Article

Retracing Phylogenetic, Host and Geographic Origins of Coronaviruses with Coloured Genomic Bootstrap Barcodes: SARS-CoV and SARS-CoV-2 as Case Studies

Alexandre Hassanin [1,*] and Opale Rambaud [1]

[1] Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, EPHE, MNHN, UA, Paris, France
* Correspondence: alexandre.hassanin@mnhn.fr

Abstract: Phylogenetic trees of coronaviruses are difficult to interpret because they undergo frequent genomic recombination.

Here, we propose a new method, named coloured genomic bootstrap (CGB) barcodes, to highlight the polyphyletic origins of human sarbecoviruses and understand their host and geographic origins. The results indicate that SARS-CoV and SARS-CoV-2 contain genomic regions of mixed ancestry originating from horseshoe bat (*Rhinolophus*) viruses. First, different regions of SARS-CoV share exclusive ancestry with five *Rhinolophus* viruses from Southwest China (RfYNLF/31C: 17.9%; RpF46: 3.3%; RspSC2018: 2.0%; Rpe3: 1.3%; RaLYRa11: 1.0%) and 97% of its genome can be related to bat viruses from Yunnan (China), supporting its emergence in *Rhinolophus* species of this province. Second, different regions of SARS-Cov-2 share exclusive ancestry with eight *Rhinolophus* viruses from Yunnan (RpYN06: 5.8%; RaTG13: 4.8%; RmYN02: 3.8%), Laos (RpBANAL103: 3.3%; RmarBANAL236: 1.7%; RmBANAL52: 1.0%; RmBANAL247: 0.7%), and Cambodia (RshSTT200: 2.3%), and 98% of its genome can be related to bat viruses from northern Laos and Yunnan, supporting its emergence in *Rhinolophus* species of this region.

Although CGB barcodes are very useful to retrace the origins of human sarbecoviruses, further investigations are needed to better apprehend the diversity of coronaviruses in bats from Cambodia, Laos, Myanmar, Thailand and Vietnam.

Keywords: coronavirus; genome; recombination; COVID-19; reservoir host; secondary host; phylogenetic support; tree reconstruction

1. Introduction

Tree reconstruction methods, such as Bayesian inference and maximum likelihood (ML), are very popular to decipher phylogenetic relationships between pathogens based on multiple sequence alignments. In particular, newly discovered coronaviruses are routinely described based on Bayesian and ML trees reconstructed from whole or partial genome alignments [1-4]. For instance, most recent studies on SARS-CoV-2, the virus involved in the COVID-19 pandemic, have published a whole-genome tree of the subgenus *Sarbecovirus* (family Coronaviridae, genus *Betacoronavirus*) in which the human virus was found closely related to RaTG13, a virus detected in a horseshoe bat of the species *Rhinolophus affinis* sampled in 2013 in the Yunnan province of China [4-7]. However, discordant placements of SARS-CoV-2 were supported in phylogenetic trees based on different genomic fragments. This was well illustrated in Zhou *et al*. [7], in which SARS-CoV-2 appeared closely related to four bat viruses, namely RaTG13, RmYN02, RpYN06, and RshSTT200, in the RNA-dependent RNA polymerase (*RdRp*) gene tree, sister-group of RmYN02 and RpYN06 in the ORF1ab tree, and linked to RaTG13 in the *Spike* gene tree. Such conflicting results between gene trees are typically explained by genomic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

recombination, a process resulting in mosaic genomes containing regions from different parental viruses. The most widely-accepted model to explain recombination in coronaviruses is the copy-choice model, also named template switching model: during RNA replication, the viral RdRp can pause on the RNA template and switch to another template, thereby generating a recombinant RNA molecule with mixed ancestry [8]. Our recent study has also suggested that circular RNAs may be involved in the process of genomic recombination [9].

Several previous studies have provided strong evidence that *Sarbecovirus* genomes are derived from a large number of past recombination events in bats [9,10] and also in humans [11,12]. This means that each *Sarbecovirus* RNA genome has a specific mosaic structure, that is, a unique combination of genomic fragments showing different evolutionary histories: some fragments may be shared with only one virus, suggesting recent ancestry; other fragments may provide only support for grouping with several divergent viruses, suggesting older ancestry (or insufficient sampling of viruses); the origin of some fragments may be very difficult to interpret due to multiple recombination events in overlapping or nested genomic locations and loss of phylogenetic signal over time (multiple nucleotide substitutions at the same site). To better interpret conflicting phylogenetic signals due to genomic recombination, we report hereinafter a new approach, named coloured genomic bootstrap (CGB) barcodes, in which a virus genome of special interest (e.g. the common ancestor of SARS-CoV-2) is represented by a succession of coloured regions showing the best phylogenetic signals, i.e. including the fewest number of closest relatives among available viral genomes. The method was applied to an alignment of 75 *Sarbecovirus* genomes to provide new insights into the phylogenetic, host and geographic origins of SARS-CoV-2 and SARS-CoV (virus involved in the 2002-2003 and 2003-2004 SARS outbreaks [13,14]).

2. Materials and Methods

2.1. Nucleotide alignment of Sarbecovirus genomes

Complete genomes available for *Sarbecovirus* in June 2022 in GenBank (https://www.ncbi.nlm.nih.gov/), GISAID (https://www.epicov.org/), and NGDC (https://ngdc.cncb.ac.cn/) databases were downloaded in Fasta format. Sequences with large stretch of missing data were removed. Several genomes showing perfect identity or high nucleotide similarity (more than 99.9% of nucleotide identity) were published for pangolin sarbecoviruses from Guangxi (5 sequences), bat sarbecoviruses from Thailand (5 sequences), Cambodia (two sequences), etc. For these clusters, a single genome was retained for our analyses. The international databases contain millions of SARS-CoV-2 genomes and hundreds of SARS-CoV genomes. For human SARS-CoV-2, we decided to include in the alignment the reference genome and one representative for each of the six variants of concern (VOC; Alpha, Beta, Gamma, Delta, Epsilon, Omicron), which were selected under NCBI Virus (ncbi.nlm.nih.gov/labs/virus/) using the following criteria: country for which the highest number of sequences was available (i.e. USA), Illumina sequencing; no stop codon in the coding sequences (cds); and no missing data. We also included two SARS-CoV-2 genomes extracted from small carnivores of the family Mustelidae, i.e. *Mustela lutreola* (European mink) and *Neovison vison* (American mink), differing by more than 0.1%. For human SARS-CoV, we included in the alignment four genomes showing more than 0.1% of nucleotide divergence. Similarly, we included three SARS-CoV-like genomes extracted from *Paguma larvata* (masked palm civet), as this small carnivore of the family Viverridae was identified as a possible intermediate host between bats and humans during the 2002-2003 and 2003-2004 SARS outbreaks [13,14]. The details on the 75 selected genomes are provided in supplementary Table S1. They include all viral lineages previously described within the subgenus *Sarbecovirus* [4-7, 15,16].

The nucleotide sequences were aligned in Geneious Prime® 2020.0.3 with MAFFT version 7.450 [17] using default parameters. Then, the alignment was corrected manually

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

on AliView 1.26 [18] based on nucleotide and amino-acid sequences using the three following criteria: (i) the number of indels was minimized because they are rarer events than nucleotide substitutions; (ii) transitions were privileged over transversions because they are more frequent; and (iii) changes between similar amino-acids (as shown by the ClustalX colour scheme) were preferred. The insertions found in only one virus were removed from the whole-genome alignment.

2.2. Phylogenetic analyses

Maximum likelihood analysis of the whole-genome alignment of sarbecoviruses was carried out using RAxML 8.2.11 [19], different GTR+G models for the three codon-positions and non-coding regions, and 1,000 bootstrap replicates. The RAxML bootstrap trees were executed in PAUP* version 4.0a [20] to construct the bootstrap 50% majority-rule consensus tree.

To examine the distribution of phylogenetic support along the whole-genome alignment, the dataset was bootstrapped under the SWB program [9] using a window of W nucleotides (W parameter) moving in steps of 50 nt (S parameter). The window size (W) is a key parameter for SWB analyses because the amount of phylogenetic signal depends on both the number of nucleotide sites and their evolutionary rates [9]. For that reason, we decided to perform five SWB analyses with the same step parameter (S) of 50 nt but using five different window sizes, i.e. 400 nt, 500 nt, 600 nt, 1000 nt and 2000 nt (Table 1). The smallest window size (W = 400 nt) was applied to detect possible changes in phylogenetic relationships due to the recombinant origin of small genomic regions, whereas the largest window size (W = 2000 nt) was used to guarantee enough phylogenetic signal (informative sites) for bootstrap analyses. The intermediate window sizes (500, 600 and 1000 nt) were used to better interpret the differences between SWB results based on the two extreme values. In the SWB program, each window bootstrap (WB) subdataset was automatically run in RAxML [19] with a GTR+G model and 100 bootstrap replicates. The SWB output is a CSV file containing the bootstrap percentages (BP) calculated for each WB subdatasets and for all the bipartitions (nodes) reconstructed during the SWB analysis. For example, the SWB_{400} output (SWB analysis based on a window of 400 nt) includes 1,213,380,595 BP values (595 WB subdatasets X 204,001 bipartitions), whereas the SWB_{2000} output includes 11,189,625 BP values (563 WB subdatasets X 19,875 bipartitions) (Table 1).

Table 1. Five sliding window bootstrap (SWB) analyses based on an alignment of 75 *Sarbecovirus* genomes (length: 30,115 nt).

	Window size (nt)	Step (nt)	WB subdatasets	SWB bipartitions[1]	SuperTRI matrix[2]	BBC bipartitions[3]	SARS-CoV bipartitions[4]	SARS-CoV-2 bipartitions[4]
1	400	50	595	204,001	750,330	1,263	215	178
2	500	50	593	150,118	649,437	1,240	208	176
3	600	50	591	116,934	577,232	1,212	213	175
4	1000	50	583	57,444	427,041	1,061	177	163
5	2000	50	563	19,875	288,711	818	142	132

[1]: Bipartitions (with bootstrap percentages (BP) caculated for each WB subdatasets) obtained under the SWB program [9];
[2]: Number of characters in the MRP matrix reconstructed using LFG [9] and SuperTRI [21] programs;
[3]: SWB bipartitions including at least one BP ≥ 50 selected under the BBC program using the SWB file as input [9];
[4]: BBC bipartitions selected under Microsoft® Excel using the BBC file as input.

The bootstrap bipartitions generated from the five SWB analyses based on different window sizes were used for SuperTRI analyses [21] to reconstruct the trees showing the most reliable phylogenetic relationships. The LFG program [9] was used to convert the SWB output file into bootstrap log files, which were then used as inputs in SuperTRI v57 [21] to construct a MRP (Matrix Representation with Parsimony) file. For example, the SWB_{400} output was converted with the LFG program into 595 bootstrap log files, and these files were further transformed into a MRP file using SuperTRI v57. In the MRP_{400} file, each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

of the 750,330 characters represents one SWB bipartition with its assigned bootstrap percentage calculated in one of the 595 window bootstrap analyses. The MRP_{400} file was then executed in PAUP* version 4.0a [20] using 1000 bootstrap replicates of weighted parsimony (with BPs of the SWB_{400} analysis assigned as weights) in order to reconstruct the SuperTRI bootstrap 50%-majority-rule consensus (SB_{400}) tree. Finally, the mean bootstrap percentages (MBP_{400}) were calculated automatically in SuperTRI v57 [21] for all nodes of the SB_{400} trees. The same approach was conducted using the SWB outputs obtained with the sliding windows of 500, 600, 1000 and 2000 nt. In total, we therefore reconstructed five SB trees with MBP values at the nodes.

2.3. Construction of genomic bootsrap barcodes

The BBC program [9] was used to select only SWB bipartitions (i.e. phylogenetic hypotheses) showing one or more BP values ≥ 50%, and to construct their corresponding genomic bootstrap (GB) barcodes. For example, the SWB_{400} output contains 204,001 bipartitions, each with 595 BP values. After BBC analysis, the GB_{400} barcodes were done for the 1,263 selected SWB_{400} bipartitions showing one or more BP values ≥ 50% (Table 1). A GB barcode is a small image representing the genome alignment and in which the N BP values (N = 595 with an alignment of 30,115 nt and a window size of 400 nt) obtained for the SWB bipartition of interest were transformed into N coloured vertical bars using the following code: green for BP ≥ 70%; grey for 30% < BP < 70%; and red for BP ≤ 30%. Using the same BBC procedure but a different SWB output (either SWB_{500}, SWB_{600}, SWB_{1000} or SWB_{2000}), we constructed 1,240 GB_{500} barcodes, 1,212 GB_{600} barcodes, 1,061 GB_{1000} barcodes, and 818 GB_{2000} barcodes (Table 1). All the 5,594 GB barcodes and the five BBC output files (BBC_{400}, BBC_{500}, BBC_{600}, BBC_{1000}, and BBC_{2000}) are available in the Open Science Framework (OSF) platform (https://osf.io/XXXXX/; the link will be provided for the final version) and some of them were reported on tree nodes of Figure 2.

2.4. Construction of coloured genomic bootstrap barcodes

A new method was developed in this study for constructing coloured genomic bootstrap (CGB) barcodes for a virus of special interest or the common ancestor of several viruses. A phylogenetic CGB barcode is a small image representing the genome of a virus (or an ancestral virus) in which the different colours show the best phylogenetic signals, i.e. containing the fewest number of closely-related viruses, detected in the different regions of the genomic alignment used for the analyses. Two other kinds of CGB barcodes were derived from the phylogenetic CGB barcodes: the host CGB barcodes showing the host origins of the closely-related viruses and the geographic CGB barcodes showing the geographic origins of the closely-related viruses. For this study, we choose to reconstruct CGB barcodes for the common ancestor of seven selected SARS-CoV genomes and for the common ancestor of nine selected SARS-CoV-2 genomes (see supplementary Table S1 for the origin of the sequences used in this study). To avoid repetitions, only the SARS-CoV-2 procedure is described below.

In the first step, the five BBC output files were imported in Microsoft® Excel. They represent five lists of SWB bipartitions showing at least one window BP value ≥ 50%. They contain the following information for each SWB bipartition: bipartition number, lists of viruses included in the bipartition, binary representations of the bipartition (*: viruses included in the bipartition; -: viruses excluded from the bipartition), BPs obtained for the N sliding window bootstrap analyses (e.g. N = 595 for SWB_{400} and N = 563 for SWB_{2000}). The five BBC outputs include the following number of SWB bipartitions: 1,263 for BBC_{400}, 1,240 for BBC_{500}, 1,212 for BBC_{600}, 1,061 for BBC_{1000}, and 818 for BBC_{2000} (Table 1). The outputs were entered sequentially into the same Excel file starting with BBC_{400} and ending with BBC_{2000}, and they were highlighted with five different background colours. To allow comparisons between the five BBC results, W and S parameters (window size and moving

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

steps) were used to calculate the median positions (pos.) for the N BPs calculated in each of the five SWB analyses.

In the second step, only bipartitions including all the nine SARS-CoV-2 sequences were selected: 178 bipartitions for BBC_{400}, 176 bipartitions for BBC_{500}, 175 bipartitions for BBC_{600}, 163 bipartitions for BBC_{1000}, and 132 bipartitions for BBC_{2000} (Table 1). For each of the five BBC lists, the single bipartition including only SARS-CoV-2 sequences was removed. Then, other bipartitions were ranked in increasing order of size, from +1 (for the bipartitions including the nine SARS-CoV-2 sequences + 1 closely-related virus) to +66 (= 75 - 9, for the single bipartition including all viruses of our dataset). To make comparisons between BPs calculated in the five SWB analyses, a new column was inserted to renumber BBC_{500}, BBC_{600}, BBC_{1000}, and BBC_{2000} bipartitions using BBC_{400} numbers as references.

In the third step, all BPs ≥ 70% were highlighted in green and all BPs comprised between 50% and 70% were highlighted in yellow green using conditional formatting options in Microsoft® Excel. We performed the comparisons starting with bipartitions +1, i.e. containing all SARS-CoV-2 sequences plus one additional virus. Due to past events of genomic recombination, we found several bipartitions +1 supporting conflicting phylogenetic relationships. For each of these bipartitions, we identified the genomic regions containing robust phylogenetic signal (GRPS) using the criteria developed later. Then we proceeded similarly for bipartitions +2 (containing all SARS-CoV-2 sequences plus two additional viruses), bipartitions +3 (containing all SARS-CoV-2 sequences plus three additional viruses), etc. By this way, we were able to identify the closest virus(es) to SARS-CoV-2 in all regions of our genome alignment. All selected GRPS contained a robust phylogenetic signal (BP ≥ 70%) in at least two WB subdatasets of the BBC_{400}, BBC_{500} or BBC_{600} results. The 5′ and 3′ ends of GRPS were extended using the following criteria: (i) by accepting BPs between 50% and 70% for BBC_{400} results; (ii) when median positions showed an average BP ≥ 50% for BBC_{400}, BBC_{500} and BBC_{600} results; and (iii) when median positions showed an average BP ≥ 50% for all the five BBC results (BBC_{400}, BBC_{500}, BBC_{600}, BBC_{1000}, and BBC_{2000}). This strategy was adopted for three major reasons: (i) GRPS of small lengths cannot be detected using SWB analyses based on the largest window sizes (i.e. 1000 nt and 2000 nt); (ii) due to stochastic variation in BP values, the comparisons between the three SWB analyses based on the smallest window sizes (i.e. 400 nt, 500 nt and 600 nt) allow us to better detect GRPS of small lengths; (iii) GRPS of large size can be erroneously interrupted if we consider only SWB analyses based on the smallest window sizes because they contain lesser amounts of phylogenetic signal [9]. It is therefore important to also make comparisons with BBC results based on the largest window sizes (i.e. 1000 nt and 2000 nt).

In the fourth step, the intervals of GRPS (5′ and 3′ median positions in the whole-genome alignment) were written in a new CSV file for each of the 52 SWB bipartitions including SARS-CoV2 sequences in which one or more GRPS were identified. A specific colour code was chosen for each of the 52 bipartitions, and the file was used as input in the CGB program (python script) to construct the 52 phylogenetic CGB barcodes of different colours. Two other files were derived from the original CSV file: (i) a file for host CGB barcodes, in which different colours were assigned to the following nine taxa: *Rhinolophus affinis*, *Rhinolophus malayanus*, *Rhinolophus marshalli*, *Rhinolophus pusillus*, *Rhinolophus shameli*, *R. affinis* + *Manis javanica*; *Rhinolophus* species, *Rhinolophus* species + *M. javanica*, bat species + *M. javanica*; and (ii) a file for geographic CGB barcodes, in which different colours were chosen for the following nine geographic areas: Cambodia; North Laos, SE Asia, Yunnan, North Laos + Yunnan, SE Asia + Yunnan, North Laos + China, SE Asia + China, and SE Asia + China + Japan. All files used to construct the CGB barcodes for SARS-CoV-2 and SARS-CoV (SWB_{400}, SWB_{500}, SWB_{600}, SWB_{1000}, SWB_{2000}, BBC_{400}, BBC_{500}, BBC_{600}, BBC_{1000}, BBC_{2000}, CGB-SARS-CoV-2 and CGB-SARS-CoV-2 files) are available at https://osf.io/XXXXX/; the link will be provided for the final version).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

3. Results

3.1. Phylogenetic analyses based on an alignment of 75 Sarbecovirus genomes

In this study, 75 genomes of the subgenus *Sarbecovirus* (supplementary Table S1) were aligned to infer phylogenetic relationships. The positions of the coding sequences were the following in our final alignment of 30,115 nucleotides (nt): 256-21,654 for ORF1ab, including the RNA-dependent RNA polymerase gene [*RdRp*] at positions 13,540-16,335; 21,664-25,557 for the spike (*S*) gene; 25,567-26,394 for ORF3a; 26,419-26,649 for the envelope (*E*) gene; 26,704-27,375 for the membrane (*M*) gene; 27,388-27,579 for ORF6; 27,589-28,101 for ORF7ab; 28,108-28,485 for ORF8; 28,504-29,775 for the nucleocapsid (*N*) gene; and 29,803-29,919 for ORF10 (Figure 1).

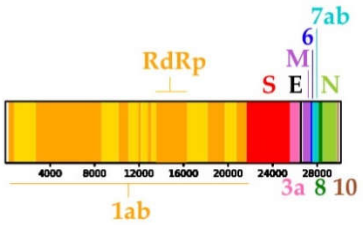

Figure 1. Positions of the coding sequences in the alignment of 75 *Sarbecovirus* genomes (30,115 nucleotides). For convenience, the scale is the same as the coloured genomic bootstrap (CGB) barcodes shown in Figures 3 and 5. Abbreviations: *E*: envelope gene; *M*: membrane gene; *N*: nucleocapsid gene; *RdRp*: RNA-dependent RNA polymerase gene; *S*: spike gene; 1ab: ORF (Open Reading Frame) 1ab; 3a: ORF3a; 6: ORF6; 7ab: ORF7ab; 8: ORF8; 10: ORF10. The alternating yellow and orange colours in ORF1ab indicate different non-structural proteins, including *RdRp*.

Two different approaches were used for phylogenetic reconstruction: whole-genome bootstrap (WGB) analysis *versus* SuperTRI bootstrap (SB) analysis. On the one hand, a classical ML approach was applied with the RAxML method using the whole-genome alignment and 1000 bootstrap replicates. The WGB 50% majority-rule consensus tree is shown in Figure 2. On the other hand, five SWB analyses were conducted using five different window sizes (400, 500, 600, 1000 and 2000 nt) and the results were used to reconstruct five SuperTRI bootstrap (SB) consensus trees (with MBP values indicated at the nodes). The five SB trees were found to be very similar, except for a few nodes. For instance, Rs4237 and Rs4247 are sister viruses in SB_{400}, SB_{500}, SB_{600} trees, whereas Rs4247 is more closely related to Rs4081 in SB_{1000} and SB_{2000} trees. Additionally, MBP values were generally lowest for SB_{400} nodes and highest for SB_{2000} nodes. Therefore, we only reported MBP_{400} and MBP_{2000} values at the nodes of the WGB tree of Figure 2. The MBP values written in black indicate nodes recovered monophyletic in all the five SB trees, whereas the MBP values highlighted in red indicate nodes not found monophyletic in SB trees. In Figure 2, we also showed the GB barcodes built from SWB_{400} and SWB_{2000} analyses for all nodes supported by $BP_{WG} \geq 90$ or recovered monophyletic in SB trees.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

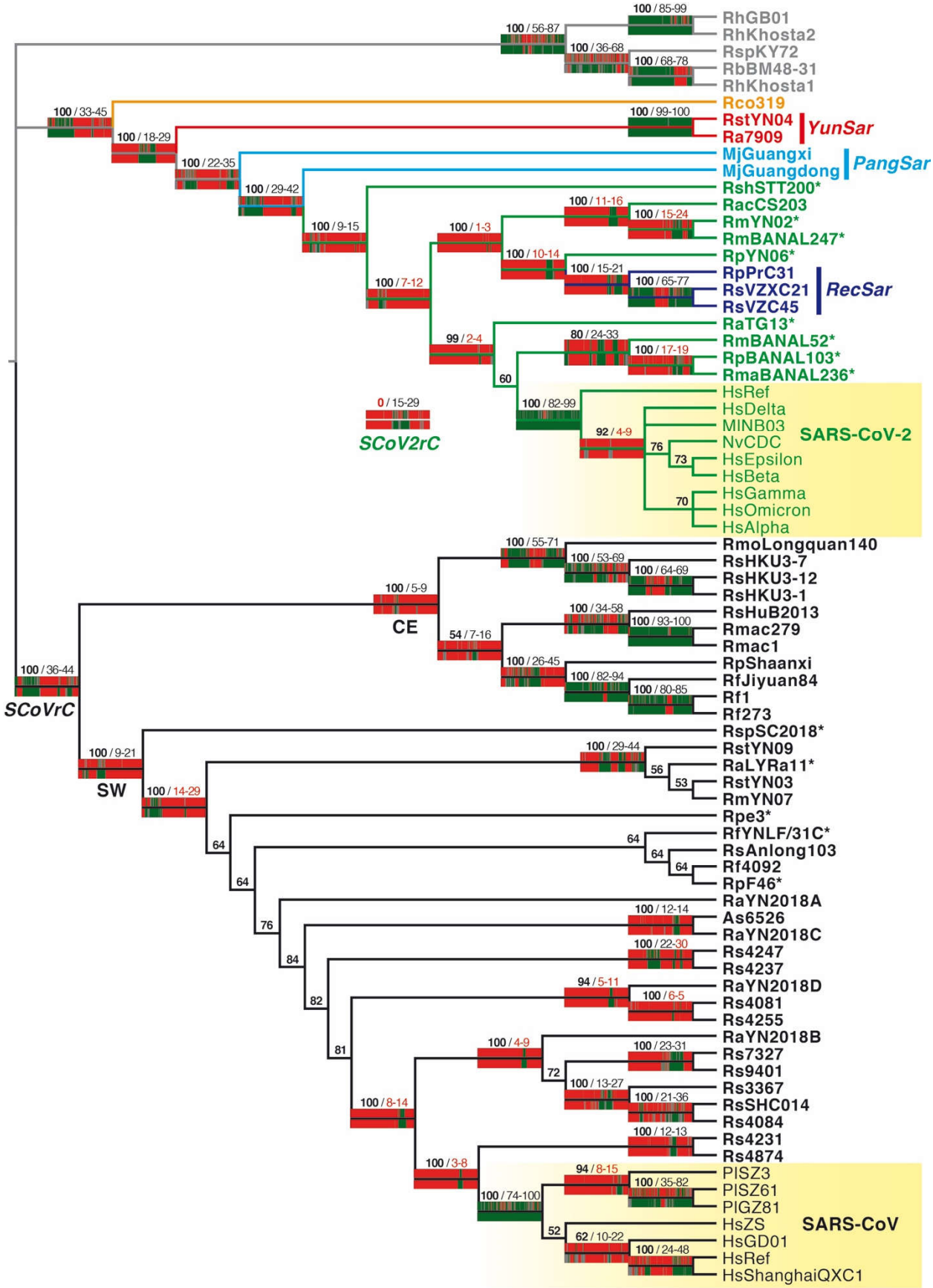

Figure 2. Whole-genome bootstrap (WGB) tree reconstructed from the alignment of 75 *Sarbecovirus* genomes. This is a 50% majority-rule consensus reconstructed after 1000 bootstrap replicates using the RAxML method and a GTR+G model for each of the four partitions of the whole-genome

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

alignment (three codon positions and non-coding regions). The bootstrap percentages (BP_{WG}) are given for all nodes in bold. The genomic bootstrap (GB) barcodes built from SWB_{400} and SWB_{2000} outputs and the BBC program [9] were reported for all nodes recovered monophyletic in the Super-TRI bootstrap (SB) trees reconstructed from SWB analyses (for these nodes, MBP_{400} and MBP_{2000} values are indicated in black at the right of the slash) and for all nodes supported by $BP_{WG} \geq 90$ that were not found monophyletic in SB trees (for these nodes, MBP_{400} and MBP_{2000} values are indicated in red at the right of the slash). The GB_{400} and GB_{2000} barcodes are displayed above and below the branches, respectively. The colours of Asian sarbecoviruses indicate to which group of synonymous nucleotide composition they belong [16]: black for SARS-CoV related coronaviruses (*SCoVrC*), green for SARS-CoV-2 related coronaviruses (*SCoV2rC*), dark blue for *RecSar* viruses showing evidence of genomic recombination between *SCoVrC* and *SCoV2rC*, light blue for pangolin sarbecoviruses, red for *YunSar* viruses, and orange for the Rco319 virus from Japan. For comparison, the GB_{400} and GB_{2000} barcodes supporting the monophyly of *SCoV2rC* are shown in the middle of the figure.

In the WGB tree of Figure 2, 48 out of the total 69 nodes (70%) are supported by $BP_{WG} \geq 90$. Robustness comparisons with SB trees revealed that 24 of these 48 robust nodes (50%) are associated with $MBP_{400/2000} \leq 30$, indicating that the phylogenetic signal is restricted to only one or several regions of the genome alignment, representing in total less than 30% of the whole genome alignment [9]. For examples, the monophyly of *RecSar*, which includes the three bat viruses RpPrC31, RsVZC45, and RsVZXC21, was supported by $MBP_{400} = 15$ and $MBP_{2000} = 21$; the sister-group relationship between SARS-CoV and Rs4231 + Rs4874 was supported by $MBP_{400} = 3$ and $MBP_{2000} = 8$; and the grouping of SARS-CoV-2 with RaTG13, RmBANAL52, RmaBANAL236, and RpBANAL103 was supported by $MBP_{400} = 2$ and $MBP_{2000} = 4$. Note that the later node was not recovered monophyletic in the five SB trees reconstructed from SWB analyses (MBP values written in red in Figure 2). For all these nodes, the GB barcodes revealed that the phylogenetic signal is restricted to one or several small genomic regions (in green, $BP \geq 70$), whereas most other genomic regions (in red, $BP \leq 30$) can support different phylogenetic relationships, as illustrated with CGB barcodes described in sections 3.2 and 3.3.

Seventeen of the 48 robust nodes (35%; $BP_{WG} \geq 90$) of the WGB tree of Figure 2 are supported by $MBP_{400/2000}$ comprised between 30 and 70, indicating that the phylogenetic signal covers many regions or one large region of the genome alignment. A first example concerns the node supporting the monophyly of SARS-CoV related coronaviruses (*SCoVrC*; $MBP_{400/2000} = 36/44$), a large group containing four human SARS-CoV, three civet SARS-CoV, and 37 bat viruses (written in black in Figure 2). As second example is the node supported by $MBP_{400/2000} = 30/45$, which includes five viral lineages showing different synonymous nucleotide composition [16] (highlighted by different colours in Figure 2): (i) SARS-CoV-2 related coronaviruses (*SCoV2rC*, in green); (ii) the three viruses showing evidence of genomic recombination between *SCoVrC* and *SCoV2rC* (*RecSar*, in dark blue); (iii) the two pangolin sarbecoviruses (in light blue); (iv) the bat sarbecoviruses from Yunnan showing a very divergent synonymous nucleotide composition (*YunSar*, in red); and (v) the Rco319 virus from Japan (in orange).

Only seven of the 48 robust nodes (15%; $BP_{WG} \geq 90$) of the WGB tree of Figure 2 are supported by $MBP_{400/2000} \geq 70$, indicating that the phylogenetic signal is present almost everywhere in the genome alignment, as shown by GB barcodes, which are green (supported by $BP \geq 70$) in most regions (GB_{400}) or all regions (most GB_{2000} barcodes) of the alignment. They include (i) the monophyly of SARS-CoV, which contains four human sequences and three civet sequences ($MBP_{400/2000} = 74/100$); (ii) the monophyly of SARS-CoV-2, which contains seven human sequences and two mink sequences ($MBP_{400/2000} = 82/99$); (iii) the monophyly of *YunSar*, which is represented by two bat sarbecoviruses from Yunnan, Ra7909 and RstYN04 ($MBP_{400/2000} = 99/100$); (iv) the sister relationship between RhGB01 and RhKhosta2 ($MBP_{400/2000} = 85/99$); (v) the sister relationship between Rmac1 and Rmac279 ($MBP_{400/2000} = 93/100$); (vi) the sister relationship between Rf1 and Rf273 ($MBP_{400/2000} = 80/85$); and (vii) their grouping with RfJiyuan84 ($MBP_{400/2000} = 82/94$). The nine SARS-CoV-2 genomes were sampled between December 2019 (HsRef, NC_045512) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

July 2022 (HsOmicron, OP010674), and we know that their most recent common ancestor (MRCA) emerged a few weeks before the end of 2019. Due to their recent divergence, these viruses have very similar genomes. Indeed, pairwise nucleotide distances are comprised between 0.04% and 0.28%. We also found very similar distances for most other closely-related viruses supported by high MBP values: between 0.11% and 0.41% for the seven SARS-CoV genomes; 0.46% between Rmac1 and Rmac279; 0.76% between Rf1 and Rf273. All these pairwise distances are the smallest calculated for our dataset. In contrast, the two other virus pairs supported by $MBP_{400/2000} \geq 70$ were found to be more distant: 2.0% between Ra7909 and RstYN04, and 12.5% between RhGB01 and RhKhosta2.

Interrelationships between the seven human and two mink SARS-CoV-2 sequences were not robust in the WGB tree of **Figure 2**, except the basal position of HsRef (BP_{WG} = 92) which was supported by one GRPS (pos. 14251-14750, with an exclusive Uracil in pos. 14,507). However, we found discordant SARS-CoV-2 relationships. For instance, HsGamma and HsOmicron were related with HsDelta based on pos. 9,851-10,350 (including an exclusive Uracil in pos. 10,127) or with HsAlpha based on pos. 28,801-29,350 (including an exclusive Cytosine in pos. 29,116).

Interrelationships between the four human and three civet SARS-CoV sequences were not robust in the WGB tree of **Figure 2**, except three nodes: (i) the sister relationship between PlSZ61 and PlGZ81 (BP_{WG} = 100), which was supported by 19 GRPS (representing 52% of the WG alignment) containing six exclusive nucleotides (pos. 4,270: A ; pos. 9,503: C; pos. 18,349: C; pos. 24,083: U; pos. 25640: A; pos. 27,499: G); (ii) the monophyly of civet SARS-CoV (PlSZ3, PlSZ61, and PlGZ81; BP_{WG} = 94), which was supported by two GRPS (pos. 25,001-26,200 and 26,401-26,950; 6% of the alignment) containing an exclusive Guanine in pos. 25,276; and (iii) the grouping of HsRef with HsShangaiQXC1 (BP_{WG} = 100), which was supported by 10 GRPS (26% of the alignment) containing three exclusive nucleotides (pos. 9,647: U ; pos. 23,497: G ; pos. 28,156: U). By comparison, the monophyly of human SARS-CoV (BP_{WG} = 52) was supported by two GRPS (pos. 22,651-24,200 and 25,251-25,750; 7% of the alignment) containing an exclusive Uracil in pos. 25,585. However, we found some GRPS supporting discordant relationships. In particular, the monophyly of human SARS-CoVs was contradicted by one GRPS supporting the grouping of the three civet SARS-CoVs with HsRef and HsShanghaiQXC1 (pos. 25801-26550; containing an exclusive Adenine in pos. 26,143), one GRPS supporting the grouping of PlSZ3 with HsRef, HsShanghaiQXC1, and HsGD0 (pos. 2,451-2,950; containing an exclusive Guanine in pos. 12,750), as well as two GRPS supporting the grouping of PlSZ61 and PlGZ81 with HsRef, HsShanghaiQXC1, and HsGD0 (pos. 17,651-18,100 and 20,901-21,400; containing an exclusive Guanine in pos. 21,161).

3.2. Coloured genomic bootstrap barcodes reconstructed for the ancestor of SARS-CoV

The phylogenetic, host and geographic CGB barcodes constructed for the common ancestor of SARS-CoV are shown in Figure 3. The phylogenetic CGB barcodes indicate that 25.6% of the SARS-CoV genome shares exclusive ancestry with five *Rhinolophus* viruses (bipartitions +1 in Figure 3) detected in three provinces of Southwest China: three viruses from Yunnan, including RfYNLF/31C sampled in *R. ferrumequinum* (eight GRPS representing 17.9% of the WG alignment), RpF46 sampled in *R. pusillus* (two GRPS; 3.3% of the alignment); RaLYRa11 sampled in *R. affinis* (one GRPS; 1.0% of the alignment); one virus from Guangxi, Rpe3 sampled in *R. pearsoni* (one GRPS; 1.3% of the alignment); and one virus from Sichuan, RspSC2018, sampled in an unidentified species of *Rhinolophus* (two GRPS; 2.0% of the alignment). If we consider the bipartitions uniting SARS-CoV with two to five closely-related viruses (n = +2, +3, +4 or +5 in Figure 3), they involve eight additional viruses, all found in *Rhinolophus* bats collected in Yunnan: RmYN07, Rs4231, Rs4237, Rs4247, Rs4874, Rs7327, Rs9401, and RstYN09. These results therefore suggest that SARS-CoV originated from horseshoe bat (genus *Rhinolophus*) viruses. This hypothesis was confirmed with our analyses of Figure 4B, which revealed that 100% of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

phylogenetic CGB barcodes reconstructed for SARS-CoV involved *Rhinolophus* viruses. As shown in Figure 4A, the most important contributors are five sarbecoviruses extracted from horseshoe bats: RfYNLF/31C (which is 65% involved), Rs7327 (63%), Rs9401 (63%), Rs4874 (62%), and Rs4084 (61%). Importantly, all the 22 viruses showing a significant contribution in SARS-CoV CGB barcodes (≥25%) belong to the *SCoVrC* lineage.

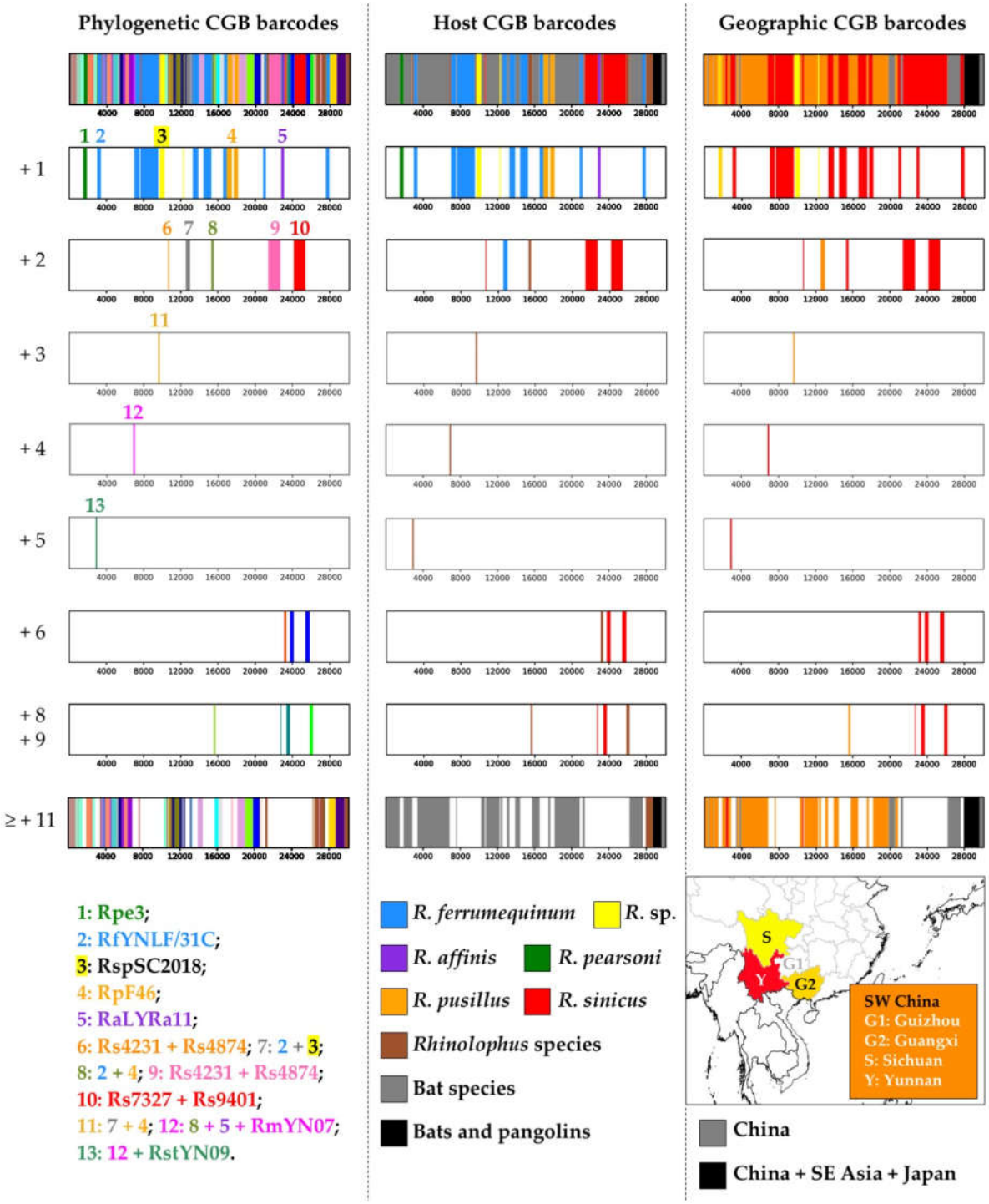

Figure 3. Coloured genomic bootstrap (CGB) barcodes constructed for the common ancestor of SARS-CoV. At the left part of the figure are shown phylogenetic CGB barcodes, in which the best phylogenetic signals are represented by different colours. To facilitate interpretation, we have also shown versions reduced to the bipartition categories +1 (n = 5), +2 (n = 5), +3 (n = 1), +4 (n = 1), +5 (n =1), +6 (n = 2), +8 and +9 (n = 4), and all bipartitions uniting SARS-CoV sequences with at least 11 other viruses (n = 30). The bat sarbecoviruses included in the 13 smallest bipartitions (categories +1

to +5) are detailed at the bottom. Similarly, the full and reduced versions of host and geographic CGB barcodes are shown in the middle and right parts of the figure. The colour codes used for host taxa and geographic areas are provided at the bottom.

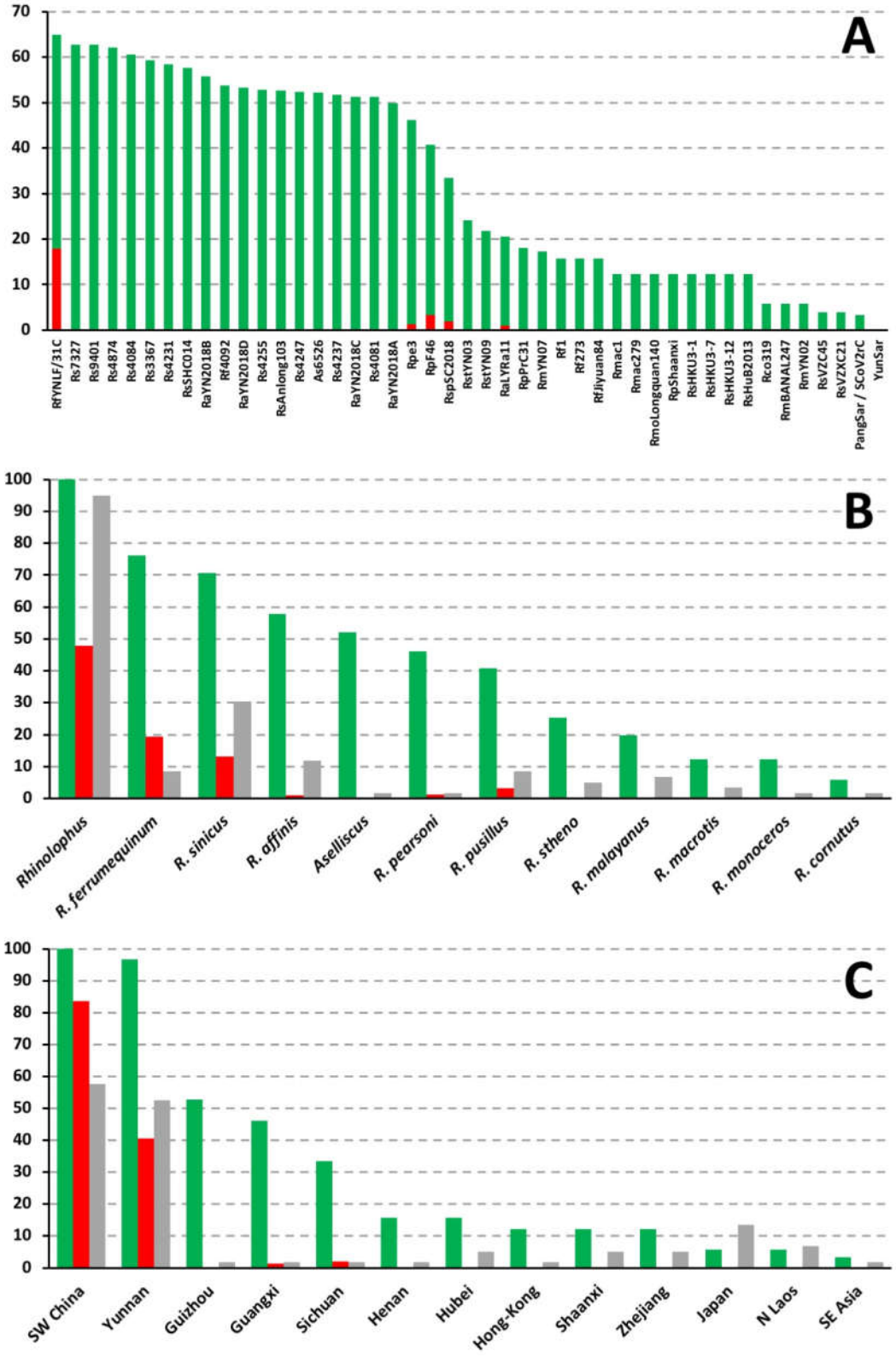

Figure 4. Percentages of whole-genome alignment including phylogenetic CGB barcodes (green histograms) shared between SARS-CoV and several bat sarbecoviruses (A) detected in different host taxa (B, green histograms) and geographic regions (C, green histograms). The red histograms

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

indicate the percentages of exclusive ancestry found in bat sarbecoviruses (A), host taxa (B), and geographic regions (C). To assess sampling efforts, the grey histograms show the proportions of bat viruses used in our dataset for the host taxa and geographic regions of interest, respectively.

Several regions of the SARS-CoV genome were found to be exclusively related to RfYNLF/31C, a virus found in *Rhinolophus ferrumequinum*. These regions represent 17.9% of the whole-genome alignment and include several dispersed fragments in ORF1ab, including two fragments in the *RdRp* gene (RfYNLF/31C; pos. 13,301-13,900 and 14,451-15,300), as well as almost the complete ORF7a (RfYNLF/31C; pos. 27,601-27,950). The genomic contribution of the other species is much lower (≤ 3.3%), except for *Rhinolophus sinicus* (13.1%), as most parts of the *Spike* gene of SARS-CoV were found to be closely related to viruses detected in this species (Rs4231 + Rs4874 in pos. 21,451-22,700; Rs7327 + Rs9401 in pos. 24,151-25,400; and Rs3367 + Rs4084 + Rs4874 + RsSHC014 + Rs7327 + Rs9401 in pos. 23,751-24,150 and 25,401-25,850). Note however that a small fragment of the *Spike* gene was found to be linked to a virus detected in *R. affinis* (RaLYRa11; pos. 22,801-23,100). Overall, the results presented in Figure 4B show that the three *Rhinolophus* species with the highest contribution to the phylogenetic CGB barcodes of SARS-CoV are *R. ferrumequinum* (76%), *R. sinicus* (71%), and *R. affinis* (58%).

The geographic CGB barcodes indicate that most regions of the SARS-CoV genome are exclusively related to bat viruses from Southwest China (84%; highlighted in orange, yellow and red in Figure 3), including Yunnan (41%), Sichuan (2%), and Guangxi (1%). Consistent with this, we found that 100% of the phylogenetic CGB barcodes reconstructed for SARS-CoV involved viruses detected in Southwest China, and the contribution of the Yunnan province is 97%, which is much more important than that of the three other provinces of Southwest China, i.e. Guizhou (53%), Guangxi (46%), and Sichuan (33%) (Figure 4C).

3.3. Coloured genomic bootstrap barcodes reconstructed for the ancestor of SARS-CoV-2

The phylogenetic, host and geographic CGB barcodes constructed for the common ancestor of SARS-Co2 are shown in Figure 5. The phylogenetic CGB barcodes indicate that 23.4% of SARS-CoV-2 genome shares exclusive ancestry with eight *Rhinolophus* viruses (bipartitions +1 in Figure 5): three from Yunnan, RaTG13 sampled in *R. affinis* (four GRPS representing 4.8% of the alignment), RmYN02 sampled in *R. malayanus* (two GRPS; 3.8% of the alignment), and RpYN06 sampled in *R. pusillus* (two GRPS; 5.8% of the alignment); four from northern Laos, RpBANAL103 sampled in *R. pusillus* (one GRPS; 3.3% of the alignment), RmaBANAL236 sampled in *R. marshalli* (one GRPS; 1.7% of the alignment), RmBANAL52 (one GRPS; 1% of the alignment) and RmBANAL247 (one GRPS; 0.7% of the alignment), both sampled in *R. malayanus*; and one from northern Cambodia, RshSTT200 sampled in *R. shameli* (two GRPS; 2.3% of the alignment). The bipartitions including between two and five additional viruses (n = +2, +3, +4 or +5 in Figure 5) involve only two additional viruses: one isolated from the Sunda pangolin (*M. javanica*), named MjGuangxi, and another found in *Rhinolophus acuminatus* from the Eastern Thailand, named RacCS203. These results therefore suggest that SARS-CoV-2 originated from horseshoe bat (genus *Rhinolophus*) viruses rather than pangolin viruses. This hypothesis was confirmed with the results of Figure 6B, which show that 100% of the phylogenetic CGB barcodes reconstructed for SARS-CoV-2 involved *Rhinolophus* viruses, whereas the contribution of *Manis* viruses remains modest (24%). As shown in Figure 6A, the most important contributors are six viruses extracted from horseshoe bats of North Laos and Yunnan: RpBANAL103 (which is 72% involved), RmBANAL52 (70%), RpYN06 (65%), RmaBANAL236 (65%), RmYN02 (63%), and RaTG13 (60%). Importantly, all the 10 viruses showing a significant contribution (≥25%) belong to the *SCoV2rC* lineage (between 72% and 35%), except RpPrC31 (31%), which belongs to the *RecSar* group [9,16].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

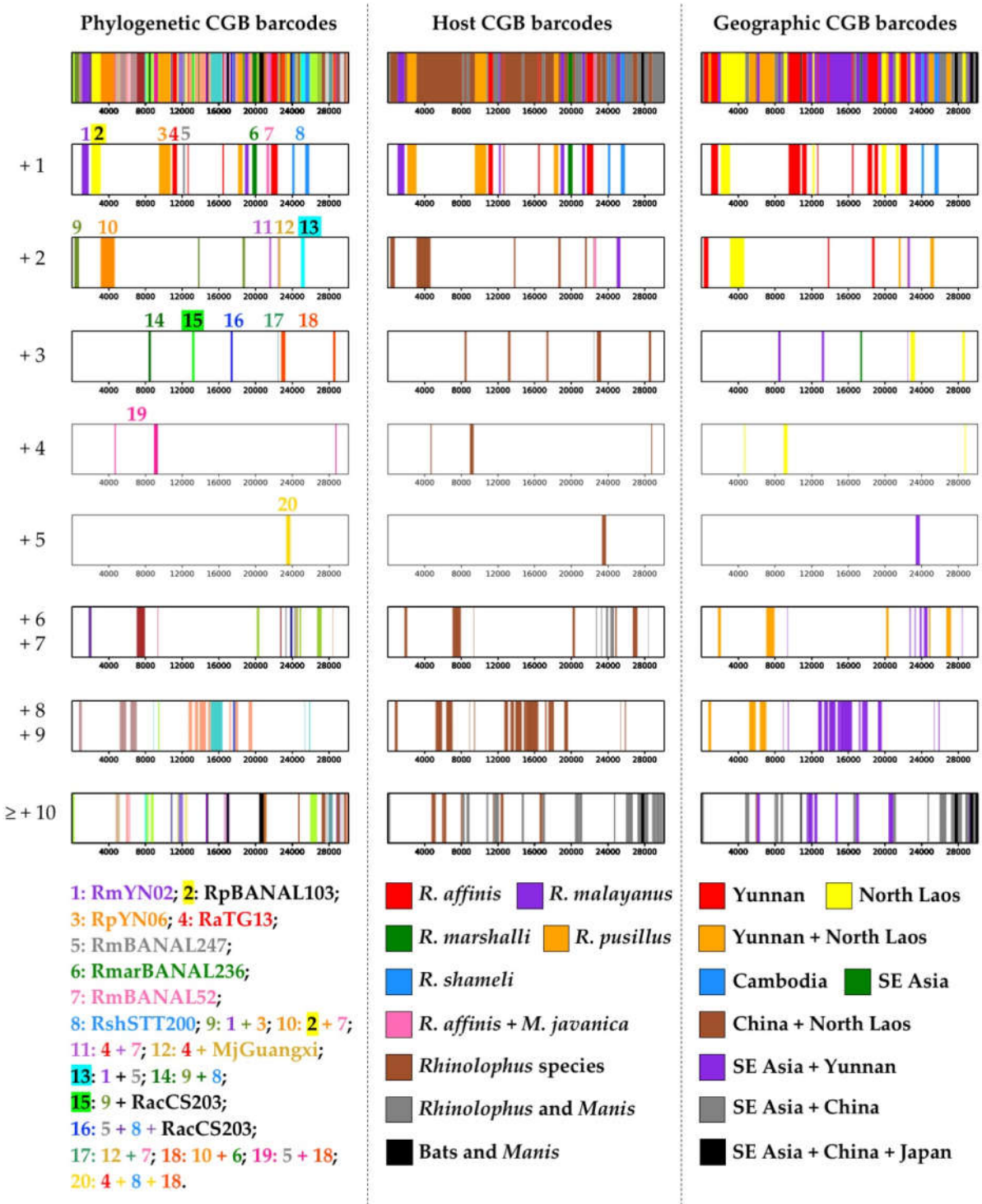

Figure 5. Coloured genomic bootstrap (CGB) barcodes constructed for the common ancestor of SARS-CoV-2. At the left part of the figure are shown phylogenetic CGB barcodes, in which the best phylogenetic signals are represented by different colours. To facilitate interpretation, we have also shown versions reduced to the bipartition categories +1 (n = 8), +2 (n = 5), +3 (n = 5), +4 (n = 1), +5 (n = 1), +6 and +7 (n = 8), +8 and +9 (n = 6), and all bipartitions uniting SARS-CoV-2 sequences with at least 10 other viruses (n = 18). The bat and pangolin sarbecoviruses included in the 20 smallest bipartitions (categories +1 to +5) are detailed at the bottom. Similarly, the full and reduced versions of host and geographic CGB barcodes are shown in the middle and right parts of the figure. The colour codes used for host taxa and geographic areas are provided at the bottom.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

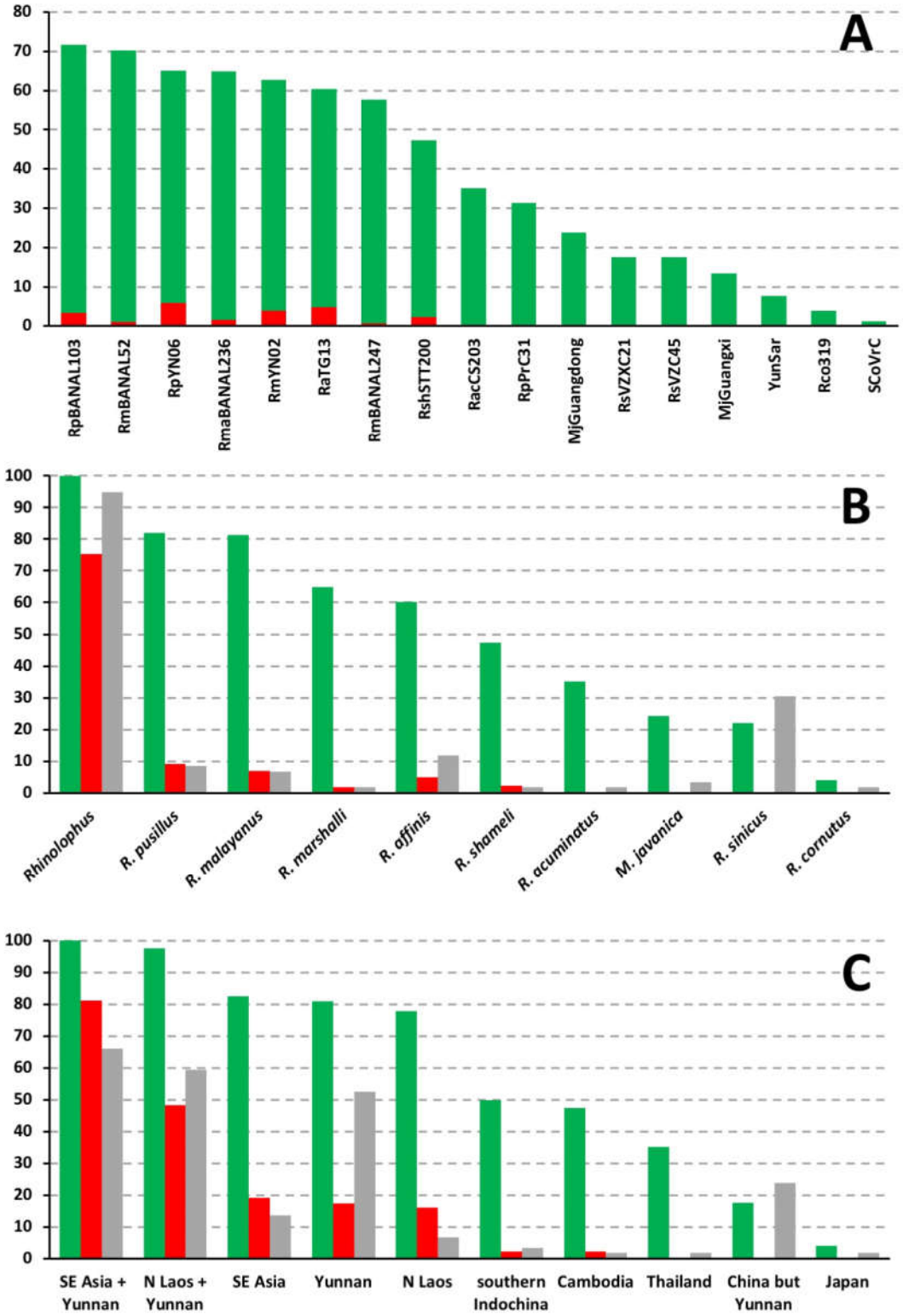

Figure 6. Percentages of whole-genome alignment including phylogenetic CGB barcodes (green histograms) shared between SARS-CoV-2 and several bat or pangolin viruses (A) detected in different host taxa (B, green histograms) and geographic regions (C, green histograms). The red histograms indicate the percentages of exclusive ancestry found in bat sarbecoviruses (A), host taxa (B), and geographic regions (C). To assess sampling efforts, the grey histograms show the proportions of sarbecoviruses used in our dataset for the host taxa and geographic regions of interest, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

Several regions of the SARS-CoV-2 genome were found to be exclusively related to viruses extracted from *R. malayanus* (6.8%; RmYN02, RmBANAL52 and RmBANAL247), *R. pusillus* (9.1%; RpYN06 and RpBANAL103), *R. affinis* (4.8%; only RaTG13), *R. shameli* (2.3%; only RshSTT200), and *R. marshalli* (1.7; only RmaBANAL236). In particular, the *Spike* gene of SARS-CoV-2 can be related to genomic fragments from viruses of different species, including *R. affinis* (RaTG13; pos. 21,701-22,400), *R. affinis* + *M. javanica* (RaTG13 + MjGuangxi; pos. 22,501-22,700), *R. shameli* (RshSTT200; pos. 24,001-24,250 and 25,401-25,850), and *R. malayanus* (RmYN02 + RmBANAL247; pos. 24,951-25,350). The results presented in Figure 6B show that the four *Rhinolophus* species with the highest contribution to the phylogenetic CGB barcodes of SARS-CoV-2 are *R. pusillus* (82%), *R. malayanus* (81%), *R. marshalli* (65%) and *R. affinis* (60%).

The geographic CGB barcodes indicate that most regions of the SARS-CoV-2 genome are exclusively related to bat viruses from Yunnan and Southeast Asia (81%), some of them being specific to Yunnan (17% highlighted in red in Figure 5) or Southeast Asia (19%; North Laos representing 16%, in yellow). Consistent with this, we found that 100% of the phylogenetic CGB barcodes reconstructed for SARS-CoV-2 involved viruses detected in Southeast Asia and Yunnan (Figure 6C). However, sampling efforts were much more important in Yunnan: 53% of the *Sarbecovirus* genomes used in our study were collected in Yunnan against 7% for those found in Laos. The contribution of bat sarbecoviruses from Laos is 78%, which is much more important than the contribution of bat sarbecoviruses found in other Southeast Asian countries: 47% for the virus from Cambodia and 35% for the virus from Thailand (Figure 6C).

4. Discussion

4.1. Phylogenetic trees versus phylogenetic CGB barcodes

One or several phylogenetic trees have been published in all reports dealing with the origin and evolution of SARS-CoV-2 [4-7,15,16]. In these studies, SARS-CoV-2 was generally found to be closely related to RaTG13, a virus detected in a *R. affinis* bat collected in Yunnan in 2013 [4-7]. However, phylogenetic analyses based on different genomic regions were found to support conflicting relationships [6,7,9,16]. This was well illustrated in the Figure 3 of Zhou *et al*. [7], in which SARS-CoV-2 appeared closely related to RaTG13, RmYN02, RpYN06, and RshSTT200 in the *RdRp* gene tree (BP = 94), sister to RmYN02 and RpYN06 in the *ORF1ab* tree (BP = 100), and linked to RaTG13 in the *Spike* gene tree (BP = 100). Our phylogenetic CGB barcodes showed that these discordances are explained by the mosaic structure of the SARS-CoV-2 genome, in which different regions support different relationships. Although SARS-CoV-2 was closely related to RaTG13 based on four genomic regions, representing only 4.8% of the WG alignment, other genomic regions provided support for 51 other phylogenetic relationships, including its grouping with RpYN06 (two regions; 5.8%), RmYN02 (two regions; 3.8%), RpBANAL103 (one region; 3.3%), RshSTT200 (two regions; 2.3%), etc (Figure 5).

Based on amino-acid sequences of the receptor-binding domain (RBD) to the cellular ACE2 receptor, Temmam *et al*. [15] published a tree showing a close relationship between SARS-CoV-2 and three *Rhinolophus* viruses detected in Laos, i.e. RmBANAL52, RpBANAL103, and RmaBANAL236. In our nucleotide whole-genome alignment, the RBD region corresponds to pos. 22,675-23,352, and phylogenetic CGB barcodes indeed confirmed that a robust phylogenetic signal exists to support the node uniting SARS-CoV-2 and the three viruses from Laos, but it is restricted to pos. 22,801-23,250 of RBD ("bipartition +3" n°18 in Figure 5). Interestingly, upstream and downstream regions support other relationships for SARS-CoV-2, including its grouping with RaTG13 and MjGuangxi (pos. 22,501-22,700) or with RshSTT200 (pos. 24,001-24,250). More generally, and focusing only on the best bipartition categories +1 to +5, other genomic regions of SARS-CoV-2 were found to be closely related to one (viruses with asterisk) or more of the following ten sarbecoviruses (20 bipartitions representing 43% of the robust phylogenetic signals in the WG alignment):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

RaTG13*, RmBANAL52*, RpBANAL103*, RmarBANAL236*, RmBANAL247*, RmYN02*, RpYN06*, RshSTT200*, MjGuangxi, and RacCS203. Although several of these bipartitions represent compatible nested bipartitions (for example, "bipartitions +1" n°1 and n°3 are nested within "bipartition +2" n°9 in Figure 5), most of them support incongruent phylogenetic relationships.

Phylogenetic CGB barcodes allowed us to detect that several gene fragments provide conflicting phylogenetic signals. The best example concerns the *Spike* gene (pos. 21,664-25,557 in our alignment) for which we found conflicting phylogenetic signals supporting the grouping of SARS-CoV-2 with RaTG13 (pos. 21,701-22,400), RshSTT200 (pos. 24,001-24,250 and 25,401-25,850), RmBANAL247 + RmYN02 (pos. 24,951-25,350), MjGuangxi + RaTG13 (pos. 22,501-22,700), RmBANAL52 + RmaBANAL236 + RpBANAL103 (pos. 22,801-23,250), and RaTG13 + RmBANAL52 + RmaBANAL236 + RpBANAL103 + RshSTT200 (pos. 23,351-23,800) (Figure 5). Similarly, we found conflicting phylogenetic signals supporting the grouping of SARS-CoV with RaLYRa11 (pos. 22,801-23,100), Rs7327 + Rs9401 (pos. 24,151-25,400), and Rs4231 + Rs4874 (pos. 21,401-22,700) (Figure 3). We conclude therefore that phylogenetic CGB barcodes provide much more reliable and accurate information than WG and gene trees for understanding the evolution of sarbecoviruses.

4.2. Intermediary hosts for SARS-CoV and SARS-CoV-2?

While many studies agree that bats are the main reservoir host for sarbecoviruses, the role of other mammals as possible intermediate hosts between bats and humans, such as small carnivores and pangolins, remains unclear and controversial [4-7,9,14,15,22]. To better understand this issue, we have included in our analyses several sarbecoviruses sequenced from captive mammals, including pangolins, minks and civets (supplementary Table S1).

In theory, pangolins could be contaminated in their natural habitat by pathogens circulating in horseshoe bats because both taxa are occasionally found together in hollow trees, burrows and possibly caves [14,22]. In contrast to bats, which are considered as asymptomatic for coronavirus, pangolins were found to be highly sensitive to sarbecovirus [23]. Also, pangolins are not gregarious like *Rhinolophus* bats; they are solitary species and the female and male meet only for reproduction. Therefore, most pangolins infected by bat sarbecovirus in the wild should be considered as evolutionary dead ends for the virus. However, the situation has changed a lot with the intense pangolin trafficking during the decade before COVID-19, as Sunda pangolins (*M. javanica*) imported illegally into China became infected in captivity [14]. By this way, at least two different pangolin sarbecoviruses were exported from Southeast Asia to China, MjGuangxi before 2017 and MjGuangdong before 2019 [14,22-24]. Despite this, we did not find any evidence of exclusive ancestry between SARS-CoV-2 and the two pangolin viruses. Moreover, the synonymous nucleotide compositions of MjGuangxi and MjGuangdong genomes were found to be similar but divergent from those of SARS-CoV-2 and bat *SCoV2rC* viruses [16], suggesting that the two pangolin viruses have evolved independently and for some time in pangolin populations either in the wild or in captivity. Although pangolins may be intermediary hosts between bats and humans for some viruses, current data and our findings do not support their involvement in the case of SARS-CoV-2.

Recent studies have provided strong evidence that SARS-CoV-2 was introduced from humans to domestic or captive carnivores, including dogs, cats, lions, tigers, and minks, and that the virus evolved very rapidly in mink farms, with several back transmissions from infected animals to humans [25,26]. The common ancestor of human and mink SARS-CoV-2 genomes was supported by high MBP values (82/99) confirming that a strong phylogenetic signal is present in all parts of the alignment (as highlighted by the green colour of GB_{400} and GB_{2000} barcodes in Figure 2). In addition, we found no genomic region providing support for the paraphyly or polyphyly of SARS-CoV-2. These results therefore

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

confirmed that human and mink SARS-CoV-2 genomes included in our study share the same MRCA, fully isolated genetically from bat *Sarbecovirus* lineages since its emergence in December 2019 or a few weeks earlier. Our analyses also revealed that different genomic regions can support conflicting relationships between SARS-CoV-2 sequences. The best example concerns the placement of Gamma and Omicron variants, which were related to the Delta variant based on pos. 9,851-10,350 or to Alpha variant based on pos. 28,801-29,350. This discordant pattern could be due to past genomic recombination, as several recombinants have already been reported in human populations [11,12].

The masked palm civet (*Paguma larvata*) was identified as the possible intermediate host transmitting SARS-CoV to humans during the SARS epidemic, which began on November 2002 in Foshan, a city about 20 km from Guangzhou (province of Guangdong, China) [13,27]. In our analyses, we have included three SARS-CoV-like genomes detected in civets maintained in captivity in the Guangdong province: two came from a wildlife market of Shenzhen before May 2003 (PlSZ3 and PlSZ61) and another was sampled in a restaurant of Guangzhou in 2003 (PlGZ81). In the WG tree of Figure 2, it is worth noting that human and civet SARS-CoVs are enclosed into a robust clade (BP_{WG} = 100) also supported by high MBP values (74/100), which indicates that the phylogenetic signal is strong in all parts of the alignment. These results therefore confirmed that human and civet SARS-CoV genomes share the same MRCA, which was fully isolated genetically from bat *Sarbecovirus* lineages. Although the node grouping the four human SARS-CoVs was not robust in the WG tree (BP_{WG} = 52), it was supported by two GRPS, one including the RBD region (pos. 22,651-24,200 in our alignment) and another covering the C-terminus of the Spike protein and the N-terminus of ORF3a (pos. 25,251-25,750). Interestingly, there are two RBD amino-acids characterizing the four human SARS-CoVs included in our alignment, F360 and T487. The last one was found to increase by 20-fold the RBD affinity for human ACE2, suggesting therefore a specific adaptation to enhance human-to-human transmission during the 2002-2003 SARS-CoV outbreak [28]. In contrast, all human and animal viruses sequenced during the 2003–2004 SARS-CoV outbreak had a Serine (instead of Threonine) at this position of the Spike protein [13,28], indicating that this new episode may have resulted from an independent viral invasion from animal to human [29]. Our analyses have shown that several discordant phylogenetic signals involving the para- or polyphyly of human SARS-CoVs were detected in other genomic regions, such as the grouping of the three civet SARS-CoVs with HsRef and HsShanghaiQXC1 or the grouping of PlSZ3 with HsRef, HsShanghaiQXC1 and HsGD01. Therefore, these results suggest that human and civet SARS-CoV were involved in past genomic recombination events in the same host and that humans and captive civets have exchanged SARS-CoV viruses. In other words, the situation was probably very similar to that recently observed for SARS-CoV-2 between humans and captive minks [25,26]. Due to the low divergence separating human and civet genomes (between 0.16% and 0.41%), however, we cannot rule out two other hypotheses involving either nucleotide homoplasy (due to convergence(s) and reversion(s)) or sequencing and genome assembly errors.

4.3. Species involved as reservoir hosts for the ancestors of SARS-CoV and SARS-CoV-2

As discussed previously [9,10], most conflicting phylogenetic signals in different genomic regions of our *Sarbecovirus* alignment can be explained by multiple past events of recombination at different periods of time and involving viruses circulating in multiple reservoir species of horseshoe bats (genus *Rhinolophus*). Host CGB barcodes obtained for both SARS-CoV and SARS-CoV-2 (Figures 3 and 5) clearly confirm that horseshoe bats are the animal reservoir in which coronaviruses related to either SARS-CoV (*SCoVrC*) or SARS-CoV-2 (*SCoV2rC*) evolve and diversify. First, all animal sarbecoviruses showing exclusive ancestry with human sarbecoviruses were extracted from *Rhinolophus* species. Second, 100% of the phylogenetic CGB barcodes reconstructed for either SARS-CoV or SARS-CoV-2 involved *Rhinolophus* viruses. The *Rhinolophus* species showing the highest

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 December 2022 doi:10.20944/preprints202212.0232.v1

contribution are *R. ferrumequinum* (76%), *R. sinicus* (71%), and *R. affinis* (58%) for SARS-CoV, and *R. pusillus* (82%), *R. malayanus* (81%), *R. marshalli* (65%) and *R. affinis* (60%) for SARS-CoV-2 (Figures 4B and 6B). All these results corroborate high and recent evolutionary dynamics of genomic recombination between sarbecoviruses circulating in several *Rhinolophus* species. As pointed out in previous publications [9,16,22], recombination implies that whole or partial genomes of two divergent viruses co-exist in a cell of the same host. Such a situation is expected to occur frequently in horseshoe bats because interspecific transmission of sarbecoviruses could be favoured by their behaviour, as several *Rhinolophus* species often nest in colonies in the same cave, and by their cave habitat, in which viral contamination could be facilitated by promiscuity within and between bat colonies, high humidity levels and cool temperatures all year round.

The ecological niches (i.e. geographic distributions predicted using climatic parameters) of bat sarbecoviruses related to either SARS-CoV (*SCoVrC*) or SARS-CoV-2 (*SCoV2rC*) have been reconstructed using an original approach combining genetic data on both viruses and bat species [22]. The results have shown that the ecological niche of *SCoVrC* extends from Southwest China and northern Myanmar, through northern Vietnam and Central China to East China, Korea and southern Japan, whereas the ecological niche of *SCoV2rC* includes four main different regions of Southeast Asia: (i) northern Laos and bordering regions; (ii) southern Laos, southwestern Vietnam, and northeastern Cambodia; (iii) the East region of Thailand and southwestern Cambodia; and (iv) the Dawna Range in central Thailand and southeastern Myanmar. Since the geographic ranges of *SCoVrC* and *SCoV2rC* are different, we assumed that these two groups of viruses generally do not circulate in the same *Rhinolophus* species assemblages. In agreement with that, geographic CGB barcodes indicate that 84% of the SARS-CoV genome showed exclusive ancestry with bat viruses from Southwest China (Figure 4C), whereas 81% of the SARS-CoV-2 genome show exclusive ancestry with bat viruses from Southeast Asia and Yunnan (Figure 6C). In addition, host CGB barcodes show that SARS-CoV shares exclusive ancestry with several viruses found in three bat species mainly distributed in China [30]: 19.4% for *R. ferrumequinum*, 13.1% for *R. sinicus*, and 1.3% for *R. pearsoni* (Figure 4B), whereas SARS-CoV-2 shares exclusive ancestry with several viruses found in three bat species mainly distributed in Southeast Asia or endemic to this region [30]: 6.8% for *R. malayanus*, 2.3% for *R. shameli*, and 1.7% for *R. marshalli* (Figure 6B).

However, the detection of recombinant viruses between *SCoVrC* and *SCoV2rC* lineages, named *RecSar* [9,16], has revealed that some bats were simultaneously infected by the two divergent virus lineages in the past. Interestingly, different viruses showing exclusive ancestry with either SARS-CoV or SARS-CoV-2 were isolated from *R. pusillus* (representing 3.3% and 9.1%, respectively) and also from *R. affinis* (representing 1.0% and 4.8%, respectively). Since these two *Rhinolophus* species are widely distributed in both China and Southeast Asia [30], their dispersal capacity is expected to be greater than that of other *Rhinolophus* species endemic to either China or Southeast Asia. Therefore, it can be argued that the rare events of genomic recombination between *SCoVrC* and *SCoV2rC* lineages may have occurred in *R. affinis* and *R. pusillus* bats, and most likely in the region where the ecological niches of *SCoVrC* and *SCoV2rC* overlap, i.e. Yunnan and adjacent regions in northern Laos [22].

 doi:10.20944/preprints202212.0232.v1

4.4. Conclusion and perspectives

Exclusive ancestry with human sarbecoviruses has currently been found in sarbecoviruses collected in the following *Rhinolophus* species (Figures 3 and 5): *R. ferrumequinum*, *R. sinicus*, and *R. pearsoni* for SARS-CoV; *R. malayanus*, *R. shameli*, and *R. marshalli* for SARS-CoV-2; and *R. pusillus* and *R. affinis* for both SARS-CoV and SARS-CoV-2. These results provide strong evidence that viral transmission within and between bat colonies of these species as well as genomic recombination have participated to the emergence of human sarbecoviruses. However, the lists of involved reservoir species cannot be considered exhaustive due to limited investigations for detecting sarbecoviruses in bats, particularly in Southeast Asia where the diversity of *Rhinolophus* species is much higher than anywhere else in the Old World [30]. Although genome data currently available support an origin of SARS-CoV in horseshoe bats of Yunnan and an origin of SARS-CoV-2 in horseshoe bats of Yunnan and northern Laos, the two hypotheses need to be further investigated by exploring *Sarbecovirus* diversity in bats of Cambodia, Laos, Myanmar, Thailand and Vietnam. Indeed, only a few bat viruses were recently published from limited number of localities in northern Cambodia (two highly similar genomes of the same virus collected in one locality) [5], eastern Thailand (five highly similar genomes of the same virus collected in one locality) [6] and northern Laos (five viruses; four localities) [15], and no sarbecovirus has yet been described from bats of Myanmar and Vietnam.

Another problem that limits our interpretation on reservoir hosts is that most studies dealing with the evolution of sarbecoviruses have generally provided very little information about the bats and caves in which the viruses were found. Recently, a new *Sarbecovirus* genome, RspSC2018 (GenBank accession: MK211374), was described from an unidentified bat at the species level collected from an unknown locality in Sichuan province [31]. Since Illumina reads were not deposited in any of the international Sequence Read Archive (SRA) databases, it was impossible to know from which host species the virus was extracted. Moreover, the quality of the virus genome assembly cannot be verified, which may also compromise some future evolutionary studies.

Our poor knowledge of horseshoe bat taxonomy also hinders understanding of their role as *Sarbecovirus* reservoir. For instance, some taxonomists have proposed to split *R. ferrumequinum* into two different species: *R. ferrumequinum* in Europe and West Asia, and *Rhinolophus nippon* in East Asia [32]. Similarly, a molecular study based on a set of ~1500 nuclear loci has suggested that the *R. sinicus* complex may contain three different species [33]: *R. sinicus*, apparently distributed from northern Vietnam to East China, through Hainan Island and Central China; *Rhinolophus septentrionalis* in Yunnan; and an undescribed species of *Rhinolophus* in Vietnam.

Supplementary Materials: Table S1: Origin of the *Sarbecovirus* genomes used in this study.

Author Contributions: Conceptualization, A.H.; methodology, A.H.; software, O.R.; formal analysis, A.H, and O.R.; investigation, A.H. and O.R.; writing—original draft preparation, A.H.; visualization, A.H and O.R.; supervision, A.H.; project administration, A.H.; funding acquisition, A.H. The two authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the "*Agence nationale de la recherche*" (AAP RA-COVID-19, grant number ANR-21-CO12-0002).

Data Availability Statement: The whole-genome alignment, the five SWB output files, the five SuperTRI MRP files, the five SB trees (SB_{400}, SB_{500}, SB_{600}, SB_{1000}, and SB_{2000}), all the 5,594 GB barcodes, and all the 303 CGB barcodes constructed in this study are available in the Open Science Framework (OSF) platform at https://osf.io/XXXXX/ (the link will be provided for the final version). The SWB, BBC, CGB and LFG programs are available at https://github.com/OpaleRambaud/GB_barcodes_project.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J., Wolthers, K. C., Wertheim-van Dillen, P. M., Kaandorp, J., Spaargaren, J., & Berkhout, B. (2004). Identification of a new human coronavirus. *Nature medicine*, 10(4), 368–373. https://doi.org/10.1038/nm102
2. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D., & Fouchier, R. A. (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *The New England journal of medicine*, 367(19), 1814–1820. https://doi.org/10.1056/NEJMoa1211721
3. Woo, P. C., Lau, S. K., Lam, C. S., Lau, C. C., Tsang, A. K., Lau, J. H., Bai, R., Teng, J. L., Tsang, C. C., Wang, M., Zheng, B. J., Chan, K. H., & Yuen, K. Y. (2012). Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. Journal of virology, 86(7), 3995–4008. https://doi.org/10.1128/JVI.06540-11
4. Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. D., Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., Zheng, X. S., … Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7
5. Delaune, D., Hul, V., Karlsson, E. A., Hassanin, A., Ou, T. P., Baidaliuk, A., Gámbaro, F., Prot, M., Tu, V. T., Chea, S., Keatts, L., Mazet, J., Johnson, C. K., Buchy, P., Dussart, P., Goldstein, T., Simon-Lorière, E., & Duong, V. (2021). A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nature communications*, 12(1), 6563. https://doi.org/10.1038/s41467-021-26809-4
6. Wacharapluesadee, S., Tan, C. W., Maneeorn, P., Duengkae, P., Zhu, F., Joyjinda, Y., Kaewpom, T., Chia, W. N., Ampoot, W., Lim, B. L., Worachotsueptrakun, K., Chen, V. C., Sirichan, N., Ruchisrisarod, C., Rodpan, A., Noradechanon, K., Phaichana, T., Jantarat, N., Thongnumchaima, B., Tu, C., … Wang, L. F. (2021). Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nature communications*, 12(1), 972. https://doi.org/10.1038/s41467-021-21240-1
7. Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Wang, Q., Hu, T., Song, H., Zhao, R., Chen, Y., Cui, M., Zhang, Y., Hughes, A. C., Holmes, E. C., & Shi, W. (2021). Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell, S0092-8674(21)00709-1. https://doi.org/10.1016/j.cell.2021.06.008
8. Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? *Nature reviews. Microbiology*, 9(8), 617–626. https://doi.org/10.1038/nrmicro2614
9. Hassanin, A., Rambaud, O., & Klein, D. (2022). Genomic bootstrap barcodes and their application to study the evolution of sarbecoviruses. Viruses, 14(2), 440. https://doi.org/10.3390/v14020440
10. Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., Rambaut, A., & Robertson, D. L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature microbiology*, 5(11), 1408–1417. https://doi.org/10.1038/s41564-020-0771-4
11. Jackson, B., Boni, M. F., Bull, M. J., Colleran, A., Colquhoun, R. M., Darby, A. C., Haldenby, S., Hill, V., Lucaci, A., McCrone, J. T., Nicholls, S. M., O'Toole, Á., Pacchiarini, N., Poplawski, R., Scher, E., Todd, F., Webster, H. J., Whitehead, M., Wierzbicki, C., COVID-19 Genomics UK (COG-UK) Consortium, … Rambaut, A. (2021). Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell, 184(20), 5179–5188.e8. https://doi.org/10.1016/j.cell.2021.08.014
12. Focosi, D., & Maggi, F. (2022). Recombination in coronaviruses, with a focus on SARS-CoV-2. Viruses, 14(6), 1239. https://doi.org/10.3390/v14061239
13. Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., Butt, K. M., Wong, K. L., Chan, K. W., Lim, W., Shortridge, K. F., Yuen, K. Y., Peiris, J. S., & Poon, L. L. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science (New York, N.Y.), 302(5643), 276–278. https://doi.org/10.1126/science.1087139
14. Hassanin A., Grandcolas P., Veron G. (2021) Covid-19: natural or anthropic origin? *Mammalia* 85:1–7. https://doi.org/10.1515/mammalia-2020-0044
15. Temmam, S., Vongphayloth, K., Baquero, E., Munier, S., Bonomi, M., Regnault, B., Douangboubpha, B., Karami, Y., Chrétien, D., Sanamxay, D., Xayaphet, V., Paphaphanh, P., Lacoste, V., Somlor, S., Lakeomany, K., Phommavanh, N., Pérot, P., Dehan, O., Amara, F., Donati, F., … Eloit, M. (2022). Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature, 604(7905), 330–336. https://doi.org/10.1038/s41586-022-04532-4
16. Hassanin A. (2022). Variation in synonymous nucleotide composition among genomes of sarbecoviruses and consequences for the origin of COVID-19. Gene, 835, 146641. https://doi.org/10.1016/j.gene.2022.146641
17. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010
18. Larsson A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531
19. Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
20. Swofford, D. L. (2003). PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland, MA: Sinauer Associates.

21. Ropiquet, A., Li, B., & Hassanin, A. (2009). SuperTRI: A new approach based on branch support analyses of multiple independent data sets for assessing reliability of phylogenetic inferences. *Comptes rendus biologies*, 332(9), 832–847. https://doi.org/10.1016/j.crvi.2009.05.001
22. Hassanin, A., Tu, V. T., Curaudeau, M., & Csorba, G. (2021). Inferring the ecological niche of bat viruses closely related to SARS-CoV-2 using phylogeographic analyses of *Rhinolophus* species. *Scientific reports*, 11(1), 14276. https://doi.org/10.1038/s41598-021-93738-z
23. Liu, P., Chen, W., & Chen, J. P. (2019). Viral metagenomics revealed sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses*, 11(11), 979. https://doi.org/10.3390/v11110979
24. Lam, T. T., Jia, N., Zhang, Y. W., Shum, M. H., Jiang, J. F., Zhu, H. C., Tong, Y. G., Shi, Y. X., Ni, X. B., Liao, Y. S., Li, W. J., Jiang, B. G., Wei, W., Yuan, T. T., Zheng, K., Cui, X. M., Li, J., Pei, G. Q., Qiang, X., Cheung, W. Y., … Cao, W. C. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, 583(7815), 282–285. https://doi.org/10.1038/s41586-020-2169-0
25. Oude Munnink, B. B., Sikkema, R. S., Nieuwenhuijse, D. F., Molenaar, R. J., Munger, E., Molenkamp, R., van der Spek, A., Tolsma, P., Rietveld, A., Brouwer, M., Bouwmeester-Vincken, N., Harders, F., Hakze-van der Honing, R., Wegdam-Blans, M. C. A., Bouwstra, R. J., GeurtsvanKessel, C., van der Eijk, A. A., Velkers, F. C., Smit, L. A. M., Stegeman, A., … Koopmans, M. P. G. (2021). Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science (New York, N.Y.), 371(6525), 172–177. https://doi.org/10.1126/science.abe5901
26. Hammer, A. S., Quaade, M. L., Rasmussen, T. B., Fonager, J., Rasmussen, M., Mundbjerg, K., Lohse, L., Strandbygaard, B., Jørgensen, C. S., Alfaro-Núñez, A., Rosenstierne, M. W., Boklund, A., Halasa, T., Fomsgaard, A., Belsham, G. J., & Bøtner, A. (2021). SARS-CoV-2 Transmission between Mink (*Neovison vison*) and Humans, Denmark. Emerging infectious diseases, 27(2), 547–551. https://doi.org/10.3201/eid2702.203794
27. Zhong, N. S., Zheng, B. J., Li, Y. M., Poon, Xie, Z. H., Chan, K. H., Li, P. H., Tan, S. Y., Chang, Q., Xie, J. P., Liu, X. Q., Xu, J., Li, D. X., Yuen, K. Y., Peiris, & Guan, Y. (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. Lancet (London, England), 362(9393), 1353–1358. https://doi.org/10.1016/s0140-6736(03)14630-2
28. Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., Wong, S. K., Huang, I. C., Xu, K., Vasilieva, N., Murakami, A., He, Y., Marasco, W. A., Guan, Y., Choe, H., & Farzan, M. (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. The EMBO journal, 24(8), 1634–1643. https://doi.org/10.1038/sj.emboj.7600640
29. Song, H. D., Tu, C. C., Zhang, G. W., Wang, S. Y., Zheng, K., Lei, L. C., Chen, Q. X., Gao, Y. W., Zhou, H. Q., Xiang, H., Zheng, H. J., Chern, S. W., Cheng, F., Pan, C. M., Xuan, H., Chen, S. J., Luo, H. M., Zhou, D. H., Liu, Y. F., He, J. F., … Zhao, G. P. (2005). Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proceedings of the National Academy of Sciences of the United States of America, 102(7), 2430–2435. https://doi.org/10.1073/pnas.0409608102
30. IUCN. 2022. The IUCN Red List of Threatened Species. Version 2022-1. https://www.iucnredlist.org. Accessed on 4 November 2022.
31. Han, Y., Du, J., Su, H., Zhang, J., Zhu, G., Zhang, S., Wu, Z., & Jin, Q. (2019). Identification of diverse bat alphacoronaviruses and betacoronaviruses in China provides new insights into the evolution and origin of coronavirus-related diseases. *Frontiers in microbiology*, 10, 1900. https://doi.org/10.3389/fmicb.2019.01900
32. Ikeda, Y., Jiang, T., Oh, H., Csorba, G., Motokawa, M. (2020) Geographic variations of skull morphology in the *Rhinolophus ferrumequinum* species complex (Mammalia: Chiroptera). *Zoologischer Anzeiger*, 288, 125e138. https://doi.org/10.1016/j.jcz.2020.08.004
33. Mao, X., Tsagkogeorga, G., Thong, V. D., & Rossiter, S. J. (2019). Resolving evolutionary relationships among six closely related taxa of the horseshoe bats (*Rhinolophus*) with targeted resequencing data. *Molecular phylogenetics and evolution*, 139, 106551. https://doi.org/10.1016/j.ympev.2019.106551