Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Gene 835 (2022) 146641

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Variation in synonymous nucleotide composition among genomes of sarbecoviruses and consequences for the origin of COVID-19

Alexandre Hassanin

Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, EPHE, MNHN, UA, Paris, France

ARTICLE INFO

Keywords:
Synonymous mutations
Recombination
RdRp selection
Manis javanica
Intermediate host
Reservoir host

ABSTRACT

The subgenus *Sarbecovirus* includes two human viruses, SARS-CoV and SARS-CoV-2, respectively responsible for the SARS epidemic and COVID-19 pandemic, as well as many bat viruses and two pangolin viruses.

Here, the synonymous nucleotide composition (SNC) of *Sarbecovirus* genomes was analysed by examining third codon-positions, dinucleotides, and degenerate codons. The results show evidence for the eight following groups: (i) SARS-CoV related coronaviruses (*SCoVrC* including many bat viruses from China), (ii) SARS-CoV-2 related coronaviruses (*SCoV2rC*; including five bat viruses from Cambodia, Thailand and Yunnan), (iii) pangolin sarbecoviruses, (iv) three bat sarbecoviruses showing evidence of recombination between *SCoVrC* and *SCoV2rC* genomes, (v) two highly divergent bat sarbecoviruses from Yunnan, (vi) the bat sarbecovirus from Japan, (vii) the bat sarbecovirus from Bulgaria, and (viii) the bat sarbecovirus from Kenya. All these groups can be diagnosed by specific nucleotide compositional features except the one concerned by recombination between *SCoVrC* and *SCoV2rC*. In particular, *SCoV2rC* genomes have less cytosines and more uracils at third codon-positions than other sarbecoviruses, whereas the genomes of pangolin sarbecoviruses show more adenines at third codon-positions. I suggest that taxonomic differences in the imbalanced nucleotide pools available in host cells during viral replication can explain the eight groups of SNC here detected among *Sarbecovirus* genomes. A related effect due to hibernating bats and their latitudinal distribution is also discussed. I conclude that the two independent host switches from *Rhinolophus* bats to pangolins resulted in convergent mutational constraints and that SARS-CoV-2 emerged directly from a horseshoe bat sarbecovirus.

1. Introduction

The Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), which is the causative agent of the coronavirus disease 2019 (COVID-19), was first detected in December 2019 in Wuhan (China) (Wu et al. 2020). By the end of March 2022, the virus had spread to 225 countries and territories, causing more than 480 millions confirmed infections and 6.1 millions deaths (https://www.worldometers.info/coronavirus/). The SARS-CoV-2 is an enveloped virus containing a positive single-stranded RNA genome of 29.9 kb (Wu et al. 2020). After entry into the host cell, two large overlapping open reading frames (ORFs), ORF1a and ORF1b, are translated into polypeptides that are cleaved into 16 non-structural proteins involved in the viral replication and transcription complex. Then, the viral replication is initiated by the synthesis of full-length negative-sense genomic copies, which serve as templates for the synthesis of genomic and subgenomic RNAs (Finkel et al. 2021; V'kovski et al 2021). The different subgenomic RNAs encode the four coronavirus structural proteins - spike (S), envelope (E), membrane (M) and nucleocapsid (N) - and a number of accessory proteins (six were predicted in the reference SARS-CoV-2 genome NC_045512: ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10; Wu et al. 2020).

According the International Committee on Taxonomy of Viruses (ICTV; https://talk.ictvonline.org/), the SARS-CoV-2 belongs to the family Coronaviridae, subfamily Orthocoronavirinae, genus *Betacoronavirus*, and subgenus *Sarbecovirus*. Phylogenetic analyses based on

Abbreviations: cds, protein coding sequences; COVID-19, coronavirus disease 2019; CTP, cytidine triphosphate; LBA, long branch attraction; nt, nucleotide; ORF, open reading frame; *PangSar*, pangolin sarbecoviruses; PCA, principal component analysis; PP, posterior probabilities; *RecSar*, three bat viruses showing evidence of genomic recombination between *SCoV2rC* and *SCoVrC* (RpPrC31, RsVZC45, and RsVZXC21); SARS-CoV-2, Severe Acute Respiratory Syndrome coronavirus 2; *SCoV2rC*, SARS-CoV-2 related coronaviruses; *SCoVrC*, SARS-CoV related coronaviruses; SNC, synonymous nucleotide composition; *YunSar*, group including two highly divergent bat viruses from Yunnan (RmYN05 and RstYN04).

E-mail address: alexandre.hassanin@mnhn.fr.

https://doi.org/10.1016/j.gene.2022.146641
Received 15 January 2022; Received in revised form 19 May 2022; Accepted 2 June 2022
Available online 11 June 2022

genomic sequences have shown that the subgenus *Sarbecovirus* includes another human virus, SARS-CoV, involved in the SARS epidemic between 2002 and 2004, two pangolin viruses, and a large diversity of bat viruses (Boni et al. 2020; Lam et al. 2020; Zhou et al. 2020, 2021; Delaune et al. 2021). Most of the bat sarbecoviruses were described from different species of the genus *Rhinolophus* (horseshoe bats) captured in caves of several provinces of China (Fan et al. 2019; Zhou H. et al. 2020, 2021). In addition, a few sarbecoviruses were detected in *Rhinolophus* species from Europe (Drexler et al. 2010), Africa (Tao and Tong 2019) and Southeast Asia (Delaune et al. 2021; Wacharapluesadee et al. 2021), suggesting that horseshoe bats of the Old World constitute the reservoir host in which sarbecoviruses have been circulating and evolving for centuries.

Five SARS-CoV-2 related coronaviruses (*SCoV2rC*), sharing between 92 and 96% of genomic identity with SARS-CoV-2, were recently sequenced from five horseshoe bat species sampled in Yunnan (*Rhinolophus affinis*, *Rhinolophus malayanus*, and *Rhinolophus pusillus*), Cambodia (*Rhinolophus shameli*) and Thailand (*Rhinolophus acuminatus*) (Zhou H. et al. 2020; Zhou P. et al. 2020; Delaune et al. 2021; Wacharapluesadee et al. 2021; Zhou et al. 2021). Based on these discoveries, the ecological niche of bat *SCoV2rC* was inferred using phylogeographic analyses of *Rhinolophus* species and it was found to include the four following geographic areas (Hassanin et al. 2021b): (i) southern Yunnan, northern Laos and bordering regions in northern Thailand and northwestern Vietnam; (ii) southern Laos, southwestern Vietnam, and northeastern Cambodia; (iii) the Cardamom Mountains in southwestern Cambodia and the East region of Thailand; and (iv) the Dawna Range in central Thailand and southeastern Myanmar. Importantly, the distribution of Sunda pangolin (*Manis javanica*) covers all these four geographic areas, as well as most other regions of mainland Southeast Asia, Borneo, Sumatra and Java (IUCN 2021). Since two sarbecoviruses related to *SCoV2rC* were sequenced from several Sunda pangolins seized in China between 2017 and 2019 (Lam et al. 2020; Xiao et al. 2020), it has been suggested that the species *Manis javanica* may have served as intermediate host between bat reservoirs and humans to import the ancestor of SARS-CoV-2 from Yunnan or Southeast Asia to the Chinese province of Hubei through wildlife trafficking (Lam et al. 2020). In accordance with the hypothesis, pangolins are known to be highly permissive to infection by sarbecoviruses (Xiao et al. 2020), as are also several small carnivores raised for fur or meat in China, such as the masked palm civet (*Paguma larvata*), raccoon dog (*Nyctereutes procyonoides*) (Guan et al. 2003) and American mink (*Neovison vison*) (Oude Munnink et al. 2021). In addition, it has been shown that pangolins collected in different geographic localities of Southeast Asia have been contaminated during their captivity (Hassanin et al. 2021a), reinforcing their possible role as intermediate host. However, all the closest relatives of SARS-CoV-2 currently known in wild animals were detected in horseshoe bats, not in pangolins or small carnivores, suggesting that the human index case was directly contaminated by a bat sarbecovirus.

The sarbecoviruses are obligate intracellular pathogens that cannot replicate without the machinery of a host cell. After entrance into the host cell, the replication of their positive-strand RNA genome is initiated by the synthesis of full-length negative-sense genomic copies, which function as templates for the generation of new RNA genomes (V'kovski et al. 2021). Since the replication process is dependent of the host cell, a viral host-shift to a new mammalian species (i.e., from the reservoir to a secondary host or from an intermediate host to a terminal host) may result in important changes in the mutational patterns driving the evolution of viral genomes. With time, such a mutational bias can affect the nucleotide content of viral genomes. Variation in nucleotide composition is generally studied at synonymous sites – where all types of mutations (at four-fold degenerate sites) or some of them (only transitions at two-fold degenerate sites) do not alter the sequences of amino acids encoded by the genes – because their evolution is assumed to be neutral or nearly so, i.e., weakly affected by natural selection (Kimura 1968; Ohta 1992). Several studies on the synonymous nucleotide composition (SNC) have detected high levels of variation among coronaviruses. Most of them have concluded that the codon usage is mainly driven by mutational bias towards A + U or U enrichment and selection against CpG dinucleotide (Kandeel et al., 2020; Tort et al., 2020; Daron & Bravo, 2021; Rice et al., 2021). However, these studies generally focused on SARS-CoV-2 features, and the variation among animal sarbecoviruses was generally not fully examined.

In the present study, the SNC was analysed in a selection of 54 *Sarbecovirus* genomes to provide new insight on the issue of the intermediate host. The three main objectives were (i) to evidence potential differences among bat *Sarbecovirus* lineages, (ii) to test whether the genomes of the two divergent pangolin sarbecoviruses have similar SNCs or not, and (iii) to determine if the genomic SNC of SARS-CoV-2 exhibits some unusual features or if it is similar to that of related sarbecoviruses found in horseshoe bats and Sunda pangolins.

2. Materials and methods

2.1. Genomic alignment of Sarbecovirus sequences

Full genomes of *Sarbecovirus* available in May 2021 in GenBank (https://www.ncbi.nlm.nih.gov/) and GISAID (https://www.epicov.org/) databases were downloaded in Fasta format. Sequences with large stretch of missing data were removed (e.g., a large fragment greater than 570 nt was missing in Rs672/2006; GenBank accession number: FJ588686). Only a single sequence was retained for similar genomes showing less than 0.1% of nucleotide divergence, such as those available for human SARS-CoV-2 (millions of sequences), pangolin sarbecoviruses from Guangxi (5 sequences), bat sarbecoviruses from Thailand (5 sequences), etc. All viral lineages previously described within the subgenus *Sarbecovirus* are included in this study (Drexler et al. 2010; Tao and Tong 2019; Murakami et al. 2020; Xiao et al. 2020; Zhou H. et al. 2020; Delaune et al. 2021; Wacharapluesadee et al. 2021; Zhou et al. 2021). The details on the 54 selected genomes are provided in Table 1. The protein coding sequences (cds) of the 54 genomes were aligned in Geneious Prime® 2020.0.3 with MAFFT version 7.450 (Katoh and Standley 2013) using default parameters. Then, the alignment was corrected manually on AliView 1.26 (Larsson 2014) based on translated and untranslated nucleotide sequences using the three following criteria: (i) the number of indels was minimized because they are rarer events than amino-acid or nucleotide substitutions; (ii) changes between similar amino-acids were preferred (using the ClustalX colour scheme available in AliView); and (iii) transitions were privileged over transversions because they are more frequent. The final alignment contains 29,550 nucleotides (nt), representing 9,850 codons. It is available in the Open Science Framework platform at https://osf.io/rv5u9/.

2.2. Phylogeny of sarbecoviruses

All phylogenetic analyses were carried out using MrBayes 3.2.7 (Ronquist et al. 2012) and different GTR + I + G models for the three codon-positions. The posterior probabilities (PP) were calculated using 10,000,000 Metropolis-coupled MCMC generations, tree sampling every 1000 generations and a burn-in of 25%.

Phylogenetic relationships were first inferred using the whole genomic alignment of 29,550 nt. Since several studies have shown evidence for multiple events of genomic recombination during the evolutionary history of sarbecoviruses (Hon et al. 2008; Boni et al. 2020), phylogenetic analyses were also conducted using the three following genomic regions of the alignment: 5′ region (positions 1–11,517; 3839 codons), central region (positions 12,970–20,289; 2440 codons), and 3′ region (positions 20,364–29,550; 3027 codons). These three genomic regions were defined after visual inspection of the translated alignment focusing on the three viruses showing evidence of recombination between *SCoVrC* and *SCoV2rC* (RpPrC31, RsVZC45, and RsVZXC21; Hu et al. 2018; Li et al. 2021; Hassanin et al., 2022).

Table 1
Origin of the *Sarbecovirus* genomes used in this study. (See above-mentioned references for further information.)

Virus name	Accession number	Host species	Geographic origin	Reference
SARS-CoV	NC_004718[1]	*Homo sapiens*	Canada	He et al. (2004)
As6526	KY417142[1]	*Aselliscus stoliczkanus*	Yunnan (China)	Hu et al. (2017)
RaLYRa11*	KF569996[1]	*Rhinolophus affinis*	Yunnan (China)	He et al. (2014)
RaYN2018A*	MK211375[1]	*Rhinolophus affinis*	Yunnan (China)	Han et al. (2019)
RaYN2018B*	MK211376[1]	*Rhinolophus affinis*	Yunnan (China)	Han et al. (2019)
RaYN2018C*	MK211377[1]	*Rhinolophus affinis*	Yunnan (China)	Han et al. (2019)
RaYN2018D*	MK211378[1]	*Rhinolophus affinis*	Yunnan (China)	Han et al. (2019)
Rf1	DQ412042[1]	*Rhinolophus ferrumequinum*[T1]	Hubei (China)	Li et al. (2005)
Rf4092	KY417145[1]	*Rhinolophus ferrumequinum*[T1]	Yunnan (China)	Hu et al. (2017)
RfJiyuan-84*	KY770860[1]	*Rhinolophus ferrumequinum*[T1]	Henan (China)	Lin et al. (2017)
RfV273*	DQ648856[1]	*Rhinolophus ferrumequinum*[T1]	Hubei (China)	Tang et al. (2006)
RfYNLF/31C*	KP886808[1]	*Rhinolophus ferrumequinum*[T1]	Yunnan (China)	Lau et al. (2015)
RmYN07*	EPI_ISL_1699447[2]	*Rhinolophus malayanus*	Yunnan (China)	Zhou et al. (2021)
Rma1*	DQ412043[1]	*Rhinolophus macrotis*[T2]	Hubei (China)	Li et al. (2005)
RmaV279*	DQ648857[1]	*Rhinolophus macrotis*[T2]	Yunnan (China)	Tang et al. (2006)
RmoLongquan140*	KF294457[1]	*Rhinolophus monoceros*[T3]	Zhejiang (China)	Lin et al. (2017)
RpF46*	KU973692[1]	*Rhinolophus pusillus*	Yunnan (China)	Wang et al. (2017)
RpShaanxi2011	JX993987[1]	*Rhinolophus pusillus*	Shaanxi (China)	Yang et al. (2013)
Rpe3	DQ071615[1]	*Rhinolophus pearsoni*	Guangxi (China)	Li et al. (2005)
Rs3367	KC881006[1]	*Rhinolophus sinicus*	Yunnan (China)	Ge et al. (2013)
Rs4081	KY417143[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs4084	KY417144[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs4231	KY417146[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs4237	KY417147[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs4247	KY417148[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs4255	KY417149[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs4874	KY417150[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs7327	KY417151[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
Rs9401	KY417152[1]	*Rhinolophus sinicus*	Yunnan (China)	Hu et al. (2017)
RsAnlong103*	KY770858[1]	*Rhinolophus sinicus*	Guizhou (China)	Lin et al. (2017)
RsHKU3-1*	DQ022305[1]	*Rhinolophus sinicus*	Hong-Kong (China)	Lau et al. (2005)
RsHKU3-7*	GQ153542[1]	*Rhinolophus sinicus*	Guangdong (China)	Lau et al. (2010)
RsHKU3-12*	GQ153547[1]	*Rhinolophus sinicus*	Hong-Kong (China)	Lau et al. (2010)
RsHuB2013*	KJ473814[1]	*Rhinolophus sinicus*	Hubei (China)	Wu et al. (2016)
RsSHC014	KC881005[1]	*Rhinolophus sinicus*	Yunnan (China)	Ge et al. (2013)
RstYN03*	EPI_ISL_1699443[2]	*Rhinolophus stheno*[T4]	Yunnan (China)	Zhou et al. (2021)
RstYN09*	EPI_ISL_1699449[2]	*Rhinolophus stheno*[T4]	Yunnan (China)	Zhou et al. (2021)
RspSC2018*	MK211374[1]	*Rhinolophus* sp.	Sichuan (China)	Han et al. (2019)
SARS-CoV-2	NC_045512[1]	*Homo sapiens*	Hubei (China)	Wu et al. (2020)
RaTG13	MN996532[1]	*Rhinolophus affinis*	Yunnan (China)	Zhou P. et al. (2020)
RacCS203	MW251308[1]	*Rhinolophus acuminatus*	Thailand	Wacharapluesadee et al. (2021)
RmYN02*	EPI_ISL_412977[2]	*Rhinolophus malayanus*	Yunnan (China)	Zhou H. et al. (2020)
RpYN06*	EPI_ISL_1699446[2]	*Rhinolophus pusillus*	Yunnan (China)	Zhou et al. (2021)
RshSTT200	EPI_ISL_852605[2]	*Rhinolophus shameli*	Cambodia	Delaune et al. (2021)
MjGuangdong*	EPI_ISL_410721[2]	*Manis javanica*	Guangdong (China)[G]	Xiao et al. (2020)
MjGuangxi*	EPI_ISL_410539[2]	*Manis javanica*	Guangxi (China)[G]	Lam et al. (2020)
RsVZXC21*	MG772934[1]	*Rhinolophus sinicus*	Zhejiang (China)	Hu et al. (2018)
RsVZC45*	MG772933[1]	*Rhinolophus sinicus*	Zhejiang (China)	Hu et al. (2018)
RpPrC31*	EPI_ISL_1098866[2]	*Rhinolophus pusillus*[T5]	Yunnan (China)	Li et al. (2021)
RmYN05	EPI_ISL_1699445[2]	*Rhinolophus malayanus*	Yunnan (China)	Zhou et al. (2021)
RstYN04*	EPI_ISL_1699444[2]	*Rhinolophus stheno*[T4]	Yunnan (China)	Zhou et al. (2021)
Rc-o319	LC556375[1]	*Rhinolophus cornutus*	Japan	Murakami et al. (2020)
RbBM48-31*	NC_014470[1]	*Rhinolophus blasii*	Bulgaria	Drexler et al. (2010)
RspKY72*	KY352407[1]	*Rhinolophus* sp.	Kenya	Tao and Tong (2019)

*original name slightly modified to be consistent with other names and to facilitate interpretations.
1: NCBI; 2: GISAID.
T: taxonomic issues; T1: currently *Rhinolophus nippon*; T2: currently *Rhinolophus episcopus*; T3 = the taxonomic assignation should be regarded as dubious because *Rhinolophus monoceros* is supposed to be endemic of Taiwan; T4: currently *Rhinolophus microglobosus* (Burgin et al. 2020); T5: currently *Rhinolophus pusillus blythi*

(Burgin et al. 2020), but included in *Rhinolophus blythi* in Li et al. (2021).
G: geographic issue; pangolins not collected in China, but more probably in Southeast Asia (exact locality unknown) (Hassanin et al. 2020a).

2.3. Analyses of nucleotide composition, dinucleotide composition and codon usage

The alignment of 54 *Sarbecovirus* genomes was used to calculate the frequency of the four nucleotides (A, C, G and U) at all third codon-positions. Nucleotide frequencies were calculated in PAUP (Swofford 2003) after exclusion of first and second codon-positions. The four variables measured were then summarised by a principal component analysis (PCA) using the FactoMineR package (Lê et al. 2008) in R version 3.6.1 (from https://www.R-project.org/). The nucleotide composition was also determined using three partitions of third codon-positions consisting in the four-fold degenerate sites (A, C, G and U percentages), purine two-fold degenerate sites (A *versus* G percentages), and pyrimidine two-fold degenerate sites (C *versus* U percentages). The eight variables were summarised by a PCA.

The genomic alignment was also used to calculate the frequency of the 16 possible dinucleotides (AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU) at second and third codon-positions (P23), and at third and first codon-positions (P31). For each of the two analyses, the 16 variables were summarised by a PCA.

Finally, the frequencies of synonymous codons (codon usage) were calculated for all amino acids except M and W, which are encoded by a single codon each (AUG and UGG, respectively). The 59 variables were summarised by a PCA.

3. Results

3.1. Nucleotide composition at third codon-positions

The four nucleotide frequencies at third codon-positions are provided in Table 2 for the main groups identified in this study (CSV file available at https://osf.io/rv5u9/). The four variables were summarised by a PCA based on the first two principal components (PCs), which contribute 89.08% and 9.84% of the total variance, respectively (Fig. 1**A**). The results allowed to distinguish the eight following groups: (i) *SCoVrC*; (ii) *SCoV2rC*; (iii) pangolin sarbecoviruses (here referred to as *PangSar*); (iv) a group here referred to as *RecSar*, which includes three bat sarbecoviruses showing evidence of genomic recombination between *SCoV2rC* and *SCoVrC* (RpPrC31, RsVZC45, and RsVZXC21; see below for more details); (v) a group here referred to as *YunSar*, including two highly divergent bat sarbecoviruses from Yunnan (RmYN05 and RstYN04); (vi) the bat sarbecovirus from Japan (Rc-o319); (vii) the bat sarbecovirus from Bulgaria (RbBM48-31); and (viii) the bat sarbecovirus from Kenya (RspKY72). As shown in Table 2, *SCoV2rC* genomes have more U nucleotides and less C nucleotides than other sarbecoviruses, whereas *SCoVrC* genomes exhibit the highest percentages of C nucleotide. The highest percentages of A nucleotide were found for the two *PangSar* genomes, whereas the lowest percentage of A nucleotide was found for the RbBM48-31 genome. The lowest percentages of G nucleotide were detected for *SCoV2rC* and *PangSar* genomes.

The nucleotide composition was also analysed at four-fold and two-fold degenerate third codon-positions (Table 2; CSV file available at https://osf.io/rv5u9/). The eight variables were summarised by a PCA based on the first two PCs, which contribute 73.07% and 20.08% of the total variance, respectively (Fig. 1**B**). The results confirmed the separation into eight groups, and some of them can be diagnosed by specific features, such as *YunSar* (less A nucleotides and more C nucleotides at four-fold degenerate third codon-positions), Rc-o319 (highest percentage of A nucleotide at four-fold degenerate third codon-positions), and RbBM48-31 (lowest percentage of A nucleotide at two-fold degenerate third codon-positions). The group uniting *SCoV2rC* and *PangSar* exhibits less G nucleotides and more U nucleotides at four-fold degenerate third codon-positions, as well as more A nucleotides at two-fold degenerate third codon-positions. For all variables at third codon-positions, *RecSar* genomes show intermediate values between *SCoV2rC* and *SCoVrC* genomes (Table 2).

3.2. Dinucleotide composition

The dinucleotide frequencies at second and third codon-positions (P23) and at third and first codon-positions (P31) are provided in Table 3 (CSV files available at https://osf.io/rv5u9/). For P23 and P31, the 16 variables were summarised by a PCA based on the first two PCs: for P23, they contribute 55.99% and 25.98% of the total variance, respectively (Fig. 2**A**); for P31, they contribute 64.13% and 17.83% of the total variance, respectively (Fig. 2**B**). The results showed a separation into the same eight groups previously identified, and some of them can be diagnosed by special features: *PangSar* genomes show less CG and GG at P23, and less CC and GU at P31; *SCoVrC* genomes are characterised by the highest percentages of GC and UC at P23; *SCoV2rC* genomes are characterised by the lowest percentages of GA at P31; *YunSar* genomes exhibit less AA, CA and GA at P23, less UC and UG at P31, more CC, CG, GU, and UA at P23, and more CG, GG, UA at P31; the Rc-o319 genome is the poorest in CU and UU at P23, whereas it is the richest in CA and GA at P23, and in AU at P31; the RbBM48-31 genome is the poorest in AA and AU at P31; the RspKY72 genome is the poorest in UC at P23 and in CU at P31, whereas it is the richest in AU and UG at P23, and in GU and UU at P31. The *SCoV2rC* and *PangSar* genomes are

Table 2
Relative frequencies of the four nucleotides at third codon-positions (3CP).

Third codon-positions	Bases	*SCoV2rC*	*PangSar*	*RecSar*	*SCoVrC*	*YunSar*	Rc-o319	RbBM48-31	RspKY72
All 3CP	A	27.8-28.3	**28.4-28.6**	26.5-27.1	24.4-25.6	24.3	27.4	**23.3**	25.0
	C	**15.6-16.1**	16.1-16.4	16.4-16.7	**18.5-20.5**	18.3	18.5	18.3	16.4
	G	12.4-13.0	12.6-12.7	13.8-14.5	15.5-16.4	15.9	15.4	16.1	15.6
	U	**43.1-43.6**	42.4-42.7	42.4-42.6	38.4-41.0	41.6	38.7	42.3	43.0
	A+U°	70.9-71.9	70.8-71.3	69.1-69.5	63.1-65.8	65.8-65.9	66.1	65.6	68.0
4-fold degenerate 3CP	A	28.9-29.3	29.9-30.0	28.5-29.3	27.5-29.8	**24.3**	**31.1**	25.1	26.3
	C	13.5-14.2	13.6-13.8	14.2-14.5	16.1-17.8	**19.2**	16.1	17.1	15.3
	G	6.4-6.9	6.2-6.7	7.2-8.0	8.1-10.2	9.6-9.7	8.6	9.1	9.2
	U	49.8-50.9	49.4-50.3	49.0-49.2	44.1-46.9	46.8-46.9	44.2	48.8	49.2
	A+U°	78.8-80.1	79.5-80.2	77.7-78.3	72.2-75.4	**71.1-71.2**	75.3	73.9	75.5
2-fold degenerate 3CP	A (*vs* G)*	66.7-68.5	67.5-67.6	61.7-63.5	55.3-58.5	57.7-57.8	58.4	**54.8**	59.6
	C (*vs* U)*	31.6-32.9	33.9-34.7	33.6-33.9	37.1-42.1	32.8	32.3	35.8	38.7

The special features of the group uniting *ScoV2rC* and *PangSar* are highlighted with light green background (percentages higher than other sarbecoviruses) or pale pink background (percentages lower than other sarbecoviruses). Similarly, coloured values indicate that one of the eight virus groups shows the highest nucleotide percentages (in green) or the lowest nucleotide percentages (in red).
°: variables not used in the PCAs of Fig. 1; *: the two variables were used in the PCA of Fig. 1.

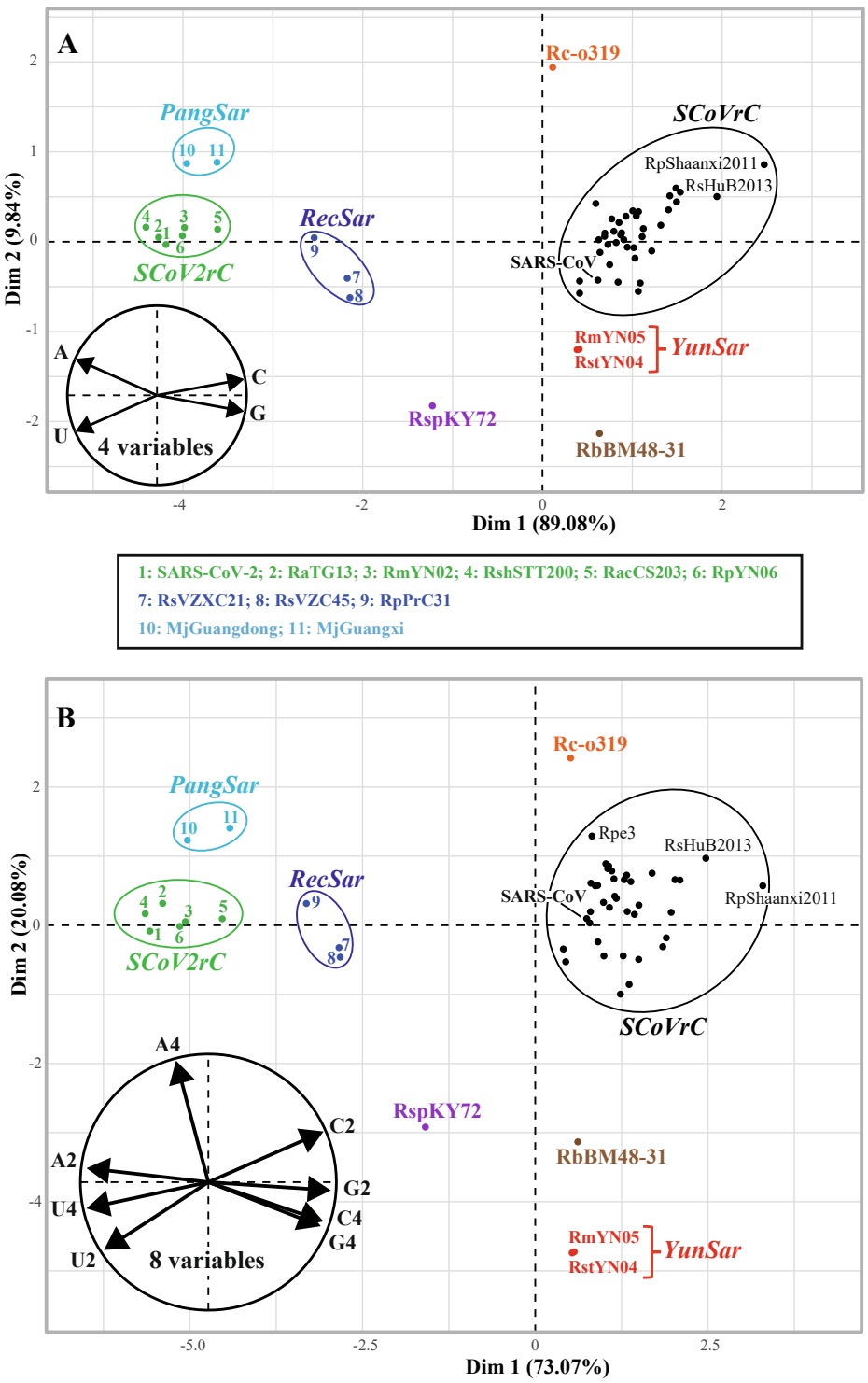

Fig. 1. Variation in nucleotide composition at third codon-positions of Sarbecovirus genomes. The genomic alignment of 29,550 nt was used to calculate the frequency of the four bases (A, C, G and U) at third codon-positions (Table 2) and the four variables measured were then summarised by a principal component analysis (PCA; Figure A). The main graph represents the individual factor map based on 54 *Sarbecovirus* genomes. The eight groups of nucleotide composition are highlighted by different colours: black for coronaviruses related to SARS-CoV (*SCoVrC*), green for coronaviruses related to SARS-CoV-2 (*SCoV2rC*), light blue for the two pangolin sarbecoviruses (*PangSar*), dark blue for bat sarbecoviruses showing evidence of genomic recombination between *SCoVrC* and *SCoV2rC* (*RecSar*), red for the two divergent bat sarbecoviruses recently identified in the Yunnan province by Zhou et al. (2021) (*YunSar*), orange for the bat sarbecoviruses from Japan, brown for the bat sarbecoviruses from Bulgaria, and purple for the bat sarbecoviruses from Kenya. The small circular graph at the bottom left represents the variables factor map. The frequency of the four bases was also calculated either at four-fold degenerate third codon-positions or at two-fold degenerate third codon-positions for either purines (A *versus* G) or pyrimidines (C *versus* U) (Table 2). The PCA obtained using these eight variables is shown in Figure B. The variables factor map is shown at the bottom left.

characterised by the highest percentages of AA at P23 and AC at P31, and by the lowest percentages of AG and CG at P23, and CC, CG, GA, GC, and GG at P31. The group uniting *SCoV2rC*, *PangSar* and *RecSar* show more AA at P31, less GC at P23, and less CC, CG, GC, and GG at P31. For all dinucleotides, *RecSar* genomes exhibit intermediate frequencies between *SCoV2rC* and *SCoVrC* genomes (Table 3).

3.3. Codon usage

The 59 variables corresponding to the relative frequencies between synonymous codons of 18 different amino acids (all except M and W) were summarised in Table 4 (CSV file available at https://osf.io/rv5u9/). The first two dimensions of the PCA contribute 44.93% and 19.37% of the total variance, respectively (Fig. 3). Here again the results confirmed the division into the eight groups previously identified, and some of them can be diagnosed by special features in codon usage: *PangSar* genomes have the lowest percentages for GCC Alanine codon; *SCoVrC* genomes have the lowest percentages for AUA Isoleucine codon, UUA Leucine codon and GUU Glycine codon, and the highest percentages for AUC Isoleucine codon, CUC and CUG Leucine codons;

Table 3
Relative frequencies of dinucleotides at second and third codon-positions (P23) and at third and first codon-positions (P31).

Dinucleotides	*SCoV2rC*	*PangSar*	*RecSar*	*SCoVrC*	*YunSar*	Rc-o319	RbBM48-31	RspKY72
P23 AA	9.3-9.7	9.5-9.6	8.6-9.0	7.6-8.2	**6.8**	8.4	7.2	7.8
P23 AC	6.0-6.3	6.3-6.5	6.2-6.4	6.4-7.4	6.2	6.9	6.8	6.1
P23 AG	4.3-4.5	4.4-4.5	5.0-5.3	5.9-6.4	6.3	5.7	6.1	5.6
P23 AU	11.0-11.3	10.6-11.0	10.9-11.1	9.6-10.8	11.3	10.3	10.6	**11.5**
P23 CA	8.2-8.4	8.7	8.2-8.4	8.0-8.7	**6.5**	**9.0**	7.4	7.7
P23 CC	2.4-2.7	2.4-2.6	2.5-2.7	2.7-3.1	**4.0**	2.9	3.4	2.8
P23 CG	0.9-1.0	**0.9**	1.1-1.3	1.1-1.6	**1.9**	1.2	1.3	1.2
P23 CU	11.0-11.4	10.8-11.0	10.8-10.9	10.0-10.9	10.5	**10.0**	10.8	10.7
P23 GA	2.9-3.1	3.2	3.1-3.1	2.8-3.1	**2.5**	**3.3**	2.8	2.6
P23 GC	2.5-2.7	2.6-2.8	2.7-3.0	**3.5-4.2**	3.0	3.4	3.4	3.1
P23 GG	1.9-2.0	**1.7-1.8**	1.9-1.9	1.9-2.1	2.1	1.9	2.0	2.1
P23 GU	8.1-8.3	8.0-8.2	7.8-8.0	6.7-7.6	**8.4**	7.2	7.9	8.1
P23 UA	7.0-7.3	6.8-7.2	6.5-6.6	5.5-5.9	**8.5**	6.7	5.9	6.9
P23 UC	4.5-4.7	4.7-4.8	4.9-4.9	**5.6-6.0**	5.0	5.4	4.8	**4.4**
P23 UG	5.3-5.6	5.4-5.7	5.8-6.0	6.3-6.6	5.6	6.6	6.7	**6.7**
P23 UU	12.7-13.1	12.6-12.8	12.8-12.8	11.8-12.3	11.4	**11.2**	13.0	12.6
P31 AA	7.5-7.9	7.7	7.3-7.5	6.2-6.8	6.1	7.2	**5.8**	6.2
P31 AC	6.8-7.0	6.9-7.1	6.4-6.5	5.9-6.4	6.0	6.4	5.9	6.0
P31 AG	8.7-8.9	8.8-9.3	8.3-8.6	7.5-8.1	7.5	8.9	7.6	8.5
P31 AU	4.5-4.8	4.7-4.8	4.4-4.5	4.1-4.6	4.6	**4.8**	**4.0**	4.2
P31 CA	6.8-7.0	7.4	7.1-7.3	7.8-8.7	6.8	8.2	7.2	7.1
P31 CC	2.2-2.4	**2.1-2.2**	2.4-2.5	2.7-3.3	3.0	2.7	2.9	2.7
P31 CG	1.8-2.0	1.7-2.1	2.1	2.5-3.0	**3.0**	2.4	2.9	2.2
P31 CU	4.6-4.9	4.9	4.8-4.9	5.1-5.8	5.5	5.3	5.3	**4.4**
P31 GA	**3.3-3.4**	3.4-3.5	3.6-3.8	3.9-4.4	3.5	4.1	4.0	3.8
P31 GC	2.4-2.6	2.4-2.5	2.7-2.9	3.4-3.8	3.4	3.3	3.6	3.3
P31 GG	3.1-3.3	3.0-3.3	3.4-3.7	3.9-4.3	**4.7**	4.0	4.1	3.8
P31 GU	3.7-3.9	**3.5-3.6**	4.0	3.8-4.2	4.2	3.9	4.4	**4.6**
P31 UA	12.1-12.5	11.6	11.7-11.9	9.8-10.8	**13.4**	10.4	12.1	12.3
P31 UC	4.7-5.0	4.9-5.1	4.9-5.0	5.0-5.4	**4.5**	4.5	4.6	4.6
P31 UG	16.3-16.6	16.3-16.5	16.1-16.2	16.0-16.7	**14.9**	15.4	16.3	16.0
P31 UU	9.4-10.1	9.5-9.7	9.6-9.7	7.5-8.4	8.8	8.4	9.3	**10.2**

The special features of the group uniting *ScoV2rC* and *PangSar* are highlighted with light green background (percentages higher than other sarbecoviruses) or pale pink background (percentages lower than other sarbecoviruses). Similarly, coloured values indicate that one of the eight virus groups shows the highest nucleotide percentages (in green) or the lowest nucleotide percentages (in red).

YunSar genomes have an atypical codon composition, as they show the lowest percentages for codons AAA, ACA, AUU, CUU, GCA, GGA, UAC, UCA, and UUG and the highest percentages for codons ACC, ACG, AUA, CCC, GCC, UCC, UCG, and UUA; the Rc-o319 genome is the poorest in ACU Threonine codons and GUU Valine codons, whereas it is the richest in ACA Threonine codons, GCA Alanine codons, and GUA Valine codons; the RbBM48-31 genome exhibits the lowest percentages for CAA Glutamine codon, CCA Proline codon, and GUA Valine codon; and the RspKY72 genome shows the lowest percentages for AUC Isoleucine codon, CAC Histidine codon, CGA Arginine codon, and CUC Leucine codon, and the highest percentages for CCU Proline codon, AGG and CGG Arginine codons, and UUG Leucine codon. The *SCoV2rC* and *PangSar* genomes are characterised by the lowest percentages for CCC Proline codon and GGC Glycine codon, and by the highest percentages for AAA Lysine codon, CAA Glutamine codon, and GAA Glutamate codon. The group uniting *SCoV2rC*, *PangSar* and *RecSar* is characterised by the lowest percantages for GUG Valine codons. For all variables, *RecSar* genomes show intermediate values between *SCoV2rC* and *SCoVrC* genomes (Table 4).

3.4. Phylogenetic relationships within the subgenus Sarbecovirus

The Bayesian tree derived from the analysis of the whole genome alignment of protein-coding sequences (29,550 nt) is shown in Fig. 4**A**. Two SNC groups were found to be monophyletic: (i) *SCoVrC*; and (iii) *YunSar*, which is composed of two divergent bat sarbecoviruses from Yunnan, i.e. RmYN05 and RstYN04. By contrast, *SCoV2rC* was found to be paraphyletic due to the inclusive placement of the two pangolin sarbecoviruses and *YunSar*. However, the branch leading to *YunSar* was found to be much more longer than other branches, suggesting a possible long branch attraction (LBA) artefact.

The genome alignment was then visualized in more details for amino-acid replacements. The three *RecSar* sequences (RpPrC31, RsVZC45, and RsVZXC21) showed high amino acid similarities with *SCoV2rC* sequences in the 5′ genomic region (positions 1–11517) and 3′ genomic region (positions 20470–29550), whereas they were found more similar to *SCoVrC* sequences in the central genomic region (positions 12970–20289). For that reason, phylogenetic analyses were also performed on these three genomic regions separately.

The Bayesian tree constructed from the 5′ genomic region is shown in Fig. 4**B**. Deep relationships were found to be congruent with the tree derived from the whole genome alignment, such as the monophyly of *SCoVrC* (PP = 1), the clade uniting *SCoV2rC*, *RecSar*, *YunSar*, and the two pangolin sarbecoviruses (PP = 1), and its sister-group relationship with Rc-o319 (PP = 1). However, several more recent relationships were discordant. For instance, the two pangolin sarbecoviruses and *YunSar* appeared more distantly related to *SCoV2rC* and *RecSar*, whereas RpPrC31 was found within the *SCoV2rC* group (PP = 1) as the sister-group of RmYN02 (PP = 1).

The Bayesian tree built from the central genomic region is shown in Fig. 4**C**. The topology was highly incongruent with other trees because the three *RecSar* viruses appeared included into the *SCoVrC* group: RsVZC45 and RsVZXC21 were closely related to RmLongquan140 (PP = 1); whereas RpPrC31 was enclosed into a large group including SARS-CoV and several bat sarbecoviruses (PP = 1). In addition, *SCoV2rC* was found to be monophyletic (PP = 1) and *YunSar* was closely related to MjGuangdong (PP = 1).

The Bayesian tree derived from the 3′ genomic region is shown in Fig. 4**D**. The results supported a large clade uniting *SCoVrC*, *SCoV2rC*, *RecSar*, Rc-o319 and the two pangolin sarbecoviruses (PP = 1) due to the

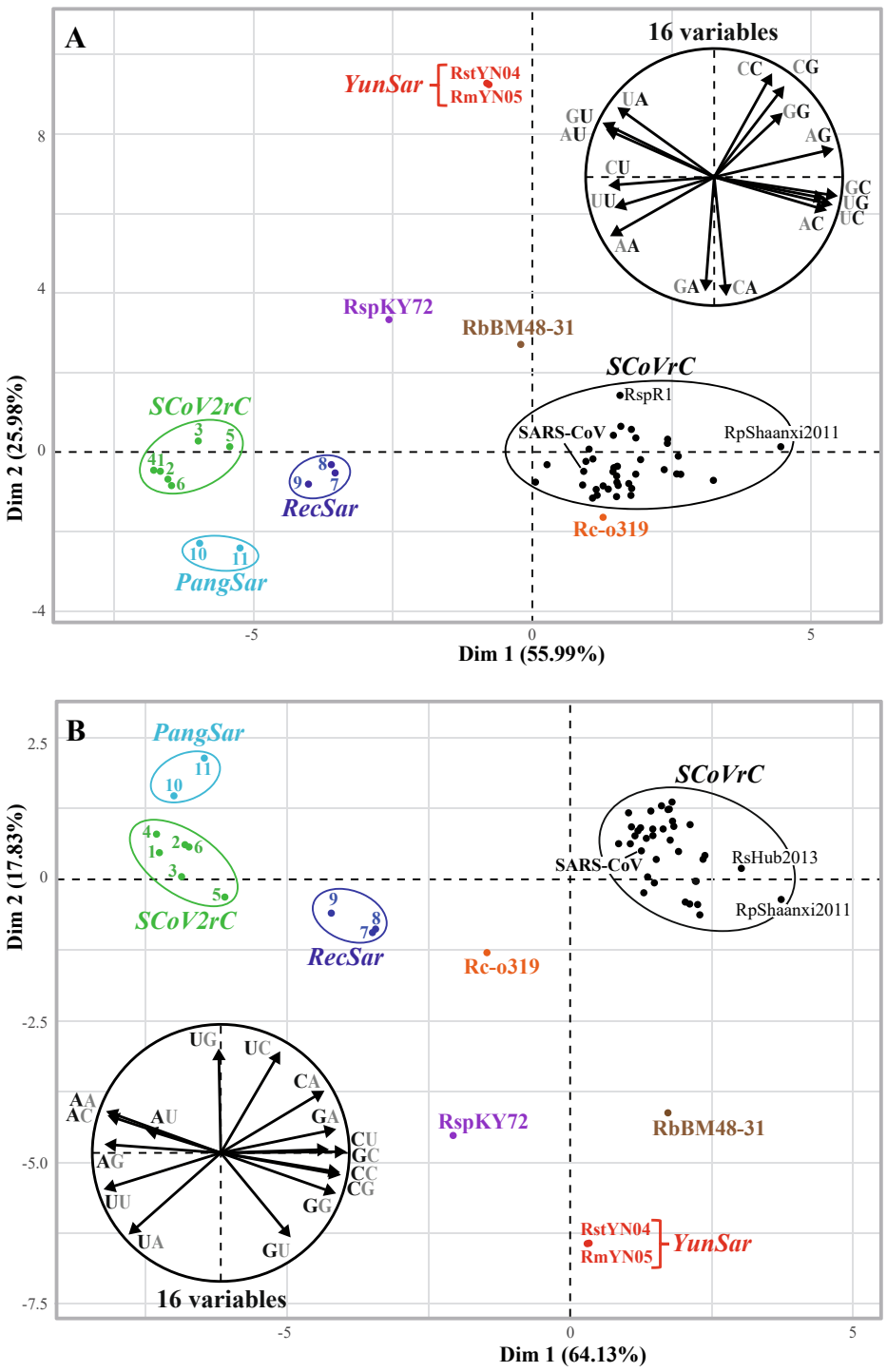

Fig. 2. Variation in dinucleotide composition among *Sarbecovirus* genomes. The genomic alignment of 29,550 nt was used to calculate the frequencies of the 16 possible dinucleotides at second and third codon positions (P23) (Table 3) and the variables were summarised by a principal component analysis (PCA; Figure A). The main graph represents the individual factor map based on 54 *Sarbecovirus* genomes. The eight groups of dinucleotide composition are highlighted by different colours as defined in Fig. 1A. The small circular graph at the top right represents the variables factor map. The dinucleotide frequencies at third and first codon positions (P31) were also calculated (Table 3) and the 16 variables were summarised by a PCA (Figure B). The variables factor map is shown at the bottom left.

divergent placement of *YunSar*. Three SNC groups were found monophyletic (PP = 1): *RecSar*, *SCoVrC*, and *YunSar*. By contrast, *SCoV2rC* was found to be polyphyletic, as RacCS203, RmYN02, and RpYN06 appeared closely related to *RecSar* (PP = 1), whereas SARS-CoV-2, RaTG13, and RshSTT200 were grouped with the two pangolin sarbecoviruses (PP = 0.8).

4. Discussion

4.1. RdRp selection of recombinant sarbecoviruses

The separate phylogenetic analyses based on 5′, central and 3′ genomic regions showed that the three *RecSar* viruses (RpPrC31, RsZC45 and RsZXC21) have emerged through two independent events of recombination involving *SCoVrC* and *SCoV2rC* genomes (Hu et al. 2018; Boni et al. 2020; Li et al. 2021): one resulted in the ancestor of RpPrC31, and the other led to the ancestor of RsZC45 and RsZXC21. On the one hand, RpPrC31 was included in the *SCoV2rC* group in the trees based on 5′ and 3′ genomic regions (Fig. 4**B and 4D**, respectively), whereas it appeared in the *SCoVrC* subgroup uniting all viruses from Southwest China in the tree based on the central genomic region (Fig. 4**C**). On the other hand, RsZC45 and RsZXC21 appeared as sister-groups in all phylogenetic trees of Fig. 4, indicating that these two viruses have shared the same evolutionary history until their recent divergence. In

Table 4
Relative codon frequencies for all amino acids (except M and W that are encoded by a single codon).

Amino acid	Codons	*SCoV2rC*	*PangSar*	*RecSar*	*SCoVrC*	*YunSar*	Rc-o319	RbBM48-31	RspKY72
A	**GCA**	25.6-29.1	30.2-32.4	25.6-27.7	25.0-29.4	**24.0-24.1**	**32.5**	25.3	28.2
	GCC	13.6-15.8	**12.2-12.6**	13.4-15.6	12.7-16.8	**19.2-19.3**	13.7	16.0	15.0
	GCG	2.6-4.2	2.6-2.9	4.8-6.0	4.5-8.1	7.5	4.2	7.2	5.6
	GCU	52.0-56.9	52.4-54.7	53.0-53.9	48.6-55.0	49.1-49.2	49.6	51.5	51.2
C	**UGC** (*vs* UGU)*	22.4-25.9	26.9-27.6	26.7-30.8	34.4-41.4	25.5-25.6	38.1	33.2	25.7
D	**GAC** (*vs* GAU)*	34.6-37.0	36.8-40.5	35.9-38.0	36.9-46.0	37.7	39.9	38.7	38.2
E	**GAA** (*vs* GAG)*	67.5-72.1	68.9-70.8	62.6-65.4	52.4-59.1	53.3	60.7	58.3	60.2
F	**UUC** (*vs* UUU)*	29.6-32.5	32.1-34.1	32.5-33.9	35.9-42.9	32.1	38.4	30.2	29.8
G	**GGA**	20.6-22.8	21.8-24.1	22.9-23.5	22.2-25.2	**15.7**	23.8	20.0	18.3
	GGC	16.5-18.4	14.3-18.4	19.2-19.5	21.8-26.5	22.4-22.6	20.4	22.4	21.2
	GGG	2.6-3.7	2.8-3.1	2.6-4.0	2.9-6.0	4.8	3.8	3.4	3.3
	GGU	57.4-58.5	57.0-58.6	53.8-54.4	**45.6-50.8**	56.9-57.1	52.1	54.1	57.2
H	**CAC** (*vs* CAU)*	28.6-32.3	29.5-33.3	29.4-31.6	29.4-38.2	32.6	36.0	34.8	**27.4**
I	**AUA**	29.6-31.5	30.4-34.9	28.2-28.6	**21.2-23.9**	**41.6-41.7**	32.3	28.1	35.5
	AUC	15.9-18.5	16.2-16.8	18.4-19.6	**19.7-24.3**	14.5	18.8	15.6	**14.3**
	AUU	50.9-53.5	48.3-53.4	52.2-53.3	53.4-58.4	**43.8-43.9**	49.0	56.2	50.2
K	**AAA** (*vs* AAG)*	65.3-69.5	63.7-66.7	59.8-63.2	52.9-57.2	**49.2-49.4**	57.6	52.3	57.4
L	**CUA**	10.5-12.3	10.3-12.3	10.1-10.4	10.8-14.6	13.9	12.5	10.5	10.8
	CUC	9.4-10.7	10.0-10.1	10.6-11.0	**12.3-15.7**	11.7-11.8	11.5	9.7	**8.2**
	CUG	4.3-6.3	5.5-6.6	6.1-6.8	**8.4-10.5**	5.9	8.3	7.2	7.6
	CUU	27.6-29.9	29.1-31.2	29.3-29.9	28.0-31.5	**22.5-22.6**	27.4	30.3	27.1
	UUA	26.0-27.4	23.8-27.2	22.9-24.8	**15.8-19.2**	**32.3**	20.2	21.5	25.3
	UUG	16.8-17.9	16.8-17.0	18.3-19.8	15.7-18.1	**13.6-13.7**	20.2	20.8	**21.1**
N	**AAC** (*vs* AAU)*	30.5-32.7	34.3-35.2	32.1-34.9	34.6-41.8	33.0	36.7	36.9	30.8
P	**CCA**	37.6-40.3	37.6-39.9	39.6-40.8	38.1-44.4	32.1	40.9	**31.2**	32.6
	CCC	6.3-8.4	7.2-7.6	8.4-10.3	9.3-13.0	**15.1**	9.2	10.7	8.8
	CCG	3.5-5.5	5.1-5.4	4.9-6.4	3.3-8.0	7.1	7.2	5.8	6.1
	CCU	48.5-50.9	47.5-49.9	43.9-45.5	39.5-45.8	45.7	42.7	52.4	**52.5**
Q	**CAA** (*vs* CAG)*	67.0-70.3	66.9-72.8	62.3-64.3	55.7-62.1	52.8	60.6	**50.9**	57.2
R	**AGA**	41.4-44.7	45.3-46.0	42.6-44.0	32.9-39.4	36.2-36.3	45.8	35.3	37.5
	AGG	13.0-15.3	11.6-12.3	13.1-14.0	11.1-16.3	15.5-15.6	12.6	14.9	**17.6**
	CGA	4.3-6.0	4.8-8.4	4.8-5.7	4.4-7.0	4.9	5.2	6.1	**4.3**
	CGC	9.7-11.7	10.5-10.7	10.0-11.6	12.2-18.7	11.2	12.3	12.4	12.2
	CGG	2.3-3.1	0.6-1.1	1.4	1.1-3.0	3.0	1.1	2.5	**3.4**
	CGU	23.4-25.4	22.8-25.9	24.9-26.3	24.2-30.2	29.0-29.2	22.9	28.9	25.0
S	**AGC**	6.1-7.5	7.0-7.5	5.9-7.6	7.2-10.6	7.0-7.1	8.5	8.2	8.9
	AGU	22.2-24.1	22.4-22.8	22.0-23.2	18.6-22.1	24.0	20.2	21.0	21.3
	UCA	26.9-27.8	26.8-27.5	26.7-29.4	25.2-29.9	**20.2-20.3**	27.1	24.8	25.5
	UCC	6.7-8.1	6.8-8.1	7.3-7.7	6.5-8.9	**12.7-12.8**	9.7	10.7	9.6
	UCG	1.8-2.5	2.3-2.4	2.0-2.5	3.2-4.8	**5.6-5.7**	3.6	2.7	1.8
	UCU	31.0-33.7	32.1-34.2	31.3-33.9	29.9-34.0	30.3	30.9	32.5	32.9
T	**ACA**	40.1-41.5	40.7-44.1	39.4-40.8	37.5-41.0	**29.7-29.8**	**44.8**	38.0	39.0
	ACC	9.0-10.6	11.2-12.1	9.5-10.9	11.1-14.5	**16.7-16.8**	12.7	15.4	11.2
	ACG	5.0-5.6	4.2-4.6	5.9-6.1	4.0-8.1	**9.6**	5.7	6.1	7.2
	ACU	43.0-45.8	40.2-43.0	43.7-44.0	40.1-44.8	43.9	**36.8**	40.6	42.6
V	**GUA**	22.0-24.1	22.2-23.3	21.9-23.1	19.0-23.2	21.2	**24.3**	**16.7**	18.2
	GUC	13.5-15.6	14.0-16.2	13.6-15.0	16.0-19.9	18.8	18.2	18.1	16.2
	GUG	13.8-15.0	12.6-16.8	16.1-17.1	17.9-21.1	17.4	18.7	20.5	20.7
	GUU	46.0-49.2	44.8-50.1	46.0-47.3	39.9-45.2	42.6	**38.9**	44.8	44.9
Y	**UAC** (*vs* UAU)*	38.8-42.0	39.5-41.7	40.4-42.5	40.0-51.7	**36.9**	45.3	43.2	37.8

The special features of the group uniting *ScoV2rC* and *PangSar* are highlighted with light green background (percentages higher than other sarbecoviruses) or pale pink background (percentages lower than other sarbecoviruses). Similarly, coloured values indicate that one of the eight virus groups shows the highest codon percentages (in green) or the lowest codon percentages (in red). *: the two variables were used in the PCA of Fig. 3.

addition, RsZC45 and RsZXC21 were closely related to *SCoV2rC* viruses in the trees based on 5′ and 3′ genomic regions (Fig. 4**B and 4D**, respectively), whereas they appeared in the *SCoVrC* subgroup uniting all viruses from East China in the tree based on the central genomic region (Fig. 4**C**). All these results indicate that two similar recombinant patterns corresponding to 5′-*SCoV2rC*-*SCoVrC*-*SCoV2rC*-3′ were independently selected in the ancestor of RpPrC31, and that of RsZC45 and RsZXC21. For RpPrC31, the recombination event was likely to occur in Yunnan, where the ecological niches of *SCoVrC* and *SCoV2rC* slightly overlap (Hassanin et al. 2021b). For the common ancestor of RsZC45 and RsZXC21, the recombination event occurred in East China, and more probably in the Zhejiang province, where the two viruses RsZC45 and RsZXC21 were discovered (Hu et al., 2018), as well as their sister virus, RmoLongquan140, in the phylogeny based on the central genomic region (Fig. 4**C**; Lin et al., 2017). It can be therefore proposed that the parental *SCoV2rC* strain of RsZC45 and RsZXC21 may have dispersed from Yunnan to East China through several generations of bats via occasional contacts in caves between populations usually found in different regions. Such a scenario involving a diffusion over several decades of new *Sarbecovirus* variants from Yunnan to other provinces of China should be tested with additional data.

The occurrence of the same 5′-*SCoV2rC*-*SCoVrC*-*SCoV2rC*-3′ pattern in two different provinces, Yunnan and Zhejiang, and into two different host species, *Rhinolophus pusillus* and *Rhinolophus sinicus* is intriguing as it suggests that the pattern was positively selected. Importantly, the central genomic region contains the cds of the RNA-dependent RNA polymerase, which plays a key role in the replication and transcription of the viral RNA genome (Gao et al., 2020). The fact that the central

Table 5
Relative frequencies of the four nucleotides at third codon-positions (3CP) of the three genomic regions.

Third codon-positions	Bases	*SCoV2rC*	*PangSar*	*RecSar*	*SCoVrC*	*YunSar*	Rc-o319	RbBM48-31	RspKY72
5' region	A	27.2-27.9	28.2-28.7	27.3-27.6	23.0-26.7	**21.6-21.7**	**29.8**	23.4	25.6
4-fold degenerate 3CP	C	11.8-12.7	12.3-13.3	11.5-12.4	15.3-18.0	**19.7-19.9**	15.4	16.4	14.4
	G	5.6-6.4	5.7-6.6	5.8-6.1	7.9-10.0	9.6	8.1	7.9	9.2
	U	53.6-54.8	51.3-53.9	54.2-55.2	47.9-51.1	48.9-49.0	**46.8**	52.3	50.7
5' region	A (*vs* G)*	65.1-66.1	64.4-65.2	62.7-65.6	51.6-55.2	56.4	58.0	52.8	55.6
2-fold degenerate 3CP	C (*vs* U)*	29.2-32.1	32.1-33.8	28.8-30.1	35.4-42.8	32.9	37.0	34.4	32.6
central region	A	30.9-31.6	32.1-32.4	31.1-32.1	30.0-33.3	28.5	**34.5**	27.3	**27.0**
4-fold degenerate 3CP	C	13.2-14.4	12.4-13.7	15.1-15.3	15.2-17.9	17.6	16.4	17.3	15.2
	G	6.0-6.5	5.9-6.1	8.6-11.4	7.4-11.1	9.7	7.2	10.8	8.4
	U	48.5-48.8	48.3-49.1	42.4-44.1	40.1-45.6	44.3	41.9	44.6	**49.4**
central region	A (*vs* G)*	71.2-74.3	70.4-71.2	57.8-59.0	54.3-58.0	63.0	57.5	57.5	58.0
2-fold degenerate 3CP	C (*vs* U)*	30.1-32.1	32.1-32.7	35.0-35.8	33.7-41.9	30.3	37.1	34.2	**29.9**
3' region	A	28.1-29.7	30.0-30.3	27.5-28.3	27.7-31.5	**24.2**	29.5	25.8	26.3
4-fold degenerate 3CP	C	15.0-16.7	15.4-15.5	16.7-17.9	16.1-20.1	**20.7**	17.9	17.7	16.5
	G	7.1-8.8	7.0-7.3	7.7-8.1	8.3-10.5	10.0	10.2	8.9	9.6
	U	45.7-48.5	47.0-47.4	45.7-48.1	39.8-45.2	45.1	42.4	47.5	47.6
3' region	A (*vs* G)*	64.6-69.2	69.1-69.4	63.3-64.3	59.0-64.7	**55.1-55.4**	64.2	55.9	63.3
2-fold degenerate 3CP	C (*vs* U)*	34.8-37.1	38.9-39.1	36.7-36.8	38.7-45.7	35.7	42.0	39.8	35.0

The special features of the group uniting *ScoV2rC* and *PangSar* are highlighted with light green background (percentages higher than other sarbecoviruses) or pale pink background (percentages lower than other sarbecoviruses). Similarly, coloured values indicate that one of the eight virus groups shows the highest nucleotide percentages (in green) or the lowest nucleotide percentages (in red).
*: the two variables were used in the PCA of Fig. 5.

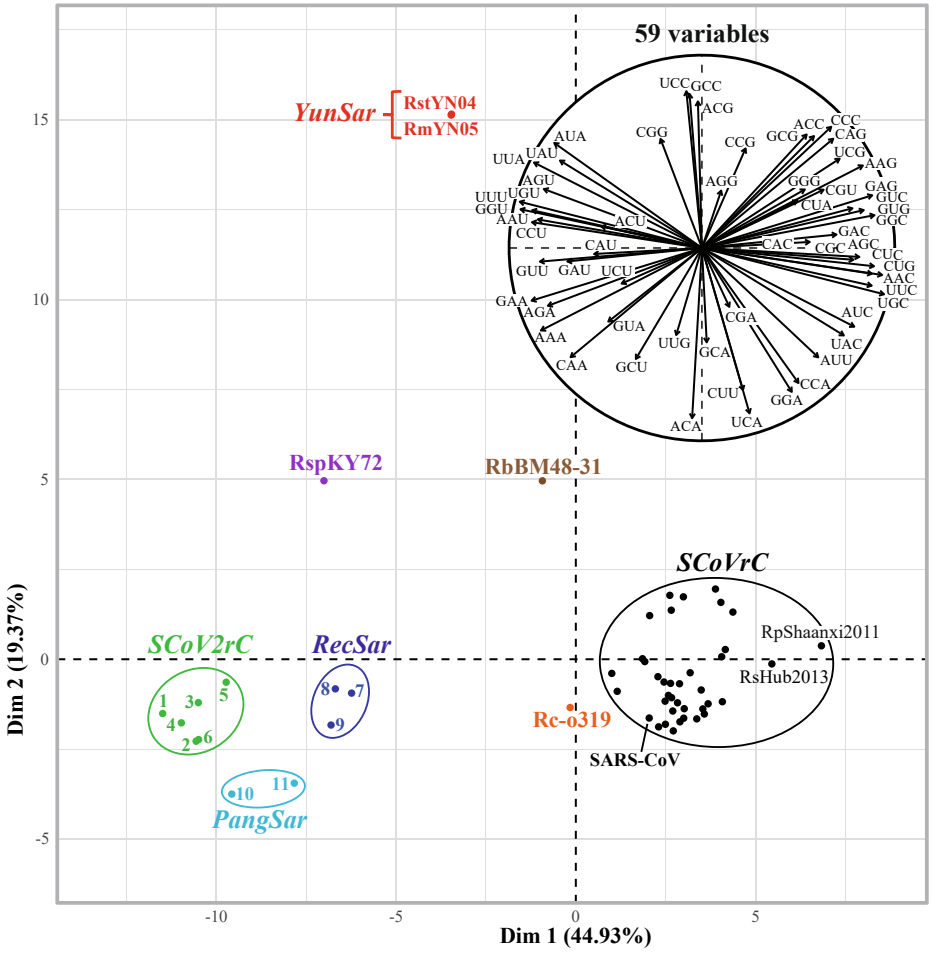

Fig. 3. Variation in codon usage among *Sarbecovirus* genomes. The alignment of 29,550 nt was used to calculate the frequencies of synonymous codons for all amino acids except M and W that are encoded by a single codon (Table 4). The 59 variables were then summarised by a principal component analysis (PCA). The main graph represents the individual factor map based on 54 *Sarbecovirus* genomes. The eight groups of codon usage are highlighted by different colours as defined in Fig. 1A. The circular graph at the top right represents the variables factor map.

genomic region was under strong selective pressure in bat host cells is another argument in favour of the hypothesis involving positive selection for a *SCoVrC* central region. It is important to note here that the tree based on the central genomic region showed a strong geographic structure: in the clade uniting *SCoVrC* and *RecSar* viruses, there are three geographic groups representing Southwest China, Central China and East China; within *SCoV2rC*, there are two geographic groups corresponding to Yunnan and South East Asia (Fig. 4**C**). A similar geographic

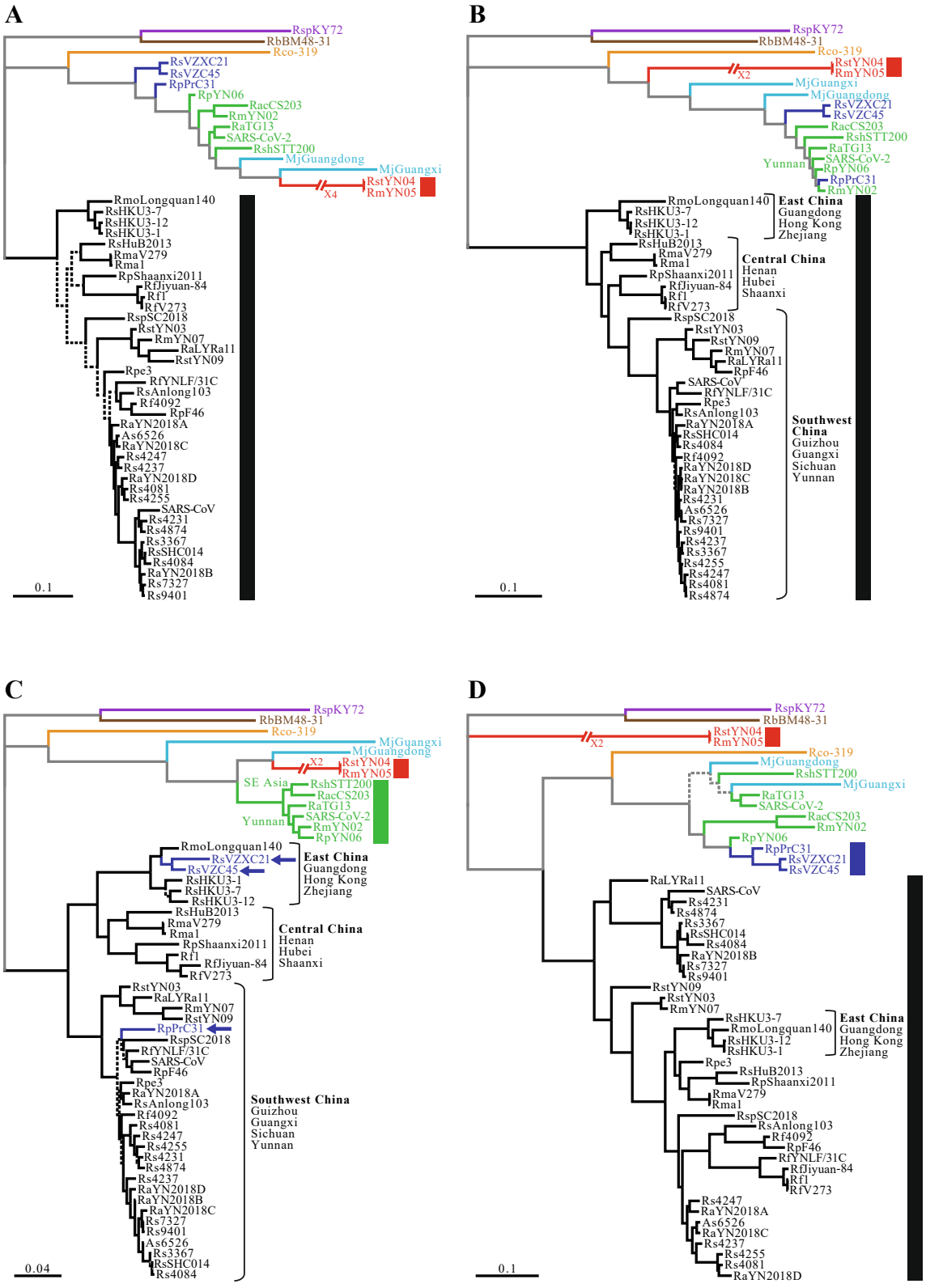

Fig. 4. Phylogenetic relationships within the subgenus *Sarbecovirus* based on (A) the whole genome alignment, (B) its 5′ region, (C) its central region, and (D) its 3′ region. To take into account the recombinant origin of the three *RecSar* viruses (RpPrC31, RsVZC45, and RsVZXC21; see main text for details), the alignment of *Sarbecovirus* genomes was partitioned into three subdatasets corresponding to the 5′ region (positions 1–11,517), central region (positions 12,970–20,289), and 3′ region (positions 20,364–29,550). All phylogenetic trees were reconstructed with MrBayes 3.2.7 (Ronquist et al., 2012) using different GTR + I + G models for the three codon-positions. They were rooted using RspKY72 and RbBM48-31 as outgroup. The eight groups of synonymous nucleotide composition (SNC) are highlighted by different colours as defined in Fig. 1A. Dash branches indicate nodes supported by posterior probability (PP) less than 0.95. All other nodes are supported by PP $\geq$ 0.95.

pattern was found in the tree based on the 5′ genomic region, except that the *SCoVrC* group of Central China and *SCoV2rC* group of SE Asia were paraphyletic (Fig. 4**B**). By contrast, the geographic pattern was poorly conserved in the tree based on the 3′ genomic region (Fig. 4**D**). Such a geographic structure was already mentioned in Boni et al. (2020) for several genomic regions, but without providing any explanation for its origin. I suggest that the geographic structure observed for the central genomic region is indicative of strong environmental selection acting on RdRp variants. According to this hypothesis, only recombinant genomes possessing RdRp variants adapted to their bat species host(s) are selected. The *YunSar* group could be a key *Sarbecovirus* lineage to test this hypothesis in the future. Indeed, *YunSar* appeared as the sister-group of MjGuangdong, and the two lineages were related to *SCoV2rC* in the tree based on the central genomic region (Fig. 4**C**). By contrast, *YunSar* was found to be much more divergent from MjGuangdong and *SCoV2rC* in the trees based on 5′ and 3′ genomic regions (Fig. 4**B and 4D**). In other words, these results suggest that the central genomic region of the ancestral *YunSar* virus has been acquired after recombination with a bat sarbecovirus more closely related MjGuangdong. Since the most likely origin of pangolin sarbecoviruses is Southeast Asia (Hassanin et al., 2021a), bats and pangolins from Laos, Myanmar, Thailand and Vietnam should be further investigated to better understand the recombinant origin of *YunSar*.

4.2. Evidence for eight groups of SNC among Sarbecovirus genomes

In this study, all analyses of nucleotide composition, dinucleotide composition and codon usage (Figs. 1-3) showed evidence for eight groups of *Sarbecovirus* genomes: (i) *SCoVrC*, including SARS-CoV and a large diversity of bat sarbecoviruses from China; (ii) *SCoV2rC*, including SARS-CoV-2 and five bat sarbecoviruses from Cambodia, Thailand and Yunnan, (iii) *PangSar*, which is composed of the two sarbecoviruses detected in Sunda pangolins seized in the Chinese provinces of Guangdong (in 2019) and Guangxi (in 2017–2018); (iv) *RecSar*, which contains the three bat sarbecoviruses showing evidence of past recombination between *SCoV2rC* and *SCoVrC* genomes; (v) *YunSar*, i.e., the two highly divergent bat sarbecoviruses recently described from Yunnan by Zhou et al. (2021; RmYN05 and RstYN04); (vi) RbBM48-31, the bat sarbecovirus from Bulgaria; (vii) RspKY72, the bat sarbecovirus from Kenya; and (viii) Rc-o319, the bat sarbecovirus from Japan. All these groups can be diagnosed by specific features (i.e., highest or lowest percentages) in nucleotide composition, dinucleotide composition, and/or codon usage (Tables 2-4). The only exception is *RecSar* for which all variables show intermediate values between those found for *SCoVrC* and *SCoV2rC*. This result can be however explained by their recombinant origin between two divergent *Sarbecovirus* lineages, *SCoV2rC* and *SCoVrC* (see also the genomic bootstrap barcodes recently published in Hassanin et al., 2022). As expected, their recombinant nature was confirmed by the separate SNC analyses of the three genomic regions: the three *RecSar* viruses clustered with *SCoV2rC* in the two PCAs based on 5′ and 3′ genomic regions (Fig. 5**A and 5C**), whereas they clustered with *SCoVrC* in the PCA based on the central genomic region (Fig. 5**B**).

4.3. Viral RNA replication in different hosts is the main evolutionary force behind the variation in SNC

The SNC is the primary factor explaining the similar results also observed at synonymous sites of dinucleotides and codons. Indeed, the two variables factor maps obtained from PCAs based on the nucleotide composition at third codon-positions (Fig. 1**A and 1B**) revealed that the variance is mainly explained by the first dimension (89.08% and 73.07%, respectively), which separates *Sarbecorvirus* genomes showing the highest A + U content at third codon positions (at the left; *SCoV2rC*, *PangSar*, *RecSar*, and RspKY72; A + U (3CP) = 71.9–68.0%) from the other ones (at the right; *SCoVrC*, RbBM48-3, Rc-o319, and *YunSar*; A + U (3CP) = 66.1–63.1%). Similar results were found in the PCAs based on

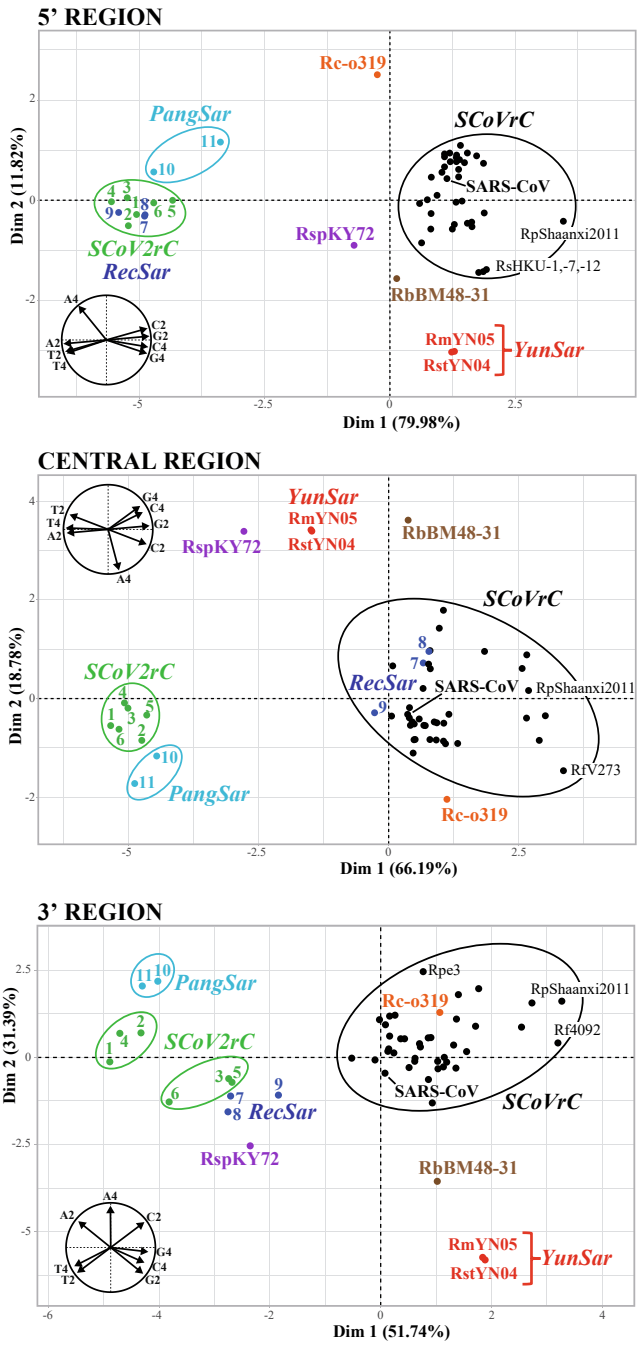

Fig. 5. Variation in synonymous nucleotide composition in three regions of *Sarbecovirus* genomes: (A) 5′ region; (B) central region; and (C) 3′ region. The alignment of *Sarbecovirus* genomes was partitioned into three sub-datasets corresponding to the 5′ region (positions 1–11,517), central region (positions 12,970–20,289), and 3′ region (positions 20,364–29,550). For each of the three genomic regions, the frequency of the four bases was calculated either at four-fold degenerate third codon-positions or at two-fold degenerate third codon-positions for either purines (A *versus* G) or pyrimidines (C *versus* U) (Table 5). The three PCAs based on eight variables and their variables factor map are shown in Figures A, B and C. The eight groups of synonymous nucleotide composition (SNC) are highlighted by different colours as defined in Fig. 1A.

dinucleotide composition (Fig. 2**A and 2B**) and codon usage (Fig. 3), with *SCoV2rC* and *PangSar* genomes exhibiting a more marked bias towards A and U nucleotides (or against C + G nucleotides) at synonymous sites. All these results support a stronger mutational bias in *SCoV2rC* and

PangSar genomes characterised by higher rates for C=>U and G=>A transitions than for the reverse mutations (U=>C and A=>G, respectively). Previous studies examining the nucleotide composition in SARS-CoV-2 genomes have all concluded to an over-representation of C=>U transitions (Rice et al. 2021; Matyášek et al. 2021). Several mechanisms have been proposed to account for the cytosine deficiency in the genome of sarbecoviruses, such as cytosine deamination resulting from the action of the host APOBEC3 system (Milewska et al. 2018), methylation of CpG dinucleotides (Xia and Kumar, 2020), or the limited availability of cytidine triphosphate (CTP), which is used not only for the viral RNA genome synthesis but also for the synthesis of the virus envelope, as well as translation and glycosylation of viral proteins (Ou et al., 2021). My results indicated that CG dinucleotides are less frequent than other dinucleotides at both P23 and P31 sites, confirming therefore the selection against CpG dinucleotide discussed in previous studies (Daron & Bravo, 2021). However, this is obviously not the main mechanism explaining the differences between the eight SNC groups. Indeed, the bias against C and G nucleotides observed in third codon-positions, which appeared more marked for *PangSar* and *SCoV2rC* (Table 2), was also detected by comparing the frequencies of dinucleotides (Table 3) or four-fold degenerate codons (amino-acids A, G, P, T, and V in Table 4). In agreement with several previous studies, my results confirmed therefore that mutational bias is the main force shaping codon usage in sarbecoviruses (Tort et al., 2020; Daron & Bravo, 2021; Simmonds and Ansari 2021). Importantly, the bias in favour of C=>U transitions (over U=>C transitions) has been observed in a wide range of mammalian RNA viruses (Simmonds and Ansari 2021), indicating that it is the result of an asymmetrical mechanism shared by all RNA viruses infecting mammals. I suggest that the replication of viral RNA genomes, which is an asymmetrical process dependent of the pool of free nucleotides available in infected cells, can explain the eight SNC patterns here observed among *Sarbecovirus* genomes. From this point of view, it is important to note that the physiological concentrations of nucleotides were found to be highly variable among mammalian species (i.e., human, mouse, rat, etc.) and tissues, with the following means and standard deviations (in μM) published in Traut (1994): ATP = 3,152 ± 1,698 > UTP = 567 ± 460 > GTP = 468 ± 224 > CTP = 278 ± 242. When a mammalian cell divides, the synthesis of nucleotides is regulated at multiple levels to maintain enough levels of free nucleotides for DNA replication (Lane and Fan 2015). By contrast, the replication of viral RNA genomes always takes place in mammalian cells in which the nucleotide concentrations are in the order ATP ≫ UTP > GTP > CTP (Traut 1994), which is supposed to promote higher mutation rates for G=>A transitions (*versus* A=>G transitions) and to a lesser extent C=>U transitions (*versus* U=>C transitions). In addition, the availability of CTP is much more reduced when RNA viruses multiply in their mammalian host cells (Ou et al. 2021), increasing therefore the rate of C=>U transitions during viral replication. Therefore, the biased concentrations in favour of UTP over CTP and of ATP over GTP can explain why *Sarbecovirus* genomes are found to be rich in U and A nucleotides at synonymous sites (Table 2). In addition, some variations in the concentrations of free nucleotides between bat reservoir species (or species assemblages) hosting sarbecoviruses may have imposed different mutational rates between the main *Sarbecovirus* lineages. In other words, I suggest that host switching is the main evolutionary force behind the variation in SNC here observed among *Sarbecovirus* genomes. This hypothesis is supported by three main arguments: (i) as detailed in the next paragraph, the eight SNC groups of *Sarbecovirus* genomes were found in different species or species assemblages (Fig. 6); (ii) the data published by Traut (1994) have revealed that the concentrations of free nucleotides can be variable between mammalian species (human, mouse and rat), suggesting that a switch to a new host species can impose different concentrations of free

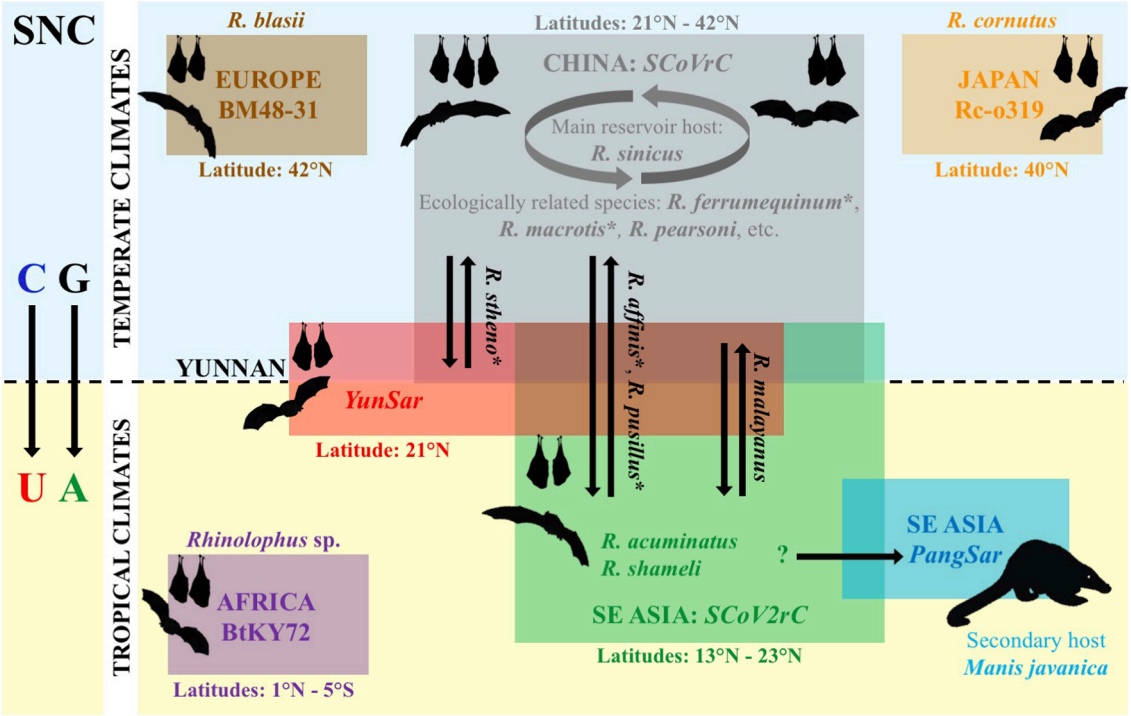

Fig. 6. Host species and latitudinal distribution of the seven groups of *Sarbecovirus* genomes showing different synonymous nucleotide compositions (SNC). The seven SNC groups of *Sarbecovirus* genomes are highlighted by coloured rectangles. The eighth SNC group, *RecSar*, was not considered here as the three genomes RpPrC31, RsZC45 and RsZXC21 showed a mixed SNC between *SCoVrC* and *SCoV2rC* (see main text for details). The abbreviation "*R.*" is used for *Rhinolophus* species. The double arrows indicate *Rhinolophus* species from which several SNC groups of *Sarbecovirus* were sequenced in previous studies. All species names concerned by taxonomic issues (see Table 1 for details) are followed by an asterisk. As shown in Table 2, the genomic bias in favour of A + U nucleotides is more marked for the SNC groups of sarbecoviruses circulating in bats (or pangolins) from tropical latitudes (BtKY72 in Sub-Saharan Africa and *SCoV2rC* and *PangSar* in Southeast Asia) than for those from temperate latitudes (BM48-31 in Europe, *SCoVrC* in China, and Rc-o319 in Japan).

nucleotides, and therefore different mutational rates; and (iii) in agreement with this view, several recent studies have concluded that the codon usage of SARS-CoV-2 adapts to the human lung environment (Li et al., 2020; Zhang et al., 2021).

The genus *Rhinolophus* currently includes between 92 and 109 insectivorous bat species (Burgin et al. 2020; IUCN 2021) that inhabit temperate and tropical regions of the Old World, with a higher biodiversity in Asia (63–68 out of the 92–109 described species) than in Africa (34–38 species), Europe (5 species) and Oceania (5 species). All *Rhinolophus* species in which sarbecoviruses were detected in previous studies (Table 1) are cave dwellers that form small groups or colonies (up to several hundreds) (IUCN 2021). As previously reviewed in Hassanin et al. (2021a,b), there is strong evidence that the genus *Rhinolophus* consitutes the reservoir host in which sarbecoviruses have evolved for centuries. The sarbecoviruses are thought to circulate among bat populations of the main reservoir host species, but other bat species may be occasionally or regularly infected. Importantly, the six groups of bat *Sarbecovirus* genomes showing different SNCs have distinct geographic distributions: China and several bordering countries for *SCoVrC* (and *RecSar*); southern Yunnan and mainland Southeast Asia for *SCoV2rC*; Yunnan for *YunSar*; Japan for Rc-o319; Bulgaria for RbBM48-31; and Kenya for RspKY72. This suggests that the six groups of sarbecoviruses have evolved in different *Rhinolophus* reservoirs, each of them being potentially represented by several ecologically related species. Out of Asia, there are two groups of sarbecoviruses, each of them known from a unique virus (Fig. 6): RbBM48-31 isolated from *Rhinolophus blasii* in Bulgaria (southeastern Europe) (Drexler et al. 2010), and RspKY72 from Kenya (East Africa), for which the *Rhinolophus* species was not identified in Tao and Tong (2019). In Asia, there are four groups of sarbecoviruses: Rc-o319, *SCoVrC*, *SCoV2rC*, and *YunSar* (Fig. 6). The Rc-o319 virus was recently discovered in Japan using fecal samples from *Rhinolophus cornutus* (Murakami et al. 2020), a species endemic to Japanese islands (Burgin et al. 2020). The high nucleotide divergence between Rc-o319 and other sarbecoviruses (between 20% and 26%) supports its evolution in allopatry due to the insular isolation of its bat reservoir. It must be however noted that the species *Rhinolophus nippon* (which is still treated as a subspecies of *Rhinolophus ferrumequinum* in the classification of IUCN, but not in that of Burgin et al. 2020) is distributed on both sides of the Sea of Japan, suggesting that some sarbecoviruses may have occasionally dispersed through long-distance flights between bat populations from the Korean Peninsula and Japan. For *SCoVrC*, many genomic sequences were published during the two last decades because sarbecoviruses have been actively sought in all Chinese provinces after the 2002–2004 SARS epidemic. Although a few *SCoVrC* viruses were detected in bat genera other than *Rhinolophus*, such as *Aselliscus* (Hu et al. 2017) or *Chaerephon* (Yang et al. 2013), the great majority of *SCoVrC* were isolated from *Rhinolophus* species, and most of them were found in *Rhinolophus sinicus*. The available data suggest therefore that *Rhinolophus sinicus* could be the main reservoir species for *SCoVrC*. In support of this hypothesis, the distribution of *Rhinolophus sinicus* (Burgin et al. 2020; IUCN 2021) fits well with the ecological niche recently inferred for *SCoVrC* (Hassanin et al. 2021b). It appears much more difficult to determine the main reservoir host species for *SCoV2rC*. Indeed, the five currently known *SCoV2rCs* were identified in five distinct species of Chiroptera: *Rhinolophus affinis*, *Rhinolophus malayanus*, and *Rhinolophus pusillus* in Yunnan (Zhou H. et al. 2020, 2021; Zhou P. et al. 2020), *Rhinolophus acuminatus* in eastern Thailand (Wacharapluesadee et al. 2021) and *Rhinolophus shameli* in northern Cambodia (Delaune et al. 2021). Two of these species, *Rhinolophus affinis* and *Rhinolophus pusillus*, are assumed to be largely distributed in China and Southeast Asia (IUCN 2021), but they belong to two different species complexes in which the taxonomy is confused and needs to be clarified (Wu et al. 2012; Soisook et al. 2016; Srinivasulu et al. 2019; Mao and Rossiter 2020). The three other species, *Rhinolophus acuminatus*, *Rhinolophus malayanus*, and *Rhinolophus shameli* are endemic to Southeast Asia (Burgin et al. 2020; IUCN 2021), although the distribution of *Rhinolophus malayanus* has been recently extended to the Yunnan province (Liang et al. 2020). As a consequence, the ecological niche predicted for *SCoV2rC* was found to be different from that of *SCoVrC* (Hassanin et al. 2021b): it includes southern Yunnan and several regions of Laos, Vietnam, Cambodia, Myanmar and Thailand. The two *YunSar* viruses here analysed (RmYN05 and RstYN04) were recently described from two different bat species, *Rhinolophus malayanus* and *Rhinolophus stheno*, collected between May 2019 and November 2020 in Mengla county, Yunnan province (Zhou et al. 2021). The *YunSar* genomes are divergent from other *Sarbecovirus* genomes (between 23% and 27%) and their SNC revealed extremes values for many variables (2/12 in Table 2; 12/32 in Table 3; 19/59 in Table 4). More recently, eight *YunSar* viruses showing 98% of genomic identity with RmYN05 and RstYN04 have been described from *Rhinolophus affinis* and *Rhinolophus stheno* collected in May 2015 in Mojiang County, Yunnan province (Guo et al. 2021). Although current data indicate that the *YunSar* group is endemic to Yunnan, other regions should be explored to better characterise its geographic distribution, including North East India, nothern Myanmar, nothern Laos, northern Thailand, and nothern Vietnam.

Biogeographically, the most striking result of this study is that the first dimension of all PCAs of Figs. 1-3 and 5 allowed to separate temperate *versus* tropical groups of *Sarbecovirus*. Indeed, two latitudinal groups can be separated in Asia (Fig. 6): the tropical group contains *SCoV2rC*, for which the ecological niche was inferred to include southern Yunnan and several regions of mainland Southeast Asia (Hassanin et al. 2021b), and *PangSar*, for which the most likely origin is Southeast Asia (Hassanin et al. 2021a); and the temperate group is composed of *SCoVrC*, for which the ecological niche was inferred to contain most southern and eastern provinces of China, as well as the Korean Peninsula, Japan, Taiwan, northeastern India, and northern regions of Myanmar and Vietnam (Hassanin et al. 2021b), Rc-o319, which is a sarbecovirus from Japan, and *YunSar*, which is currently endemic to Yunnan. Similarly, two latitudinal groups can be separated in the western Old World (Fig. 6): RspKY72 from Kenya in East Africa *versus* RbBM48-31 from Bulgaria in Europe. I suggest that hibernation of bat reservoirs could explain the SNC differences here observed between *Sarbecovirus* genomes from temperate *versus* tropical latitudes. Indeed, most temperate species of *Rhinolophus* found in China, Europe and Japan have to hibernate in winter (Burgin et al. 2020; IUCN 2021) when insect populations become significantly less abundant. By contrast, *Rhinolophus* species found at intertropical latitudes, i.e., between the Tropics of Capricorn (23°S) and Cancer (23°N), do not hibernate because insects are available all year round. In temperate climates, bat hibernation can impact the SNC of *Sarbecovirus* genomes *via* two possible mechanisms: (i) viral replication can be significantly reduced in hibernating bats, and this may explain the lower winter prevalence of coronaviruses in hibernating bats (e.g., Lo et al. 2020 for Korean bats); and (ii) the concentrations of free nucleotides available in the cells of hibernating bats can be strongly modified due to the reduction and remodelling of many metabolic pathways (Andrews 2007).

4.4. Why SCoV2rC and PangSar show similar but different SNCs?

All PCAs of this study showed that the SNCs of *SCoV2rC* genomes are similar to those found for *PangSar* genomes. Indeed, the two groups share extreme values (highest or lowest percentages) for many variables (Tables 2-4), including the highest percentages of A nucleotide and lowest percentage of G nucleotide at third codon-positions, the highest percentages of U nucleotide and lowest percentages of C and G nucleotides at four-fold degenerate third codon-positions, as well as the highest percentages of A nucleotide at two-fold degenerate third codon-positions. As previously discussed, these results suggest higher levels of C=>U and G=>A transitions in the genomes of *SCoV2rC* and *PangSar* than in those of other viral lineages (i.e., RbBM48-31, Rc-o319, RspKY72, *SCoVrC*, and *YunSar*). All these elements suggest that *SCoV2rC* and *PangSar* have originally evolved in the same bat reservoir, which

may have included several ecologically related species of *Rhinolophus* distributed in the ecological niche of *SCoV2rC*, i.e., in southern Yunnan and several regions of mainland Southeast Asia (Hassanin et al. 2021b). This hypothesis implies that pangolins are secondary hosts for sarbecoviruses, which is corroborated by the diversity of *Sarbecovirus* lineages found in *Rhinolophus* species (*SCoVrC*, *SCoV2rC*, RbBM48-31, Rc-o319, RspKY72, and *YunSar*) and by the internal placement of the two divergent pangolin sarbecoviruses in the phylogenetic trees (Fig. 4; Zhou H. et al. 2020, 2021; Zhou P. et al. 2020; Delaune et al. 2021; Wacharapluesadee et al. 2021).

Despite their high SNC similarities, *SCoV2rC* and *PangSar* genomes appeared in two different clusters in all PCAs (Figs. 1-3, 5). In addition, the two groups exhibit special features. On the one hand, the six *SCoV2rC* genomes show more U nucleotides and less C nucleotides than *PangSar* genomes at third codon-positions (Table 2). On the other hand, the two *PangSar* genomes (MjGuangdong and MjGuangxi) exhibit more A nucleotides than *SCoV2rC* genomes at third codon-positions (Table 2). Importantly, none of the phylogenetic analyses of Fig. 4 supported the monophyly of *PangSar*: based on the 5′ genomic region, MjGuangdong was grouped with *SCoV2rC* and *RecSar* (PP = 1); based on the central genomic region, MjGuangdong was allied to *YunSar* (PP = 1), with *SCoV2rC* as their sister-group (PP = 1); and based on the 3′ genomic region, MjGuangdong and MjGuangxi were included into the same clade than *SCoV2rC* and *RecSar* (PP = 1). The MjGuangdong genome shares between 88% and 90% of identity with *SCoV2rC* genomes, but only 85% with the MjGuangxi genome. Taken together, all these results suggest that the similar SNCs observed in the two *PangSar* genomes (MjGuangdong and MjGuangxi) were not inherited from a common ancestor, but have been rather acquired by convergence, most probably as a consequence of the shift from their original bat reservoir(s) to Sunda pangolins. As discussed previously, the replication process is dependent of the pool of free nucleotides available in the host cell. As a consequence, variations in the concentrations of ATP, CTP, GTP and UTP among mammalian species (Traut 1994) may have imposed different mutational pressures in case of viral host-shift to a new mammalian species, i.e., from bats to pangolins. In this regard, it is important to note that the human SARS-CoV-2 genome (NC_045512, patient admitted to the Central Hospital of Wuhan on 26 December 2019; Wu et al. 2021) was never grouped to *PangSar* genomes in the PCAs based on SNC (Figs. 1–3, 5), whereas it was always closely associated with bat *SCoV2rC* genomes. The results support therefore that SARS-CoV-2 emerged directly from a bat sarbecovirus, without pangolin intermediate host.

CRediT authorship contribution statement

Alexandre Hassanin: Conceptualization, Methodology, Formal analysis, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability.

All data used for the analyses (multiple genome alignment and csv files used for PCAs) are publicly available in the Open Science Framework platform at https://osf.io/rv5u9/.

Funding

This research was funded by the AAP RA-COVID-19, grant number ANR-21-CO12-0002.

Acknowledgements

I would like to thank the two anonymous reviewers for their helpful comments on the first version of the manuscript.

References

Andrews, M.T., 2007. Advances in molecular biology of hibernation in mammals. BioEssays : news and reviews in molecular, cellular and developmental biology 29 (5), 431–440. https://doi.org/10.1002/bies.20560.

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. Nat. Microbiol. 5 (11), 1408–1417. https://doi.org/10.1038/s41564-020-0771-4.

Burgin, C.J., Wilson, D.E., Mittermeier, R.A., Rylands, A.B., Lacher, T.E., Sechrest, W., 2020. Illustrated Checklist of the Mammals of the World, Vol. 2. Lynx Edicions, Barcelona.

Daron, J., Bravo, I.G., 2021. Variability in codon usage in coronaviruses is mainly driven by mutational bias and selective constraints on CpG dinucleotide. Viruses 13 (9), 1800. https://doi.org/10.3390/v13091800.

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. Nat. Commun. 12 (1), 6563. https://doi.org/10.1038/s41467-021-26809-4.

Drexler, J.F., Gloza-Rausch, F., Glende, J., Corman, V.M., Muth, D., Goettsche, M., Seebens, A., Niedrig, M., Pfefferle, S., Yordanov, S., Zhelyazkov, L., Hermanns, U., Vallo, P., Lukashev, A., Müller, M.A., Deng, H., Herrler, G., Drosten, C., 2010. Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. J. Virol. 84 (21), 11336–11349. https://doi.org/10.1128/JVI.00650-10.

Fan, Y., Zhao, K., Shi, Z.-L., Zhou, P., 2019. Bat Coronaviruses in China. Bat Coronaviruses in China. Viruses 11 (3), 210. https://doi.org/10.3390/v11030210.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Stein, D., Israeli, O., Beth-Din, A., Melamed, S., Weiss, S., Israely, T., Paran, N., Schwartz, M., Stern-Ginossar, N., 2021. The coding capacity of SARS-CoV-2. Nature 589 (7840), 125–130. https://doi.org/10.1038/s41586-020-2739-1.

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., Wang, T., Sun, Q., Ming, Z., Zhang, L., Ge, J.i., Zheng, L., Zhang, Y., Wang, H., Zhu, Y., Zhu, C., Hu, T., Hua, T., Zhang, B., Yang, X., Li, J., Yang, H., Liu, Z., Xu, W., Guddat, L.W., Wang, Q., Lou, Z., Rao, Z., 2020. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 368 (6492), 779–782.

Ge, X.Y., Li, J.L., Yang, X.L., Chmura, A.A., Zhu, G., Epstein, J.H., Mazet, J.K., Hu, B., Zhang, W., Peng, C., Zhang, Y.J., Luo, C.M., Tan, B., Wang, N., Zhu, Y., Crameri, G., Zhang, S.Y., Wang, L.F., Daszak, P., Shi, Z.L., 2013. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature 503 (7477), 535–538. https://doi.org/10.1038/nature12711.

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.

Guo, H., Hu, B., Si, H.-R., Zhu, Y., Zhang, W., Li, B., Li, A., Geng, R., Lin, H.-F., Yang, X.-L., Zhou, P., Shi, Z.-L., 2021. Identification of a novel lineage bat SARS-related coronaviruses that use bat ACE2 receptor. Emerging Microbes Infect. 10 (1), 1507–1514.

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. Front. Microbiol. 10, 1900. https://doi.org/10.3389/fmicb.2019.01900.

Hassanin, A., Grandcolas, P., Veron, G., 2021a. Covid-19: natural or anthropic origin? Mammalia 85, 1–7. https://doi.org/10.1515/mammalia-2020-0044.

Hassanin, A., Tu, V.T., Curaudeau, M., Csorba, G., 2021b. Inferring the ecological niche of bat viruses closely related to SARS-CoV-2 using phylogeographic analyses of Rhinolophus species. Sci. Rep. 11 (1), 14276. https://doi.org/10.1038/s41598-021-93738-z.

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.

He, B., Zhang, Y., Xu, L., Yang, W., Yang, F., Feng, Y., Xia, L., Zhou, J., Zhen, W., Feng, Y.e., Guo, H., Zhang, H., Tu, C., Perlman, S., 2014. Identification of diverse alphacoronaviruses and genomic characterization of a novel severe acute respiratory syndrome-like coronavirus from bats in China. J. Virol. 88 (12), 7070–7082. https://doi.org/10.1128/JVI.00631-14.

He, R., Dobie, F., Ballantine, M., Leeson, A., Li, Y., Bastien, N., Cutts, T., Andonov, A., Cao, J., Booth, T.F., Plummer, F.A., Tyler, S., Baker, L., Li, X., 2004. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem. Biophys. Res. Commun. 316 (2), 476–483. https://doi.org/10.1016/j.bbrc.2004.02.074.

Hon, C.C., Lam, T.Y., Shi, Z.L., Drummond, A.J., Yip, C.W., Zeng, F., Lam, P.Y., Leung, F.C., 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. J. Virol. 82 (4), 1819–1826. https://doi.org/10.1128/JVI.01926-07.

Hu, B., Zeng, L.-P., Yang, X.-L., Ge, X.-Y., Zhang, W., Li, B., Xie, J.-Z., Shen, X.-R., Zhang, Y.-Z., Wang, N., Luo, D.-S., Zheng, X.-S., Wang, M.-N., Daszak, P., Wang, L.-F., Cui, J., Shi, Z.-L., Drosten, C., 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. 13 (11), e1006698. https://doi.org/10.1371/journal.ppat.1006698.

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y.i., Ye, F., Yang, L.u., Ding, C., Zhu, X., Lv, R., Zhu, J., Hassan, B., Feng, Y., Tan, W., Wang, C., 2018. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerging Microbes Infect. 7 (1), 1–10. https://doi.org/10.1038/s41426-018-0155-5.

IUCN 2021. The IUCN Red List of Threatened Species. Version 2021-1. https://www.iucnredlist.org. Downloaded on 15 July 2021.

Kandeel, M., Ibrahim, A., Fayez, M., Al-Nazawi, M., 2020. From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. J. Med. Virol. 92 (6), 660–666. https://doi.org/10.1002/jmv.25754.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780. https://doi.org/10.1093/molbev/mst010.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217 (5129), 624–626. https://doi.org/10.1038/217624a0.

Lam, T.-Y., Jia, N.a., 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.-M., Li, L.-F., Sun, F.-F., Qin, S.i., Huang, J.-C., Leung, G.M., Holmes, E.C., Hu, Y.-L., Guan, Y.i., 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.

Lane, A.N., Fan, T.W., 2015. Regulation of mammalian nucleotide metabolism and biosynthesis. Nucleic Acids Res. 43 (4), 2466–2485. https://doi.org/10.1093/nar/gkv047.

Larsson, A., 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics (Oxford, England) 30 (22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531.

Lau, S.K.P., Feng, Y., Chen, H., Luk, H.K.H., Yang, W.-H., Li, K.S.M., Zhang, Y.-Z., Huang, Y.i., Song, Z.-Z., Chow, W.-N., Fan, R.Y.Y., Ahmed, S.S., Yeung, H.C., Lam, C. S.F., Cai, J.-P., Wong, S.S.Y., Chan, J.F.W., Yuen, K.-Y., Zhang, H.-L., Woo, P.C.Y., Perlman, S., 2015. Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination. J. Virol. 89 (20), 10532–10547. https://doi.org/10.1128/JVI.01048-15.

Lau, S.K., Li, K.S., Huang, Y., Shek, C.T., Tse, H., Wang, M., Choi, G.K., Xu, H., Lam, C.S., Guo, R., Chan, K.H., Zheng, B.J., Woo, P.C., Yuen, K.Y., 2010. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. J. Virol. 84 (6), 2808–2819. https://doi.org/10.1128/JVI.02219-09.

Lau, S.K., Woo, P.C., Li, K.S., Huang, Y., Tsoi, H.W., Wong, B.H., Wong, S.S., Leung, S.Y., Chan, K.H., Yuen, K.Y., 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. PNAS 102 (39), 14040–14045. https://doi.org/10.1073/pnas.0506735102.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: An R package for multivariate analysis. J. Stat. Softw. 25, 1–18. https://doi.org/10.18637/jss.v025.i01.

Li, L.L., Wang, J.L., Ma, X.H., Sun, X.M., Li, J.S., Yang, X.F., Shi, W.F., Duan, Z.J., 2021. A novel SARS-CoV-2 related coronavirus with complex recombination isolated from bats in Yunnan province. China. Emerging microbes & infections 10 (1), 1683–1690. https://doi.org/10.1080/22221751.2021.1964925.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B.T., Zhang, S., Wang, L.F., 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science (New York, N.Y.) 310 (5748), 676–679. https://doi.org/10.1126/science.1118391.

Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., Jiang, W., 2020. GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. Molecular genetics and genomics : MGG 295 (6), 1537–1546. https://doi.org/10.1007/s00438-020-01719-0.

Liang, J., Hea, X., Peng, X., Xie, H., Zhang, L., 2020. First record of existence of *Rhinolophus malayanus* (Chiroptera, Rhinolophidae) in China. Mammalia 84 (4), 362–365. https://doi.org/10.1515/mammalia-2019-0062.

Lin, X.D., Wang, W., Hao, Z.Y., Wang, Z.X., Guo, W.P., Guan, X.Q., Wang, M.R., Wang, H. W., Zhou, R.H., Li, M.H., Tang, G.P., Wu, J., Holmes, E.C., Zhang, Y.Z., 2017. Extensive diversity of coronaviruses in bats from China. Virology 507, 1–10. https://doi.org/10.1016/j.virol.2017.03.019.

Lo, V.T., Yoon, S.-W., Noh, J.Y., Kim, Y., Choi, Y.G., Jeong, D.G., Kim, H.K., 2020. Long-term surveillance of bat coronaviruses in Korea: Diversity and distribution pattern. Transboundary and emerging diseases 67 (6), 2839–2848. https://doi.org/10.1111/tbed.13653.

Mao, X., Rossiter, S.J., 2020. Genome-wide data reveal discordant mitonuclear introgression in the intermediate horseshoe bat (*Rhinolophus affinis*). Mol. Phylogenet. Evol. 150, 106886. https://doi.org/10.1016/j.ympev.2020.106886.

Matyášek, R., Řehůřková, K., Berta Marošiová, K., Kovařík, A., 2021. Mutational Asymmetries in the SARS-CoV-2 Genome May Lead to Increased Hydrophobicity of Virus Proteins. Genes 12 (6), 826. https://doi.org/10.3390/genes12060826.

Milewska, A., Kindler, E., Vkovski, P., Zeglen, S., Ochman, M., Thiel, V., Rajfur, Z., Pyrc, K., 2018. APOBEC3-mediated restriction of RNA virus replication. Sci. Rep. 8 (1), 5960. https://doi.org/10.1038/s41598-018-24448-2.

Murakami, S., Kitamura, T., Suzuki, J., Sato, R., Aoi, T., Fujii, M., Matsugo, H., Kamiki, H., Ishida, H., Takenaka-Uema, A., Shimojima, M., Horimoto, T., 2020. Detection and Characterization of Bat Sarbecovirus Phylogenetically Related to SARS-CoV-2. Japan. Emerging infectious diseases 26 (12), 3025–3029. https://doi.org/10.3201/eid2612.203386.

Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23 (1), 263–286. https://doi.org/10.1146/annurev.es.23.110192.001403.

Ou, Z., Ouzounis, C., Wang, D., Sun, W., Li, J., Chen, W., Marlière, P., & Danchin, A. (2020). A Path toward SARS-CoV-2 Attenuation: Metabolic Pressure on CTP Synthesis Rules the Virus Evolution. Genome biology and evolution, 12(12), 2467–2485. https://doi.org/10.1093/gbe/evaa229.

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., van der Poel, W.H.M., 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.

Rice, A. M., Castillo Morales, A., Ho, A. T., Mordstein, C., Mühlhausen, S., Watson, S., Cano, L., Young, B., Kudla, G., & Hurst, L. D. (2021). Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design. Molecular biology and evolution, 38(1), 67–83. https://doi.org/10.1093/molbev/msaa188.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61 (3), 539–542. https://doi.org/10.1093/sysbio/sys029.

Simmonds, P., Ansari, M.A., 2021. Extensive C->U transition biases in the genomes of a wide range of mammalian RNA viruses; potential associations with transcriptional mutations, damage- or host-mediated editing of viral RNA. PLoS Pathog. 17 (6), e1009596 https://doi.org/10.1371/journal.ppat.1009596.

Soisook, P., Karapan, S., Srikrachang, M., Dejtaradol, A., Nualcharoen, K., Bumrungsri, S., Lin Oo, S.S., Aung, M.M., Bates, P.J.J., Harutyunyan, M., Buś, M.M., Bogdanowicz, W., 2016. Hill forest dweller: A new cryptic species of *Rhinolophus* in the '*pusillus* group' (Chiroptera: Rhinolophidae) from Thailand and Lao PDR. Acta Chiropterologica 18 (1), 117–139. https://doi.org/10.3161/15081109ACC2016.18.1.005.

Srinivasulu, C., Srinivasulu, A., Srinivasulu, B., Jones, G., Vincenot, C., 2019. Integrated approaches to identifying cryptic bat species in areas of high endemism: The case of *Rhinolophus andamanensis* in the Andaman Islands. PLoS ONE 14 (10), e0213562. https://doi.org/10.1371/journal.pone.0213562.

Swofford, D. L. (2003). PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland, MA: Sinauer Associates.

Tang, X.C., Zhang, J.X., Zhang, S.Y., Wang, P., Fan, X.H., Li, L.F., Li, G., Dong, B.Q., Liu, W., Cheung, C.L., Xu, K.M., Song, W.J., Vijaykrishna, D., Poon, L.L., Peiris, J.S., Smith, G.J., Chen, H., Guan, Y., 2006. Prevalence and genetic diversity of coronaviruses in bats from China. J. Virol. 80 (15), 7481–7490. https://doi.org/10.1128/JVI.00697-06.

Tao, Y., Tong, S., 2019. Complete Genome Sequence of a Severe Acute Respiratory Syndrome-Related Coronavirus from Kenyan Bats. Microbiology resource announcements 8 (28), e00548–e00619. https://doi.org/10.1128/MRA.00548-19.

Tort, F.L., Castells, M., Cristina, J., 2020. A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. Virus Res. 283, 197976. https://doi.org/10.1016/j.virusres.2020.197976.

Traut, T.W., 1994. Physiological concentrations of purines and pyrimidines. Mol. Cell. Biochem. 140 (1), 1–22. https://doi.org/10.1007/BF00928361.

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., Thiel, V., 2021. Coronavirus biology and replication: implications for SARS-CoV-2. Nat. Rev. Microbiol. 19 (3), 155–170. https://doi.org/10.1038/s41579-020-00468-6.

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.-W., Sirichan, N., Ruchisrisarod, C., Rodpan, A., Noradechanon, K., Phaichana, T., Jantarat, N., Thongnumchaima, B., Tu, C., Crameri, G., Stokes, M.M., Hemachudha, T., Wang, L.-F., 2021. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. Nat. Commun. 12 (1) https://doi.org/10.1038/s41467-021-21240-1.

Wang, L., Fu, S., Cao, Y., Zhang, H., Feng, Y., Yang, W., Nie, K., Ma, X., Liang, G., 2017. Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern China. Emerging Microbes Infect. 6 (1), 1–8. https://doi.org/10.1038/emi.2016.140.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3.

Wu, Y., Motokawa, M., Harada, M., Thong, V.D., Lin, L.-K., Li, Y.-C., 2012. Morphometric variation in the *pusillus* group of the genus *Rhinolophus* (Mammalia: Chiroptera: Rhinolophidae) in East Asia. Zoolog. Sci. 29 (6), 396–402. https://doi.org/10.2108/zsj.29.396.

Wu, Z., Yang, L., Ren, X., Zhang, J., Yang, F., Zhang, S., Jin, Q., 2016. ORF8-Related Genetic Evidence for Chinese Horseshoe Bats as the Source of Human Severe Acute Respiratory Syndrome Coronavirus. J. Infect. Dis. 213 (4), 579–583. https://doi.org/10.1093/infdis/jiv476.

Xia X. (2020). Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. Molecular biology and evolution, 37(9), 2699–2705. https://doi.org/10.1093/molbev/msaa094.

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J. J., Li, N., Guo, Y., Li, X., Shen, X., Zhang, Z., Shu, F., Huang, W., Li, Y., Zhang, Z., Chen, R. A., Wu, Y. J., Peng, S. M., Huang, M., Xie, W. J., … Shen, Y. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature, 583(7815), 286–289. https://doi.org/10.1038/s41586-020-2313-x.

Yang, L., Wu, Z., Ren, X., Yang, F., He, G., Zhang, J., Dong, J., Sun, L., Zhu, Y., Du, J., Zhang, S., Jin, Q., 2013. Novel SARS-like betacoronaviruses in bats, China, 2011. Emerg. Infect. Dis. 19 (6), 989–991. https://doi.org/10.3201/eid1906.121648.

Zhang, Y., Jin, X., Wang, H., Miao, Y., Yang, X., Jiang, W., & Yin, B. (2021). Compelling Evidence Suggesting the Codon Usage of SARS-CoV-2 Adapts to Human After the Split From RaTG13. Evolutionary bioinformatics online, 17, 11769343211052013. https://doi.org/10.1177/11769343211052013.

Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C., Hughes, A.C., Bi, Y., Shi, W., 2020a. 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. https://doi.org/10.1016/j.cub.2020.05.023.

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). https://doi.org/10.1016/j.cell.2021.06.008, 00709-1.

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.i., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 2020b. 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.