar threat of emergence in the human population and other animals, will be by increasing the intensity of our sampling.

Keywords. SARS-CoV-2, *Sarbecoviruses*, bats, origins, host range, coronaviruses, recombination, China, Cambodia, *Rhinolophus,* pangolins

One year since the emergence of SARS-CoV-2, the origins of this new pandemic human coronavirus remains an apparent mystery. First detected in association with an unusual respiratory disease outbreak in December 2019 at a wet market in Wuhan city, Hubei province, China (Li et al. 2020) no definitive animal progenitor has been identified. Environmental samples taken from the Huanan Seafood Wholesale Market in question have only revealed evidence for human infections and the finding of cases with no linkage to this location suggests it may not be 'ground zero' for the SARS-CoV-2 spillover event. This coupled with the distance (~1500 km) from Wuhan to Yunnan province where the closest evolutionary related coronaviruses circulating in horseshoe bats have been identified (Zhou et al. 2020), has fed, without evidence, a conspiracy theory about the origins of SARS-CoV-2 being the Wuhan Institute of Virology.

Without question SARS-CoV-2 is a member of the *Sarbecovirus* subgenus of *Betacoronaviruses* found in horseshoe bat hosts (family *Rhinophilidae*) and a sister lineage of SARS-CoV (Figure 1A), the causative agent of the SARS outbreak in 2002-3 (Gorbalenya et al. 2020). Here, we focus on the broader set of 'nCoV' *Sarbecoviruses* that cluster with SARS-CoV-2 in phylogenetic analysis and perform recombination detection analysis on a whole genome alignment of all the available *Sarbecoviruses* (Figure 1A). This identified 16 recombination breakpoints that can be used to split the alignment into 17 putatively non-recombinant genomic regions from which evolutionary history can be inferred. To clearly characterise the recombination patterns between viruses in the same clade as SARS-CoV-2 and those in the sister clade, which includes SARS-CoV, we have attributed each virus in

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

each of the 17 regions to either being in the nCoV clade or the non-nCoV clade (closer to SARS-CoV), similarly defined in MacLean et al. (2020).

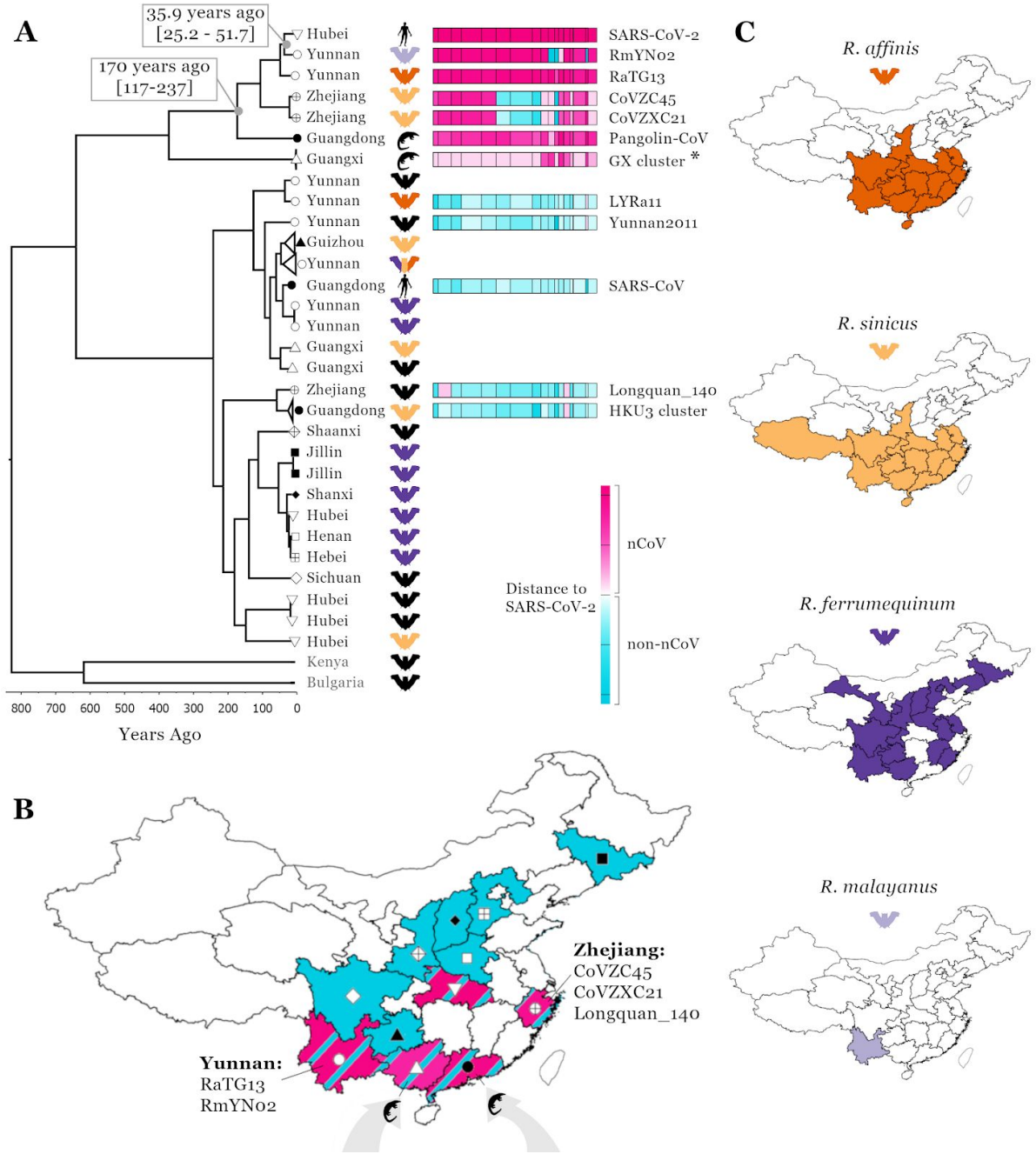

Figure 1. Linking SARS-CoV-2 related *Sarbecovirus* to geography and host ranges. Recombination analysis on a whole genome alignment of 69 *Sarbecoviruses* identified 16 recombination breakpoints (A), see methods. The fourth region was used for a molecular clock analysis (see scale below phylogenetic tree) and the median node age is presented for two key nodes including 95% HPD intervals; a representative set of 32 viruses are shown in the phylogenetic tree. To illustrate recombination patterns between viruses in the same clade as SARS-CoV-2 and those in other clades of the tree, we have attributed each region of each virus to either being in the nCoV clade (the clade SARS-CoV-2 is found in, pink) or the non-nCoV clade (closer to SARS-CoV, blue). The colour shade corresponds to genetic distance, see key. The 'GX cluster' (asterisk) includes five viruses sampled in Sunda pangolins in Guangxi (P1E, P2V, P4L, P5E, P5L; see Table S1). A map of China (B) is shown with colours corresponding to regions *Sarbecoviruses* have been sampled: blue, the non-nCoV clade and pink, the nCoV clade. Symbols correspond to provinces at tree tips in A. Host ranges of four potential hosts of the proximal SARS-CoV-2 ancestor (C); ranges from Smith and Xie (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While the two genetically closest relatives to SARS-CoV-2 identified so far are the bat *Sarbecovirus*es RaTG13 and RmYN02 (Zhou et al. 2020; 2020b), both recombinants from samples collected in Yunnan (Figure 1B), they are estimated to have shared a common ancestor with SARS-CoV-2 about 40/50 years ago (Boni et al. 2020; Wang et al. 2020; MacLean et al. 2020) so are too distant to be SARS-CoV-2's progenitors. Importantly, three recombinant bat *Sarbecovirus*es, CoVZC45, CoVZXC21 and Longquan_140, the next closest SARS-CoV-2 relatives in the nCoV clade (for most of their genomes for CoVZC45 and CoVZXC21, except for four regions on Orf1ab and Spike, and for two parts of Longquan_140's genome) were all sampled in Zhejiang a coastal province in Eastern China (Hu et al. 2018; Lin et al. 2017) (Figure 1 and S3). Interestingly, in the second nCoV part of Longquan_140's genome, the HKU3 set of closely related bat *Sarbecoviruses* sampled in Hong Kong (bordering Guangdong province) also cluster within the nCoV clade. Longquan_140 and HKU3 also share a recent common ancestor for most parts of their genomes, suggesting that this recombinant region was acquired in their shared ancestor.

This high prevalence of recombination among *Sarbecoviruses*, the bringing together of evolutionary divergent genome regions in co-infected hosts to form a hybrid virus, is typical of many RNA viruses and for coronaviruses provides a balance to their relatively slow evolutionary rate (Graham and Baric 2010). Recombinants with parts of their genomes shared with the SARS-CoV-2 progenitor (between 40 and 100 years ago, Figure 1A) are distributed on both sides of China (a distance of ~2000 km) indicating the urgent need to broaden the geographical region being searched for the SARS-CoV-2's immediate animal ancestor and avoiding being overly focussed on the Yunnan location of the two closest *Sarbecoviruses* RaTG13 and RmYN02.

The finding that Sunda (also known as Malayan) pangolins, *Manis javanica,* non-native to China, are the other mammal species from which *Sarbecoviruses* related to SARS-CoV-2 have been sampled in Guangxi and Guangdong provinces in the southern part of China (Lam et al. 2020; Xiao et al. 2020), indicates these animals are being infected in this part of the country (Figure 1B). 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 is that these *Sarbecoviruses* infected the pangolins after being trafficked into Southern China, consistent with the respiratory distress they exhibit (Liu et al. 2019; Xiao et al. 2020) and the lack of evidence of infection of Sunda pangolins in Malaysia (Lee et al. 2020).

Pangolin's susceptibility to an apparently new human coronavirus is not surprising given the well-documented generalist nature of SARS-CoV-2 (e.g., Conceicao et al. 2020), which has been found to readily transmit to multiple mammals with similar ACE2 receptors and poses a grave risk of reverse-zoonosis as has been seen most notably with human to mink transmissions (Oude Munnink et al. 2021). Twelve months since SARS-CoV-2's first characterisation, the lack of finding of an intermediate reservoir species indicates that the important evolution that gave rise to this coronavirus, with this 'ready-to-go' nature in many animal species including humans, occurred in horseshoe bats (MacLean et al. 2020). Although, the recent confirmation of a *Sarbecovirus* sampled from a Chinese pangolin, *Manis pentadactyla*, in Yunnan, sampled in 2017 (GISAID ID EPI_ISL_610156, authors: Jian-Bo Li, Hang Liu, Ting-Ting Yin, Min-Sheng Peng and Ya-Ping Zhang of the State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences), does raise the question, are Chinese pangolins infected

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

endemically? However, this may also have been an imported animal. The key, and urgent question to minimise future spillovers, is thus **not**, how did SARS-CoV-2 get from Yunnan to Hubei, but rather which bat or other animal species are harbouring nCoV *Sarbecoviruses*?

Focus should be placed on *R. affinis* and *R. sinicus* as they are the only horseshoe bat species that have ranges sufficiently dispersed across China to account for the geographical spread of i) bat Sarbecovirus recombinants in the West and East of China, ii) infected imported pangolins in the South, iii) bat Sarbecovirus recombinant links to Southern China, and iv) SARS-CoV-2 emergence towards Hubei in Central China (Figure 1B). Strikingly, the ranges of these species are overlapping, especially for *R. affinis* and *R. sinicus* across the regions of China where all the nCoV-associated viruses have been collected (Figure 1C). The other possible horseshoe bat hosts, *R. ferrumequinum,* is not found in large parts of Central or Southern China, while *R. malayanus* is found in the West part of China only. *R. malayanus's* association appears to be due to parts of the SARS-CoV-2 lineage being exchanged into this species that is found predominantly in countries on the Southwest of China (Myanmar, Thailand, Cambodia, Laos, Viet Nam, and Peninsular Malaysia; Bates et al. 2018) and only in Yunnan where RmYN02 was sampled (Zhou et al. 2020b).

The evidence for *R. affinis* being the prime suspect for SARS-CoV-2's original host is supported by its finding in shared roosts with *R. sinicus* and *R. ferrumequinum* in Yunnan and *R. sinicus* in Guangxi (Luo et al. 2013), providing opportunities for host switches, co-infections and thus recombination between the sarbecoviruses that they carry. Latinne et al. (2020) recently published a large-scale sampling expedition of coronaviruses across bats in China. Although only short RdRp fragments were sequenced, reconstructing the phylogeny for the novel viruses reveals a cluster of seven identical sarbecoviruses within the nCoV clade, close to SARS-CoV-2 (Figure S1), with the identifiable *Rhinolophus* species being *R. affinis*.

Based on the analysis of the sarbecovirus and host data presented here, we propose that horseshoe bat population sampling should focus on the known ranges of probable bat hosts and sample (carefully) in likely subterranean environments spread across China (Luo et al. 2013). Sampling strategies will also need to consider the distinct subspecies of *R. affinis* and *R. sinicus* as the delineators of genetically meaningful host populations for viruses for which there are two on mainland China for *R. affinis*: *himalayanus* and *macrurus* (Mao et al. 2010). While *R. affinis* seems to be the most likely significant reservoir host of SARS-CoV-2's proximal ancestor, we urge caution due to the generalist nature of the nCoV *Sarbecoviruses*. For example *R. pusillus* is broadly distributed across China with limited full viral genomes identified from this species, despite metagenomic analysis finding a high diversity of viruses including coronaviruses (Hu et al. 2017). It is also possible that Chinese pangolins, given their host range across Southern China (Challender et al. 2019), are the 'missing' intermediate host of the SARS-CoV-2 proximal ancestor. We urge that future sampling is not restricted to specific bat hosts or to just bat species and encompasses a range of indigenous mammals that we now know can be infected by coronaviruses (Fan et al. 2019). The recent reporting of two bat sarbecoviruses closely related to SARS-CoV-2 from samples collected in Cambodia from *R. shameli* (Mallapaty 2020), if confirmed by whole-genome analysis, would necessitate further extending the search for the SARS-COV-2 progenitor into Southeast Asia. Intriguingly the host range of *R. affinis* stretches not only across China but all the way into Southeast Asia including Cambodia.

One unique SARS-CoV-2 signature that has raised considerable interest due to its functional importance in virus cell entry is the furin cleavage site (FCS) on its Spike protein (Hoffmann

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

et al. 2020). The 12 nucleotide insertion producing this FCS is absent at that site from all of the other 68 *Sarbecoviruses* sampled, raising concerns about potential 'non-natural' laboratory origins. However, similar protease cleavage sites at this part of Spike have been reported in a number of coronaviruses, indicative of a significant role in S1/S2 and S2' protease-mediated cleavage of Spike (Garry and Gallaher, 2020). The finding of related cleavage sites supports the natural origin of the insertion. Gallaher (Gallaher, 2020) has provided one explanation for how the FCS arose through a copy-choice recombination error between the proximal ancestor of SARS-CoV-2 and a yet unsampled Betacoronavirus. This event would suggest circulation of the SARS-CoV-2 progenitor with another unsampled virus, at times infecting the same individual. Specifically, a short region of sequence of homology between SARS-CoV-2 and RmYN02 (Figure S2 and see Zhou et al. 2020b and Lytras et al. 2020) supports the copy-choice origin of the FCS in a co-infected bat. As mentioned above RmYN02 is a recombinant virus with most of its Spike gene belonging to the non-nCoV clade (recombinant regions 10 and 11, Figure 1A). Since this recombination pattern is absent from the SARS-CoV-2 genome, the event responsible took place after the two virus lineages diverged, less than about 40 years ago so relatively recently.

Another region of the Spike protein critical for emergence in humans is the receptor binding domain. The ability of the nCoV *Sarbecovirses* to use human ACE2 is confusing due to the recombination nature of Spike in these viruses. Specifically, RaTG13 has a divergent RBM region (Boni et al. 2020), while RmYN02 is non-nCoV like in this region. However, that the pangolin sarbecoviruses can use hACE2 efficiently (Thomson et al. 2020), a probable recent nCoV acquisition from bats, demonstrates members of the nCoV clade of viruses can infect humans if given the opportunity. This capability of bat *Sarbecoviruses* to infect human cells was well documented prior to the emergence of SARS-CoV-2 in 2019 (Ge et al. 2013; Menachery et al. 2015).

Although, beyond relatively rare detection of SARS-like antibodies in rural communities in China (Li et al. 2019; Wang et al. 2018), that SARS-CoV variants have not, to our knowledge, seeded outbreaks in humans before, would indicate there are limited human exposure to these viruses suggesting ecological barriers to emergence (Plowright et al. 2017). One possible recent disruption that would have caused widespread and unusual movement of animals in China, breaking this 'barrier', was the dramatic shortage of pork products in 2019 (Mason-D'Croz et al. 2020) as a result of African swine fever virus (ASFV) infecting 100s of millions of pigs in China (30-50% of the population) and the very sharp rise in the price of pork. The culling of large numbers of pigs, together with regional control measures of hog/pork movement, led to increased live animal movements in China. As pork is the major food source in China, such movements on a large scale will have potentially brought humans into increased contact with *Sarbecovirus* infected animals as i) exotic meats replaced pork, ii) animals from rural locations were brought to city markets from rural locations and/or iii) by infected meat being transported in cold chain processes. Given the reality of frequent human-animal contact, routine characterisation of respiratory infections would seem a sensible precaution to prevent future emergence of *Sarbecoviruses*.

Conclusion. The currently available data, although sparse, illustrates a complex history behind the natural evolution of SARS-CoV-2, governed by co-circulation of related coronaviruses, over at least the last 100 years, across the bat populations from East-to-West/Central and Southern China and multiple recombination events imprinted on the genomes of these viruses. The evidence of recombination events between viruses sampled in different parts of China and from different bat hosts indicate frequent movement

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of the viruses between areas as a result of the different bat populations that carry them. Although the presence of occasional intermediate hosts, even between bats and humans in the case of SARS-CoV-2 cannot be discounted, the long geographic ranges covered by the recombination patterns would require a reservoir host with a wide geographical range. All the evidence points to this host being Chinese horseshoe bats. Having presented evidence in support for *R. affinis*'s importance, it should be noted at least 20 different *Rhinolophus* species are distributed across China (four being endemic to China) leaving many species for which the viruses are unknown. Also, the generalist nature of sarbecoviruses means wild or farmed animals (e.g., minks) could facilitate transmission of viruses from bats to humans. The risk of future emergence of a new SARS-CoV-2 nCoV strain is too high to restrict sampling strategies.

Methods

The whole genome sequences of the 69 *Sarbecoviruses* used in this analysis (Table S1) were aligned and the open reading frames (ORF) of the major protein-coding genes were defined based on SARS-CoV-2 annotation. To minimise alignment error codon-level alignments of the ORFs were created using MAFFT (Katoh et al. 2005) 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 and Others 1999).

The resulting alignment was examined for recombination breakpoints using the Genetic Algorithm for Recombination Detection (GARD) method (Pond et al. 2006) and likelihood was evaluated using the Akaike Inference Criterion (AIC). This analysis provided 16 likely breakpoints (positions corresponding to the SARS-CoV-2 reference genome Wuhan-Hu-1 in order: 926, 3327, 5115, 8886, 11676, 14306, 18473, 20082, 21470, 22566, 23311, 24279, 25449, 26054, 27845 and 28257) based on which the whole-genome alignment was split into 17 putatively non-recombinant regions. Phylogenetic reconstruction of each region was performed using RAxML-NG (Kozlov et al. 2019) under a GTR+Γ model. Node support was determined using the Transfer Bootstrap Expectation (TBE) (Lemoine et al. 2018) with 1000 replicates for each tree. All constructed phylogenies are presented in Figure S3.

Almost all phylogenies have the same overall topology with the BtKY72 and BM48-31 viruses lineage being the outgroup and two sister lineages separating in the ingroup clade, one containing SARS-CoV-2 and the other SARS-CoV along with their related viruses. The nCoV clade was defined as the monophyletic grouping of SARS-CoV-2 with the non-nCoV clade as its ingroup sister lineage (similar to the clade definition used in MacLean et al. (2020). The phylogenies of three non-recombinant regions - 10, 12 and 13 - had topologies incongruent to that of all the other genomic regions. In particular, in these special cases the clade containing SARS-CoV-2 clusters within parts of the non-nCoV clade (Figure S3). The three regions are parts of the Spike ORF, indicating that the tree incongruencies are either a result of selection signatures that disrupt the phylogenetic reconstruction, or deeper signatures of Spike recombination not detected by GARD. The nCoV clade was visually determined for these regions based on its definition in the other non-recombinant regions' trees. A final special case is non-recombinant region 16 for which four viruses - JTL2012, JTMC15, BtKY72, BM48-31 - are missing more than 90% of their sequence. These viruses were excluded from that alignment before reconstructing the phylogeny to avoid artefactual signals due to limited genetic information. Phylogenies were visualised using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and ETE 3 (Huerta-Cepas et al. 2016).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To illustrate the distance of each virus from SARS-CoV-2, while distinguishing whether the virus in question is part of the nCoV clade or the non-nCoV clade, we use an arbitrary tip distance scale normalised between all phylogenies so distances are comparable between regions. For each maximum likelihood tree, the tip distance between each tip and SARS-CoV-2 is calculated using ETE 3 as d1 for members of the nCoV clade and d2 for members of the non-nCoV clade. The distances are then normalised so that for nCoV clade members they range between 0.1 and 1.1 (1.1 being SARS-CoV-2 itself and 0.1 being the most distant tip from SARS-CoV-2 within the nCoV clade) and between -0.1 and -1.1 for non-nCoV members (-0.1 being the closest non-nCoV virus to SARS-CoV-2 and -1.1 the most distant), as follows:

$$d'_1 = 1.1 - \frac{d_1}{d_{1,max}} \qquad \text{(1:nCoV)}$$

$$d'_2 = -0.1 - \frac{d_2 - d_{2,min}}{d_{2,max} - d_{2,min}} \qquad \text{(2:non-nCoV)}$$

With d'1 and d'2 being the normalised 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.

To provide temporal information to the phylogenetic history of the viruses, we performed a Bayesian phylogenetic analysis on non-recombination region 4, using BEAST (Bouckaert et al. 2019). This region was selected due to its length, being the second longest non-recombinant region in the analysis (3764 bp), and because it represents one of the non-recombinant regions where the CoVZC45/CoVZXC21 lineage clusters within the nCoV clade. Based on the observation of an increased evolutionary rate specific to the deepest branch of the nCoV clade reported in MacLean et al. (2020), 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 $5x10^{-4}$ and standard deviation $2x10^{-4}$ 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+Γ substitution model was specified independently for each partition. The maximum likelihood phylogeny reconstructed previously for non-recombinant region 4 was used as a starting tree. A constant size coalescent model was used for the tree prior and a lognormal prior with a mean of 6 and standard deviation of 0.5 was specified on the population size. Two independent MCMC runs were performed for 250 million states for the dataset.

Geographical and genomic visualisation was performed using D3 and JavaScript in Observable.

Acknowledgements

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 supplementary material. DLR and JH are funded by the MRC (MC_UU_1201412) and DLR by the WT (220977/Z/20/Z). SL is funded by an MRC studentship.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

Bates, P., Bumrungsri, S., Csorba, G. & Soisook, P'. 2018. '*Rhinolophus Malayanus*: *IUCN Red List of Threatened Species*. https://doi.org/10.2305/iucn.uk.2019-3.rlts.t19551a21978424.en.

Boni, Maciej F., Philippe Lemey, Xiaowei Jiang, Tommy Tsan-Yuk Lam, Blair W. Perry, Todd A. Castoe, Andrew Rambaut, and David L. Robertson. 2020. 'Evolutionary Origins of the SARS-CoV-2 *Sarbecovirus* Lineage Responsible for the COVID-19 Pandemic'. *Nature Microbiology* 5 (11): 1408–17.

Bouckaert, Remco, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, et al. 2019. 'BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis'. *PLoS Computational Biology* 15 (4): e1006650.

Challender, D., S. Wu, P. Kaspal, A. Khatiwada, A. Ghose, N. Ching-Min Su, and T. Laxmi Suwal. 2019. '*Manis Pentadactyla*. The IUCN Red List of Threatened Species 2019: E. T12764A123585318'.

Conceicao, Carina, Nazia Thakur, Stacey Human, James T. Kelly, Leanne Logan, Dagmara Bialy, Sushant Bhat, et al. 2020. 'The SARS-CoV-2 Spike Protein Has a Broad Tropism for Mammalian ACE2 Proteins'. *PLoS Biology* 18 (12): e3001016.

Fan, Yi, Kai Zhao, Zheng-Li Shi, and Peng Zhou. 2019. "Bat Coronaviruses in China." *Viruses* 11 (3) 210.

Gallaher, William R. 2020. 'A Palindromic RNA Sequence as Common Breakpoint Contributor to Copy-Choice Recombination in SARS-CoV-2'. https://doi.org/10.21203/rs.3.rs-46379/v1.

Garry, Robert F., and William R. Gallaher. 2020. 'Naturally Occurring Indels in Multiple Coronavirus Spikes'. https://virological.org/t/naturally-occurring-indels-in-multiple-coronavirus-spikes/560.

Ge, Xing-Yi, Jia-Lu Li, Xing-Lou Yang, Aleksei A. Chmura, Guangjian Zhu, Jonathan H. Epstein, Jonna K. Mazet, et al. 2013. "Isolation and Characterization of a Bat SARS-like Coronavirus That Uses the ACE2 Receptor." *Nature* 503 (7477): 535–38.

Gorbalenya, A., S. Baker, R. Baric, R. de Groot, Christian Drosten, A. Gulyaeva, B. Haagmans, et al. 2020. 'Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming It SARS-CoV-2'. *Nature Microbiology* 2020: 03–04.

Graham, Rachel L., and Ralph S. Baric. 2010. 'Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission'. *Journal of Virology* 84 (7): 3134–46.

Hall, Tom A. 1999. 'BioEdit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/NT'. In *Nucleic Acids Symposium Series*, 41:95–98.

Hoffmann, Markus, Hannah Kleine-Weber, and Stefan Pöhlmann. 2020. 'A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells'. *Molecular Cell* 78 (4): 779–84.e5.

Hu, Dan, Changqiang Zhu, Lele Ai, Ting He, Yi Wang, Fuqiang Ye, Lu Yang, et al. 2018. 'Genomic Characterization and Infectivity of a Novel SARS-like Coronavirus in Chinese Bats'. *Emerging Microbes & Infections* 7 (1): 154.

Hu, Dan, Changqiang Zhu, Yi Wang, Lele Ai, Lu Yang, Fuqiang Ye, Chenxi Ding, et al. 2017. 'Virome Analysis for Identification of Novel Mammalian Viruses in Bats from Southeast China'. *Scientific Reports* 7 (1): 10917.

Huerta-Cepas, Jaime, François Serra, and Peer Bork. 2016. 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data'. *Molecular Biology and Evolution* 33 (6): 1635–38.

Katoh, Kazutaka, Kei-Ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. 'MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment'. *Nucleic Acids Research* 33 (2): 511–18.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Kozlov, Alexey M., Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. 2019. 'RAxML-NG: A Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference'. *Bioinformatics* 35 (21): 4453–55.

Lam, Tommy Tsan-Yuk, Na Jia, Ya-Wei Zhang, Marcus Ho-Hin Shum, Jia-Fu Jiang, Hua-Chen Zhu, Yi-Gang Tong, et al. 2020. 'Identifying SARS-CoV-2-Related Coronaviruses in Malayan Pangolins'. *Nature* 583 (7815): 282–85.

Latinne, Alice, Ben Hu, Kevin J. Olival, Guangjian Zhu, Libiao Zhang, Hongying Li, Aleksei A. Chmura, et al. 2020. 'Origin and Cross-Species Transmission of Bat Coronaviruses in China'. *Nature Communications* 11 (1): 4235.

Lee, Jimmy, Tom Hughes, Mei-Ho Lee, Hume Field, Jeffrine Japning Rovie-Ryan, Frankie Thomas Sitam, Symphorosa Sipangkui, et al. 2020. "No Evidence of Coronaviruses or Other Potentially Zoonotic Viruses in Sunda Pangolins (*Manis Javanica*) Entering the Wildlife Trade via Malaysia." EcoHealth 17 (3): 406–18.

Lemoine, F., J. -B. Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, and O. Gascuel. 2018. 'Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data'. *Nature*. 556 (7702): 452–456.

Li, Hongying, Emma Mendelsohn, Chen Zong, Wei Zhang, Emily Hagan, Ning Wang, Shiyue Li, et al. 2019. 'Human-Animal Interactions and Bat Coronavirus Spillover Potential among Rural Residents in Southern China'. *Biosafety and Health* 1 (2): 84–90.

Lin, Xian-Dan, Wen Wang, Zong-Yu Hao, Zhao-Xiao Wang, Wen-Ping Guo, Xiao-Qing Guan, Miao-Ruo Wang, et al. 2017. 'Extensive Diversity of Coronaviruses in Bats from China'. *Virology* 507 (July): 1–10.

Li, Qun, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, et al. 2020. 'Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia'. *The New England Journal of Medicine* 382 (13): 1199–1207.

Liu, Ping, Wu Chen, and Jin-Ping Chen. 2019. 'Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (*Manis Javanica*)'. *Viruses* 11 (11).

Luo, Jinhong, Tinglei Jiang, Guanjun Lu, Lei Wang, Jing Wang, and Jiang Feng. 2013. 'Bat Conservation in China: Should Protection of Subterranean Habitats Be a Priority?' *Oryx: The Journal of the Fauna Preservation Society* 47 (4): 526–31.

MacLean, Oscar A., Spyros Lytras, Steven Weaver, Joshua B. Singer, Maciej F. Boni, Philippe Lemey, Sergei L. Kosakovsky Pond, and David L. Robertson. 2020. 'Natural Selection in the Evolution of SARS-CoV-2 in Bats, Not Humans, Created a Highly Capable Human Pathogen'. *BioRxiv* https://doi.org/10.1101/2020.05.28.122366.

Mallapaty, Smriti. 2020. "Coronaviruses Closely Related to the Pandemic Virus Discovered in Japan and Cambodia." *Nature* 588 (7836): 15–16.

Mao, Xiu Guang, Guang Jian Zhu, Shuyi Zhang, and Stephen J. Rossiter. 2010. 'Pleistocene Climatic Cycling Drives Intra-Specific Diversification in the Intermediate Horseshoe Bat (*Rhinolophus Affinis*) in Southern China'. *Molecular Ecology* 19 (13): 2754–69.

Mason-D'Croz, Daniel, Jessica R. Bogard, Mario Herrero, Sherman Robinson, Timothy B. Sulser, Keith Wiebe, Dirk Willenbockel, and H. Charles J. Godfray. 2020. 'Modelling the Global Economic Consequences of a Major African Swine Fever Outbreak in China'. *Nature Food*. 1:221–228.

Menachery, Vineet D., Boyd L. Yount Jr, Kari Debbink, Sudhakar Agnihothram, Lisa E. Gralinski, Jessica A. Plante, Rachel L. Graham, et al. 2015. "A SARS-like Cluster of Circulating Bat Coronaviruses Shows Potential for Human Emergence." *Nature Medicine* 21 (12): 1508–13.

Oude Munnink, Bas B., Reina S. Sikkema, David F. Nieuwenhuijse, Robert Jan Molenaar, Emmanuelle Munger, Richard Molenkamp, Arco van der Spek, et al. 2021. "Transmission of

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.22.427830; this version posted January 22, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

SARS-CoV-2 on Mink Farms between Humans and Mink and back to Humans." *Science* 371 (6525): 172–77.

Plowright, Raina K., Colin R. Parrish, Hamish McCallum, Peter J. Hudson, Albert I. Ko, Andrea L. Graham, and James O. Lloyd-Smith. 2017. 'Pathways to Zoonotic Spillover'. *Nature Reviews. Microbiology* 15 (8): 502–10.

Pond, S. L. Kosakovsky, S. L. Kosakovsky Pond, D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost. 2006. 'GARD: A Genetic Algorithm for Recombination Detection'. *Bioinformatics* 22 (24): 3096–3098.

Lytras, S, Oscar MacLean, and David L Robertson. 2020. 'The *Sarbecovirus* Origin of SARS-CoV-2's Furin Cleavage Site'. https://virological.org/t/the-sarbecovirus-origin-of-sars-cov-2-s-furin-cleavage-site/536.

Smith, AT and Yan Xie. 2013. 'Mammals of China Edited'. 2013. *The Quarterly Review of Biology* 88 (4): 363–363.

Suyama, Mikita, David Torrents, and Peer Bork. 2006. 'PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments'. *Nucleic Acids Research* 34 (Web Server issue): W609–12.

Thomson, E. C., L. E. Rosen, J. G. Shepherd, and R. Spreafico. 2020. 'The Circulating SARS-CoV-2 Spike Variant N439K Maintains Fitness While Evading Antibody-Mediated Immunity'. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.11.04.355842v1.abstract.

Wang, Hongru, Lenore Pipes, and Rasmus Nielsen. 2020. 'Synonymous Mutations and the Molecular Evolution of SARS-Cov-2 Origins'. Virus Evolution veaa098, https://doi.org/10.1093/ve/veaa098.

Wang, Ning, Shi-Yue Li, Xing-Lou Yang, Hui-Min Huang, Yu-Ji Zhang, Hua Guo, Chu-Ming Luo, et al. 2018. 'Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China'. *Virologica Sinica* 33 (1): 104–7.

Xiao, Kangpeng, Junqiong Zhai, Yaoyu Feng, Niu Zhou, Xu Zhang, Jie-Jian Zou, Na Li, et al. 2020. 'Isolation of SARS-CoV-2-Related Coronavirus from Malayan Pangolins'. *Nature* 583 (7815): 286–89.

Xu, Ling, Jing Guan, Wilson Lau, and Yu Xiao. 2016. 'An Overview of Pangolin Trade in China'. *TRAFFIC September* 2016: 1–10.

Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. 'A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin'. *Nature* 579 (7798): 270–73.

Zhou, Hong, Xing Chen, Tao Hu, Juan Li, Hao Song, Yanran Liu, Peihan Wang, et al. 2020b. 'A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein'. *Current Biology: CB* 30 (11): 2196–2203.e3.