



**HAL**  
open science

## Phylogeography of horseshoe bat sarbecoviruses in Vietnam and neighbouring countries. Implications for the origins of SARS - CoV and SARS - CoV -2

Alexandre Hassanin, Vuong Tan Tu, Tamás Görföl, Lam Quang Ngon, Phu Van Pham, Chu Thi Hang, Tran Anh Tuan, Matthieu Prot, Etienne Simon-Loriere, Gábor Kemenesi, et al.

### ► To cite this version:

Alexandre Hassanin, Vuong Tan Tu, Tamás Görföl, Lam Quang Ngon, Phu Van Pham, et al.. Phylogeography of horseshoe bat sarbecoviruses in Vietnam and neighbouring countries. Implications for the origins of SARS - CoV and SARS - CoV -2. *Molecular Ecology*, 2024, 33 (18), pp.e17486. 10.1111/mec.17486 . hal-04892094

**HAL Id: hal-04892094**

**<https://hal.science/hal-04892094v1>**

Submitted on 17 Jan 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Phylogeography of horseshoe bat sarbecoviruses in Vietnam and neighbouring countries. Implications for the origins of SARS-CoV and SARS-CoV-2



Alexandre Hassanin<sup>1</sup>  | Vuong Tan Tu<sup>2</sup> | Tamás Görföl<sup>3</sup> | Lam Quang Ngon<sup>2</sup> | Phu Van Pham<sup>2</sup> | Chu Thi Hang<sup>2</sup> | Tran Anh Tuan<sup>2</sup> | Mathieu Prot<sup>4</sup> | Etienne Simon-Lorière<sup>4</sup> | Gábor Kemenesi<sup>3</sup> | Gábor Endre Tóth<sup>3</sup> | Laurent Moulin<sup>5</sup> | Sébastien Wurtzer<sup>5</sup>

<sup>1</sup>Institut de Systématique, Évolution, Biodiversité (ISYEB), SU, MNHN, CNRS, EPHE, UA, Sorbonne Université, Paris, France

<sup>2</sup>Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology, Hanoi, Vietnam

<sup>3</sup>National Laboratory of Virology, Szentágotthai Research Centre, University of Pécs, Pécs, Hungary

<sup>4</sup>G5 Evolutionary Genomics of RNA Viruses, Institut Pasteur, Université Paris Cité, Paris, France

<sup>5</sup>R&D Laboratory, Direction Recherche, Développement et Qualité de l'Eau, Eau de Paris, Ivry-sur-Seine, France

## Correspondence

Alexandre Hassanin, Institut de Systématique, Évolution, Biodiversité (ISYEB), SU, MNHN, CNRS, EPHE, UA, Sorbonne Université, Paris, France.  
Email: [alexandre.hassanin@mnhn.fr](mailto:alexandre.hassanin@mnhn.fr)

## Funding information

Agence Nationale de la Recherche, Grant/Award Number: ANR-21-CO12-0002; Vietnam Academy of Science and Technology, Grant/Award Number: KHCBTD.02/22-24 and QTHU01.01/22-23; National Research, Development, and Innovation Fund of Hungary, Grant/Award Number: NKFIH FK137778 and RRF-2.3.1-21-2022-00010; János Bolyai Research Scholarship of the Hungarian Academy of Sciences, Grant/Award Number: BO/00825/21

**Handling Editor:** Camille Bonneaud

## Abstract

Previous studies on horseshoe bats (*Rhinolophus* spp.) have described many coronaviruses related to SARS-CoV (*SARSCoVr*) in China and only a few coronaviruses related to SARS-CoV-2 (*SARSCoV2r*) in Yunnan (southern China), Cambodia, Laos and Thailand. Here, we report the results of several field missions carried out in 2017, 2021 and 2022 across Vietnam during which 1218 horseshoe bats were sampled from 19 locations. Sarbecoviruses were detected in 11% of faecal RNA extracts, with much more positives among *Rhinolophus thomasi* (46%). We assembled 38 *Sarbecovirus* genomes, including 32 *SARSCoVr*, four *SARSCoV2r*, and two recombinants of *SARSCoVr* and *SARSCoV2r* (*RecSar*), one showing a Spike protein very similar to SARS-CoV-2. We detected a bat co-infected with four coronaviruses, including two sarbecoviruses. Our analyses revealed that *Sarbecovirus* genomes evolve in Vietnam under strong geographical and host constraints. First, we found evidence for a deep separation between viruses from northern Vietnam and those from central and southern Vietnam. Second, we detected only *SARSCoVr* in *Rhinolophus thomasi*, both *SARSCoVr* and *SARSCoV2r* in *Rhinolophus affinis*, and only *RecSar* in *Rhinolophus pusillus* captured close to the border with China. Third, the bias in favour of Uracil in synonymous third codon positions of *SARSCoVr* extracted from *R. thomasi* showed a negative correlation with latitudes. Our results also provided support for an emergence of SARS-CoV in horseshoe bats from northern Yunnan and emergence of SARS-CoV-2 in horseshoe

Alexandre Hassanin and Vuong Tan Tu contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Molecular Ecology* published by John Wiley & Sons Ltd.

bats from northern Indochina subtropical forests (southern Yunnan, northern Laos and north-western Vietnam).

#### KEYWORDS

Chiroptera, coronavirus genome, molecular evolution, phylogeny, recombination, Southeast Asia

## 1 | INTRODUCTION

The COVID-19 pandemic, with more than 700 million cases and 7 million deaths, was caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Wu et al., 2020). More than 4 years after the first cases detected in Wuhan (Hubei province, China) in December 2019, the origin of SARS-CoV-2 still remains unresolved. However, the three main hypotheses all imply a key role for horseshoe bats (Hassanin, 2022; Hassanin, Grandcolas, et al., 2021; Lam et al., 2020; Temmam et al., 2023; Worobey et al., 2022): (i) direct transmission from bats to humans, (ii) transmission via an intermediate animal such as small carnivores, pangolins, etc. or (iii) bat coronavirus escaped from a laboratory.

Horseshoe bats (family Rhinolophidae) are small or medium-sized insectivores characterised by their typical flat horseshoe-shaped nose-leaf which is employed in echolocation for navigating and hunting. There are 114 species, all included into a single genus (*Rhinolophus*), which are distributed in the Old World with much more species diversity in Southeast Asia (Simmons & Cirranello, 2020). Horseshoe bats are reservoir hosts for coronaviruses of the subgenus *Sarbecovirus* (family Coronaviridae, genus *Betacoronavirus*), that is, those causing Severe Acute Respiratory Syndrome (SARS) diseases, including SARS-CoV, involved in the 2002–2004 epidemics, and SARS-CoV-2, involved in the ongoing COVID-19 pandemic (Li et al., 2005; Zhou et al., 2020). Dozens of SARS-CoV related coronaviruses (SARSCoVr) have been detected in many horseshoe bat species collected in different provinces of China (Han et al., 2019; Lau et al., 2005; Wu et al., 2023). Recently, a few SARS-CoV-2 related coronaviruses (SARSCoV2r) were described from several horseshoe bat species: *Rhinolophus acuminatus* from East Thailand (RacCS203; Wacharapluesadee et al., 2021); *Rhinolophus affinis* sampled in the Yunnan province of China (RaTG13; Zhou et al., 2020); *Rhinolophus malayanus* sampled in Yunnan (RmYN02; Zhou et al., 2021) and northern Laos (RmBANAL52 and RmBANAL247; Temmam et al., 2022); *R. marshalli* from northern Laos (RmaBANAL236; Temmam et al., 2022); *Rhinolophus pusillus* sampled in Yunnan (RpYN06; Zhou et al., 2021) and northern Laos (RpBANAL103; Temmam et al., 2022); and *Rhinolophus shameli* from northern Cambodia (RshSTT200; Delaune et al., 2021).

The genome sequences of the two lineages SARSCoVr and SARSCoV2r differ by about 20% of nucleotide divergence. In addition, they show different synonymous nucleotide compositions (SNCs), suggesting their evolution in different *Rhinolophus* species assemblages and/or environmental conditions (Hassanin, 2022). Despite

these differences, several recombinant viruses between SARSCoVr and SARSCoV2r (named *RecSar*) have been detected in horseshoe bats sampled in Yunnan and Zhejiang provinces of China (Hassanin et al., 2022; Li et al., 2021). By comparing the ecological niches inferred for SARSCoVr and SARSCoV2r, it has been suggested that recombination of these two lineages can occur only in bats circulating in southern Yunnan, northern Laos and northern Vietnam (Hassanin, Tu, et al., 2021), and that the common ancestor of the two *RecSar* viruses identified in eastern China's Zhejiang province (RpZXC21 and RpVZC45) originated in Yunnan and then spread eastward to Zhejiang (Hassanin et al., 2022). At least two other *Sarbecovirus* lineages, also characterised by different SNCs and high nucleotide divergence (Hassanin, 2022), were previously detected in Asian horseshoe bats: the first was mainly found in *Rhinolophus steno* from Yunnan (lineage named *YunSar*) (Guo et al., 2021; Hassanin et al., 2022; Zhou et al., 2021) and the second in *Rhinolophus cornutus* from Japan (Murakami et al., 2022). Several more distant sarbecoviruses were also described in African and European horseshoe bats (Alkhovsky et al., 2022; Crook et al., 2021; Drexler et al., 2010; Tao & Tong, 2019). All recent studies have therefore confirmed that horseshoe bats are the main reservoir hosts of sarbecoviruses and that Sunda pangolins (*Manis javanica*), in which two different viruses were detected (Lam et al., 2020; Liu et al., 2019), are secondary hosts. Although pangolins have been considered as possible intermediate hosts between bats and humans to explain the origin of COVID-19 (Lam et al., 2020), several studies have recently concluded that SARS-CoV-2 emerged either directly from bats (Hassanin, 2022; Temmam et al., 2022) or from another intermediate host, such as the mammal genera present at the Huanan market in November and December 2019, that is, *Erinaceus*, *Nyctereutes*, *Rhizomys*, etc. (Liu et al., 2023; Worobey et al., 2022). A more extensive inventory of bat sarbecoviruses in Southeast Asia therefore appears crucial for identifying the key evolutionary steps leading to human epidemics. In 2021, predictions based on ecological niches suggested that SARSCoVr should also be found in horseshoe bats in northern Vietnam and northern Myanmar, and that SARSCoV2r should be present in horseshoe bats living in four areas (Hassanin et al., 2021b): (i) northern Laos and bordering regions in northern Myanmar, Yunnan and north-western Vietnam; (ii) southern Laos, south-western Vietnam and north-eastern Cambodia; (iii) the Cardamom Mountains in south-western Cambodia and the East region of Thailand; and (iv) the Dawna Range in central Thailand and south-eastern Myanmar. These ecological inferences were confirmed in northern Laos in 2022 with the description of four SARSCoV2r viruses in horseshoe bats collected in

Vientiane province (Temmam et al., 2022). In the present study, we aimed to verify whether *SARSCoVr* and *SARSCoV2r* are indeed present in Vietnam by exploring *Sarbecovirus* diversity in 1218 horseshoe bats captured in 2017, 2021 and 2022 at 19 geographical locations representing 12 provinces across the country (Figure S1).

## 2 | MATERIALS AND METHODS

### 2.1 | Bat sampling, RNA extraction and detection of sarbecoviruses

The bats were captured with mist-nets and harp-traps during several field surveys in Vietnam in 2017, 2021 and 2022. The species were identified using phenotypic traits and body measurements using field guides (Francis, 2019; Kruskop, 2013). For each bat, we collected a wing biopsy preserved in 95% ethanol for bat DNA taxonomy (the mitochondrial cytochrome c oxidase subunit 1 gene [mt-Co1] was sequenced and analysed as in Tu et al., 2017) and one or several faecal droppings stored in RNA*later* (Invitrogen, France) and conserved in liquid nitrogen for virology. All bats were released after sampling except a few injured animals that were euthanised and stored in ethanol for vouchers at IEBR.

For some bats, we obtained very small amount of faecal material. To ensure a good yield of RNA extraction, we have therefore opted for pools containing between two and five faecal samples from females or males of the same species collected in the same karst (sample codes are listed in Table 1). Based on previous studies (Trujillo et al., 2021; Wurtzer et al., 2021), potential pathogens were inactivated by pasteurisation (60°C, 1 h) just before RNA extraction. Then, they were lysed using a Fisherbrand Bead Mill 24 (Thermo Fisher Scientific, Waltham, USA) and extracted using QIASymphony PowerFecal Pro kit on a QIASymphony automated extractor (QIAGEN, Hilden, Germany) and a custom protocol allowing additional washing step before nucleic acid elution (Wurtzer et al., 2020).

The detection of sarbecoviruses was carried out on the E gene by real time reverse transcription PCR (RT-qPCR) following Corman et al. (2020) using a QIAgility (Qiagen, Hilden, Germany) with Fast virus 1-step Master mix 4x (Lifetechnologies, France). Thermal cycling was performed on a QuantStudio 5 instrument (Applied Biosystems, France) at 50°C for 5 min for reverse transcription, followed by 95°C for 20s and then 45 cycles of 95°C for 5s, 58°C for 40s.

### 2.2 | Next generation sequencing and genome assembly

All samples showing positive RT-qPCR for the E gene with a cycle threshold <30 were used to generate sequencing libraries. RNA extracts were quantified using the BioAnalyzer 2100 instrument with RNA 6000 Nano and Pico kits (Agilent Technologies, France).

Although a few samples were sequenced as in Delaune et al. (2021) (Table 1), most libraries were constructed and sequenced at the 'Institut du Cerveau et de la Moelle épinière' (Paris, France). Briefly, 11 microliters of RNA were used to deplete eukaryote/prokaryote ribosomal RNA and globine RNA before library preparation with stranded Total RNA Prep with Ribozero Plus kit (Illumina, USA) and sequencing using NovaSeq 6000 S1 Reagent kit (300cycles). The Illumina reads were mapped in Geneious Prime® 2020.0.3 using 20% of maximum mismatches per read on 63 reference genomes representing a large diversity of coronaviruses of the subfamily Orthocoronavirinae (Table S1). By this way, we were able to detect *Sarbecovirus* sequences and assemble several contigs. Then, some contigs were further identified by BLASTn search in NCBI ([blast.ncbi.nlm.nih.gov/](http://blast.ncbi.nlm.nih.gov/)) and the genomes were assembled using iterative mapping and de novo approaches.

To produce full *Sarbecovirus* genomes without missing data, PCRs were performed for a few samples of interest. Based on similarities with SARS-CoV-2, the ARTIC v3 protocol (DNA Pipelines R&D et al., 2020) was used for Rp22DB159 and Rp22DB167: 5 µL of RNA extract were used to generate cDNA using Superscript IV VILO Master Mix (Invitrogen) and amplicons were produced using the ARTIC v3.1 primer set. The two libraries were prepared using NEBNext Ultra II DNA Library Prep kit (New England Biolabs, France) and sequenced using MiSeq Reagent kit v2 (500 cycles) (Illumina, USA). To fill in the few missing genomic regions in Ra21CB8 and Rp22DB159, we designed specific primer pairs (Table S2) and cDNAs were used to produce amplicons using Q5 high fidelity Master Mix (New England Biolabs, France). A library was done with an equimolar pool of amplicons and MiSeq sequencing was conducted as detailed above.

The Rt17DN420 genome was amplified using several primer pairs (Table S2) and amplicons were end repaired and tailed with the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs, USA). Barcodes from the EXP-NBD196 kit (Oxford Nanopore Technologies, UK) were ligated to the end-prepped DNA with NEBNext Ultra II Ligation Module (New England Biolabs, USA). The pooled, barcoded samples were jointly cleaned with Ampure XP beads (Beckman Coulter, USA) and AMII sequencing adapters were ligated with NEBNext Quick Ligation Module. After quantification with Qubit dsDNA HS Assay Kit (Invitrogen, USA), the library was sequenced on a R9.4.1 (FLO-MIN106D) flow cell.

### 2.3 | Whole-genome alignment of sarbecoviruses

Complete *Sarbecovirus* genomes available in December 2022 in GenBank and GISAID databases were downloaded in Fasta format. Sequences with a large stretch of missing data were removed. Only a single sequence was retained for highly similar genomes (<0.1% of nucleotide divergence), such as SARS-CoV-2 (millions of sequences), pangolin viruses from Guangxi (5 sequences), etc. The details on the 111 selected genomes are provided in Table S3. They include all lineages previously described within the subgenus *Sarbecovirus* and 15



TABLE 1 *Sarbecovirus* genomes sequenced in this study.

Virus name	Date	Sample code(s)	Host	S	G	Total reads	Virus reads	L	GenBank	M
SARSCoVr										
Ra21CB8	12/2021	PO8+25+26	<i>R. affinis</i>	♂	1	145,979,440 <sup>No</sup> 12,071,222 <sup>P,Mi</sup>	31,150 NA	29,717	OR233291	0
Ra22DB107	06/2022	DB107	<i>R. affinis</i>	♀	3	147,368,172 <sup>No</sup>	12,188	29,703	OR233292	0
Ra22DB163	06/2022	DB163	<i>R. affinis</i>	♀	3	221,895,122 <sup>No</sup>	104,383	29,747	OR233325	0
Ra22DB173	06/2022	DB173	<i>R. affinis</i>	♀	3	193,766,498 <sup>No</sup>	20,270	29,737	OR233326	9
Ra22DB191	06/2022	DB191	<i>R. affinis</i>	♀	3	173,806,908 <sup>No</sup>	66,299	29,703	OR233304	0
Ra22QT27	11/2022	QT27+31+40+41	<i>R. affinis</i>	♂	13	180,980,390 <sup>No</sup>	4364	29,556	OR233315	11
Rt17DN420	06/2017	DN420	<i>R. thomasi</i>	♀	19	6,999,560 <sup>Mi</sup> 1,270,725 <sup>Na</sup>	911,000 NA	29,602	OR233295	0
Rt21LC39	11/2021	TP39	<i>R. thomasi</i>	♂	2	29,044,598 <sup>Ne</sup>	1,130,241	29,682	OR233308	0
Rt21LC92	11/2021	TP92	<i>R. thomasi</i>	♀	2	132,378,662 <sup>No</sup>	217,686	29,729	OR233314	0
Rt21LC192	11/2021	TP192+231+253	<i>R. thomasi</i>	♀	2	22,327,756 <sup>Ne</sup>	32,665	29,613	OR233305	1
Rt22CB395	02/2022	PO395+399	<i>R. thomasi</i>	♀	1	152,809,588 <sup>No</sup>	278,791	29,688	OR233293	0
Rt22DB31	06/2022	DB31	<i>R. thomasi</i>	♀	5	90,350,822 <sup>No</sup>	2,203,542	29,682	OR233294	0
Rt22DB38	06/2022	DB38	<i>R. thomasi</i>	♀	5	136,035,328 <sup>No</sup>	406,353	29,682	OR233310	0
Rt22LC371	03/2022	TP371+374+388	<i>R. thomasi</i>	♂	2	18,404,194 <sup>Ne</sup>	36,196	29,525	OR233306	16
Rt22LC376	03/2022	TP376+385	<i>R. thomasi</i>	♀	2	22,865,114 <sup>Ne</sup>	41,488	29,602	OR233307	10
Rt22LC378	03/2022	TP378+379+398	<i>R. thomasi</i>	♂	2	32,866,958 <sup>Ne</sup>	1,025,362	29,688	OR233296	0
Rt22QB8	11/2022	QB8+9+14	<i>R. thomasi</i>	♂	12	174,749,536 <sup>No</sup>	18,066	29,668	OR233299	0
Rt22QB78	11/2022	QB78	<i>R. thomasi</i>	♀	12	155,258,714 <sup>No</sup>	35,838	29,668	OR233298	0
Rt22QT36	11/2022	QT36+43+44	<i>R. thomasi</i>	♀	13	138,019,446 <sup>No</sup>	142,184	29,671	OR233318	0
Rt22QT46	11/2022	QT46+47+55+57	<i>R. thomasi</i>	♂	13	192,686,690 <sup>No</sup>	66,199	29,671	OR233319	0
Rt22QT48	11/2022	QT48+49+52	<i>R. thomasi</i>	♀	13	159,045,256 <sup>No</sup>	1,423,140	29,671	OR233320	0
Rt22QT53	11/2022	QT53+54+56	<i>R. thomasi</i>	♂	13	137,570,206 <sup>No</sup>	145,282	29,671	OR233321	0
Rt22QT124	11/2022	QT124+125+130+141	<i>R. thomasi</i>	♀	13	89,480,094 <sup>No</sup>	164,385	29,671	OR233300	0
Rt22QT161	11/2022	QT161+162+168+171	<i>R. thomasi</i>	♀	13	152,891,738 <sup>No</sup>	114,649	29,671	OR233316	0
Rt22QT178	11/2022	QT178+180+188+189	<i>R. thomasi</i>	♀	13	174,847,582 <sup>No</sup>	89,235	29,671	OR233317	0
Rt22SL9	07/2022	SL9+20	<i>R. thomasi</i>	♀	8	214,976,536 <sup>No</sup>	65,824	29,699	OR233303	0
Rt22SL58	07/2022	SL58+60	<i>R. thomasi</i>	♂	10	225,779,484 <sup>No</sup>	14,621,456	29,693	OR233312	0
Rt22SL67	07/2022	SL67	<i>R. thomasi</i>	♂	10	246,212,216 <sup>No</sup>	6437	29,695	OR233309	12
Rt22SL85	07/2022	SL85+86+88	<i>R. thomasi</i>	♀	10	199,789,662 <sup>No</sup>	60,245,942	29,693	OR233313	0
Rt22SL92	07/2022	SL92+95+97	<i>R. thomasi</i>	♀	10	220,472,398 <sup>No</sup>	2,309,119	29,693	OR233297	0
Rt22SL115	07/2022	SL115+116+119	<i>R. thomasi</i>	♀	10	240,657,066 <sup>No</sup>	9,488,696	29,693	OR233301	0
Rt22SL130	07/2022	SL130+131+132	<i>R. thomasi</i>	♂	10	174,282,568 <sup>No</sup>	1,273,016	29,693	OR233311	0
SARSCoV2r										
Ra22QT77	11/2022	QT77+81+94+97	<i>R. affinis</i>	♂	13	159,549,534 <sup>No</sup>	802,557	29,751	OR233324	0
Ra22QT106	11/2022	QT106+107+108+109	<i>R. affinis</i>	♀	13	145,804,146 <sup>No</sup>	28,507	29,751	OR233322	0
Ra22QT135	11/2022	QT135+138+140	<i>R. affinis</i>	♂	13	169,407,334 <sup>No</sup>	107,501	29,751	OR233323	0
Ra22QT137	11/2022	QT137+139	<i>R. affinis</i>	♀	13	160,939,448 <sup>No</sup>	70,872	29,751	OR233328	0
RecSar										
Rp22DB159	06/2022	DB159	<i>R. pusillus</i>	♀	3	125,632,212 <sup>No</sup> 13,461,204 <sup>A,Mi</sup> 12,071,222 <sup>P,Mi</sup>	6725 NA NA	29,823	OR233302	0
Rp22DB167	06/2022	DB167	<i>R. pusillus</i>	♀	3	167,424,614 <sup>No</sup> 14,541,622 <sup>A,Mi</sup>	694 NA	29,751	OR233327	11

Abbreviations: <sup>A</sup>, ARTIC v3 multiplex PCR; G, geographical locations (Figure S1); L, genome length; M, percentage of missing data; <sup>Mi</sup>, MiSeq Illumina system; <sup>Na</sup>, Nanopore system; <sup>Ne</sup>, NextSeq Illumina system; <sup>No</sup>, NovaSeq Illumina system; <sup>P</sup>, PCR amplifications (see main text for more details); S, sex.

new viruses from Vietnam. As indicated in Table S3, several virus names were slightly modified to be consistent with other names and to facilitate interpretations. The nucleotide sequences were aligned in Geneious Prime® 2020.0.3 with MAFFT 7.450 (Katoh & Standley, 2013) using default parameters. Then, the alignment was corrected manually on AliView 1.26 (Larsson, 2014) as explained in Hassanin and Rambaud (2023). In agreement with previous studies, several *Sarbecovirus* genomes extracted from horseshoe bats captured in Africa and Europe were used as outgroup (Hassanin et al., 2022; Zhou et al., 2021).

## 2.4 | Analysis of synonymous nucleotide composition

The SNC of *Sarbecovirus* genomes was studied using an alignment reduced to all protein genes (29,541 nt; ORF1ab, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N, and ORF10; ORF7a and ORF7b overlap by four nucleotides). Nucleotide frequencies were calculated as detailed in Hassanin et al. (2022) and the variables were summarised by a principal component (PC) analysis using the FactoMineR package (Lê et al., 2008) in R version 3.6.1 (from <https://www.R-project.org/>).

## 2.5 | Phylogenetic analyses

Maximum likelihood (ML) analyses were carried out on RAxML 8.2.11 (Stamatakis, 2014) using partitioned GTR+G models on three alignments: whole genome (30,103 nucleotides [nt]; partitions: non-coding regions and three codon positions of protein genes), RNA-dependent RNA polymerase (RdRp) gene (2796 nt; partitions: three codon positions) and Spike gene (3900 nt; partitions: three codon positions).

To examine the distribution of phylogenetic support along the genome alignment, the dataset of 111 sarbecoviruses and 30,103 nt was bootstrapped under the SWB program (Hassanin et al., 2022) following the procedure detailed in Hassanin and Rambaud (2023): five SWB analyses were conducted using different window sizes (400, 500, 600, 1000 and 2000 nt) moving in steps of 50 nt. In the SWB program, each window sub-dataset was automatically run in RAxML (Stamatakis, 2014) with a GTR+G model and 100 bootstrap replicates. The SWB output file contains the window bootstrap percentages (WBP) calculated for each window sub-datasets and for all the bipartitions (nodes) reconstructed during the SWB analysis. For example, the SWB analysis based on a window size of 400 nt generated 377,102,075 WBP values, that is, 633,785 bipartitions with WBP values (between 0% and 100%) for each of the 595 window sub-datasets (Table S4). Following Hassanin et al. (2022), the LFG program was used to convert the SWB<sub>400</sub> output into 595 bootstrap log files (lists of bootstrap bipartitions with their WBP values, from 1% to 100%), which were then used as inputs in SuperTRI v57 (Ropiquet et al., 2009) to construct an MRP (Matrix Representation

with Parsimony) file. The MRP<sub>400</sub> file is a Nexus matrix of 111 sarbecoviruses and 1,298,453 binary characters (each of them represents a bipartition found during the SWB<sub>400</sub> analysis); it contains an assumption block to assign the cumulated WBP to all characters. The MRP<sub>400</sub> file was then executed in PAUP 4.0a (Swofford, 2021) using 1000 bootstrap replicates of weighted parsimony (with cumulated WBP values used as weights) to construct a SuperTRI bootstrap 50% majority-rule consensus (SB<sub>400</sub>) tree, which is a supertree showing the phylogenetic relationships supported by the largest genomic fragments. In parallel, we also built SB<sub>500</sub>, SB<sub>600</sub>, SB<sub>1000</sub> and SB<sub>2000</sub> supertrees.

## 2.6 | Construction of genomic bootstrap (GB) barcodes

A GB barcode is a small image representing the nucleotide alignment in which the genomic regions containing a robust phylogenetic signal (GRPS) were coloured in green whereas the regions with no robust signal were coloured in red. The GB barcodes were constructed for several nodes of interest using the procedure detailed in Hassanin et al. (2022). The SWB results based on five different window sizes were analysed to identify the intervals of GRPS as previously explained in Hassanin and Rambaud (2023). Then, the CGB program (Hassanin & Rambaud, 2023) was used to draw the GB barcodes.

## 2.7 | Construction of coloured genomic bootstrap (CGB) barcodes

A phylogenetic CGB barcode is a small image representing the genome of a virus in which the different colours show the best robust phylogenetic signals, that is, the bipartitions containing the fewest number of closely related viruses, detected in different regions of the alignment. The phylogenetic CGB barcodes were constructed for several viruses of interest using the SWB results based on five window sizes and the procedure published in Hassanin and Rambaud (2023). Firstly, the BBC program (Hassanin et al., 2022) was used to select only SWB bipartitions showing one or more WBP values  $\geq 50\%$  (e.g. 633,785 SWB<sub>400</sub> bipartitions were reduced to 2452 BBC<sub>400</sub> bipartitions, Table S4). Then, only BBC bipartitions including the virus of interest were selected. For example, to reconstruct the CGB of SARS-CoV-2, we extracted 216 BBC<sub>400</sub> bipartitions, 210 BBC<sub>500</sub> bipartitions, etc. (Table S4). Then, the bipartitions were ranked in Excel in increasing order of size, from category '+1' (bipartitions including SARS-CoV-2 and one closely related virus) to category '+110' (the single bipartition including all viruses of our dataset). To make comparisons between WBPs calculated in the five SWB analyses, all WBP<sub>400</sub>, WBP<sub>500</sub>, WBP<sub>600</sub>, WBP<sub>1000</sub> and WBP<sub>2000</sub>  $\geq 70\%$  were highlighted in green and all WBPs between 50% and 70% were highlighted in yellow green using conditional formatting options in Microsoft® Excel. We performed the comparisons starting with bipartitions of the category '+1'. Due to past

events of genomic recombination, we found several bipartitions supporting conflicting phylogenetic relationships, for example, SARS-CoV-2 + RaTG13, SARS-CoV-2 + Rp22DB159, etc. For each bipartition +1, we identified the intervals of genomic regions containing a robust phylogenetic signal (GRPS) using previously published criteria (Hassanin & Rambaud, 2023). Then, we proceeded similarly by analysing other genomic fragments for bipartitions +2 (SARS-CoV-2 and two closely-related viruses), bipartitions +3 (SARS-CoV-2 and three closely-related viruses), etc. In this way, we were able to identify the closest virus(es) to SARS-CoV-2 in all regions of the genome alignment. In the last step, the intervals of GRPS (5' and 3' median positions in the whole-genome alignment) were written in a new CSV file for all bipartitions including SARS-CoV-2 in which one or more GRPS were identified as the best phylogenetic signals (i.e. bipartitions containing the fewest number of closely related viruses). A specific colour code was chosen for each bipartition and the file was used in the CGB program (Hassanin & Rambaud, 2023) to construct the phylogenetic CGB barcode of SARS-CoV-2.

The geographical CGB barcodes (showing the geographical origins of the closely related viruses) were derived from the original phylogenetic CGB barcodes by choosing different colours for viruses collected in distinct geographical areas.

Following Hassanin and Rambaud (2023), the CGB barcodes were used to calculate the phylogenetic contributions of viruses ( $C_T$ ), those of geographical areas ( $C_{TG}$ ), and those of host species ( $C_{TH}$ ). For instance, the  $C_{TG}$  of Yunnan in the CGB barcode of SARS-CoV-2 was calculated by summing the GRPS intervals in the whole-genome alignment in which SARS-CoV-2 was found to be closely related to one or several viruses from Yunnan and other geographical areas. Then, the sum was multiplied by 100 and divided by the total length of our alignment (30,103nt) to obtain the percentage contribution. The exclusive contributions ( $C_E$ ,  $C_{EG}$  and  $C_{EH}$ ) were calculated using only bipartitions showing exclusive ancestry with the virus(es) of interest. McNemar's Chi-squared tests were used to compare two geographical contributions (e.g. Laos vs. Yunnan) or two host contributions (e.g. *R. malayanus* vs. *R. pusillus*).

## 2.8 | SimPlot analyses

The Rp22DB159 genome was compared to a selection of nine *Sarbecovirus* genomes using SimPlot++ (Samson et al., 2022) with a TN93 distance model and a sliding window of 1000nt moving in steps of 100nt along the whole-genome alignment.

## 3 | RESULTS

### 3.1 | *Sarbecovirus* detection

Bats were captured in Vietnam during five field expeditions in June 2017 ( $n=31$ ), October–December 2021 ( $n=419$ ), February–March 2022 ( $n=361$ ), June 2022 ( $n=155$ ) and October–November 2022

( $n=252$ ). Thirteen *Rhinolophus* species were identified based on an integrative taxonomy approach combining morphological, acoustic and molecular data. More than 30 individuals were sampled for seven species: *R. affinis* ( $n=192$ ), *R. episcopus* ( $n=206$ ), *R. malayanus* ( $n=72$ ), *R. pearsonii* ( $n=89$ ), *R. pusillus* ( $n=252$ ), *R. siamensis* ( $n=111$ ) and *R. thomasi* ( $n=236$ ).

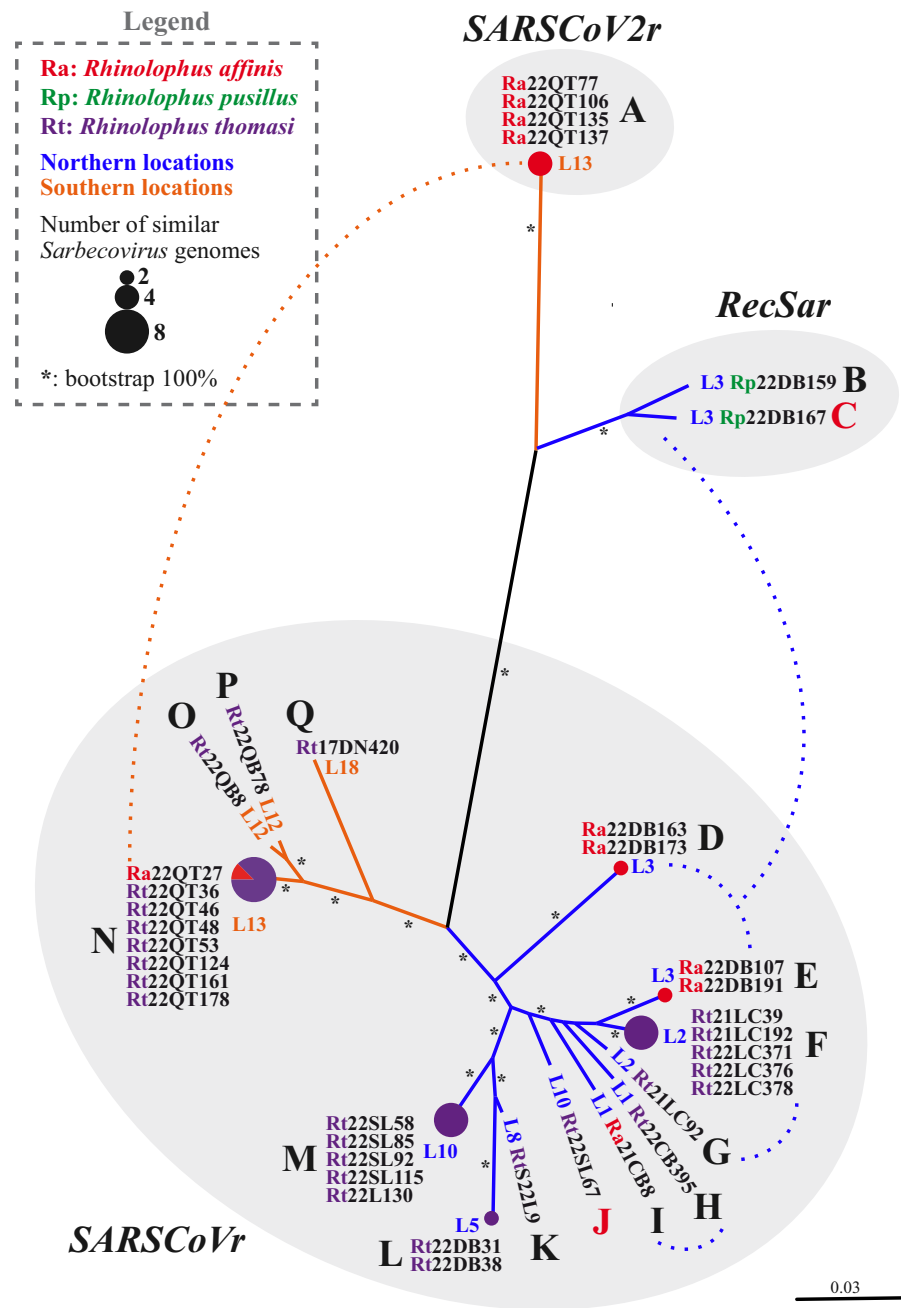
Faecal samples were investigated based on locations ( $n=19$ ), bat species ( $n=13$ ), sex (females versus males), and collection dates. In total, 547 pools were used for RNA extraction. We obtained 59 RT-qPCR amplifications of the E gene with a cycle threshold value  $<30$ , representing 11% of the samples. These included 46 samples of *R. thomasi* ( $n=99$ ; 46% of positive PCRs), 10 samples of *R. affinis* ( $n=99$ ; 10%), two samples of *R. pusillus* ( $n=99$ ; 2%) and one sample of *R. malayanus* ( $n=36$ ; 3%).

### 3.2 | New *Sarbecovirus* genomes

RNA sequencing generated thousands or millions of *Sarbecovirus* reads for most of the 45 sequenced libraries. However, some libraries did not provide enough data to perform genome assembly. In total, we were able to assemble 38 *Sarbecovirus* genomes (Table 1). The viruses were named using the following rules: the first two letters represent an abbreviation of the bat species; the two first numbers indicate the year of sampling; the two uppercase letters indicate the Vietnamese province (Figure S1) followed by a field number. Twenty-eight *Sarbecovirus* genomes were completely assembled with high read-depth mean coverage ( $>100\times$ ) after shotgun metagenomic sequencing. For 10 libraries, however, the coverage was too low and only a partial genome with several missing regions was assembled. Multiple amplifications by PCRs (using the ARTIC v3 set and/or primers specially designed for this study, Table S2) were conducted to complete four genomes of special interest. Assembling a full genome without missing data was possible for Ra21CB8, Rp22DB159 and Rt17DN420. In contrast, Rp22DB167 has posed problems because metagenomic sequencing revealed the existence of four coronaviruses in the faecal sample coming from a single bat (code: DB167), including one decacovirus, one rhinacovirus and two sarbecoviruses. The number of *Sarbecovirus* reads was too low compared to *Rhinacovirus* reads (694 vs. 51,354), thus preventing the extraction of separate contigs for the two sarbecoviruses. Co-infection with two sarbecoviruses was also detected after amplification with the ARTIC kit v3. BLAST searches were performed on divergent reads identified in several homologous regions: the hits revealed that the two sarbecoviruses were similar to Rp22DB159 and several viruses from Yunnan (RpPrC31, RpHN2021G, etc.) (Figure S2). Because of these difficulties, we choose to construct a 85% majority-rule consensus sequence for Rp22DB167, which was included in our first phylogenetic analysis (Figure 1), but not in subsequent analyses (to avoid possible artefacts due to this unreliable 'artificial' sequence).

The unrooted tree based on our 38 *Sarbecovirus* genomes (Figure 1) provided evidence for 17 viruses. Strains of the same virus showed highly similar genomes (nucleotide distances  $\leq 0.07\%$ );

**FIGURE 1** Distance tree of the 38 *Sarbecovirus* genomes assembled for this study. The tree was built in PAUP 4.0a (Swofford, 2021) with the Neighbour-Joining method using a genomic alignment of 29,979 nucleotides. The three divergent lineages of *Sarbecovirus* are highlighted in grey, including the 32 viruses closely related to SARS-CoV (SARSCoVr), the four viruses closely related to SARS-CoV-2 (SARSCoV2r), and the two recombinant viruses (*RecSar*) between SARSCoVr and SARSCoV2r lineages. The bold letters indicate the 17 sarbecoviruses identified in this study. For all of them, except C and J (written in red), we assembled one or several complete genomes with no missing data. Locations (L1, L2, L3, etc.) where bats were sampled are numbered as in the map provided in Figure S1 and viruses from the same location are linked by dashed lines.



they were generally collected in the same cave or karst network (e.g. locations N°2 and N°3) and some were sampled after a period of 4 months (e.g. Rt21LC39 in November 2021 and Rt22LC371 in March 2022). The 17 sarbecoviruses were separated by nucleotide distances ranging from 1.2% to 21%. Three divergent clusters can be recognised in Figure 1: SARSCoVr (32 genomes representing 14 viruses), SARSCoV2r (four genomes representing a single virus), and *RecSar* (two sequences representing three viruses; see above for Rp22DB167). Pairwise nucleotide distances are as follows: 10%–11% between SARSCoV2r and *RecSar*; 16%–18% between SARSCoVr and *RecSar*; and 20%–21% between SARSCoVr and SARSCoV2r.

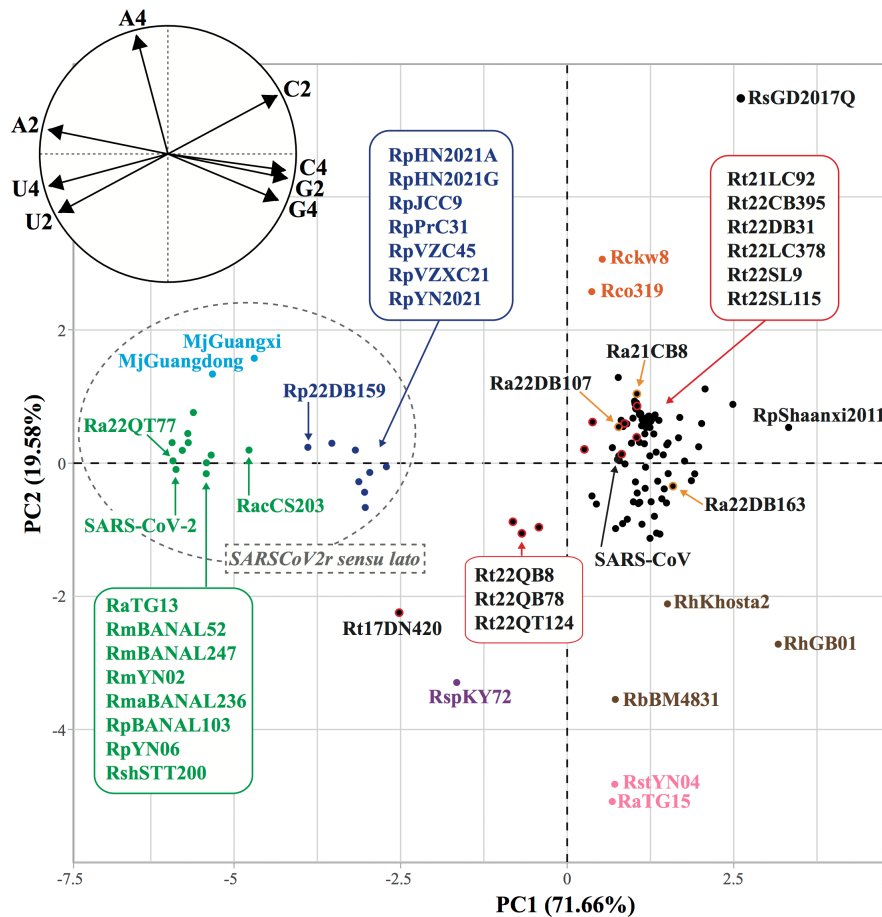
In some locations, we detected different viruses: in location N°13, we found one SARSCoV2r in four *R. affinis* samples and one SARSCoVr in seven *R. thomasi* samples and one *R. affinis* sample; in

flying bats captured close to the border with China (location N°3), we sequenced two *RecSar* in *R. pusillus* and two SARSCoVr in *R. affinis*.

### 3.3 | Synonymous nucleotide composition

Nucleotide frequencies were calculated at four-fold and two-fold degenerate third codon positions (abbreviations: 4X-P3 and 2X-P3 respectively) (Table S5) and these eight variables were summarised by a PC analysis based on the first two dimensions, which contribute 72% and 20% of the total variance respectively (Figure 2).

The sarbecoviruses from Vietnam are distributed in different SNC groups: (i) Ra22QT77 is close to SARS-CoV-2 and other bat SARSCoV2r; (ii) Rp22DB159 is close to *RecSar* collected in *R.*



**FIGURE 2** Variation in synonymous nucleotide composition (SNC) at third codon positions of *Sarbecovirus* genomes. The alignment of protein-coding genes (29,541 nt; 111 *Sarbecovirus* genomes) was used to calculate the frequency of the four bases either at four-fold degenerate third codon positions (A4, C4, G4 and U4 variables) or at two-fold degenerate third codon positions for either purines (A2 and G2 variables) or pyrimidines (C2 and U2 variables) (Table S5). The main graph represents the individual factor map obtained from the PC analysis based on the eight variables. The small circular graph at the top left represents the variables factor map. The eight SNC groups previously identified (Hassanin, 2022) are highlighted by different colours: (i) black for coronaviruses related to SARS-CoV (SARSCoVr), (ii) green for coronaviruses related to SARS-CoV-2 (SARSCoV2r), (iii) light blue for the two pangolin sarbecoviruses (*PangSar*), (iv) dark blue for bat viruses showing evidence of genomic recombination between SARSCoVr and SARSCoV2r (*RecSar*), (v) pink for the bat sarbecoviruses from the Yunnan province showing a divergent SNC (*YunSar*), (vi) orange for bat sarbecoviruses from Japan, (vii) brown for bat sarbecoviruses from Europe and (viii) purple for the bat sarbecovirus from Kenya. The 15 sarbecoviruses collected in Vietnam are distinguished as follows: the three SARSCoVr from *Rhinolophus affinis* are indicated by black circles with orange outline; the ten SARSCoVr from *Rhinolophus thomasi* are indicated by black circles with red outline; the positions of Ra22QT77 (SARSCoV2r) and Rp22DB159 (*RecSar*) are shown by an arrow.

*pusillus* from three provinces of China, that is, Hunan (RpHN2021A and RpHN2021G), Yunnan (RpJCC9, RpPrC31 and RpYN2021) and Zhejiang (RpVZC45 and RpVZXC21); (iii) all nine SARSCoVr from northern Vietnam are grouped with SARS-CoV and all other bat SARSCoVr except RsGD2017Q; (iv) the three SARSCoVr from central Vietnam (Rt22QB8, Rt22QB78 and Rt22QT124) are separated from the main SARSCoVr group; and (v) the sarbecovirus sampled farthest south in Vietnam, Rt17DN420, occupies a more isolated position. Some SNC groups showed specific features (Table S5): SARSCoV2r and pangolin viruses have the highest percentages of U at 4X-P3 and the highest percentages of A at 2X-P3; viruses of the clade named SARSCoV2r sensu lato (including SARSCoV2r, pangolin viruses and *RecSar*) (Hassanin et al., 2022) have the lowest percentages of C and G at 4X-P3; *YunSar* viruses (RaTG15 and RstYN04) have the highest

percentages of C and lowest percentages of A at 4X-P3; European sarbecoviruses have the highest percentages of G at 2X-P3.

In Figure 2, the clustering of Vietnamese SARSCoVr extracted from *R. thomasi* in three geographical groups (north, central and south) suggests a negative correlation between the percentages of U nucleotides and the latitudinal distribution of bat populations. By reporting the percentages of pyrimidines at synonymous sites of SARSCoVr genomes against the latitudes of the sampling sites, we found a negative correlation between the latitudes and percentages of U at both 2X-P3 and 4X-P3 ( $R^2=0.982$  and  $0.918$  respectively; Figure S3) and a positive correlation between the latitudes and percentages of C at both 2X-P3 and 4X-P3 ( $R^2=0.983$  and  $0.839$ , respectively; Figure S3). These results therefore indicate that SARSCoVr genomes evolve in *R. thomasi* of Vietnam under strong



latitudinal pressure, the viruses having more C-to-U mutations at lower latitudes than higher latitudes. By analysing the dinucleotide composition at second and third codon positions of SARSCoVr genomes isolated from *R. thomasi*, we showed that the bias toward C-to-U mutation concerns all types of dinucleotides (AC<sub>2</sub>-to-AU<sub>2</sub>, CC<sub>2</sub>-to-CU<sub>2</sub>, GC<sub>2</sub>-to-GU<sub>2</sub> and UC<sub>2</sub>-to-UU<sub>2</sub>) (Figure S4).

### 3.4 | Phylogenetic trees of sarbecoviruses

All traditional phylogenetic methods, such as Bayesian inference and ML estimation, assume that all regions of the sequences share a common underlying evolutionary history (Posada et al., 2002). This basic assumption is violated in the case of bat sarbecoviruses because different small genomic regions can support strikingly discordant phylogenetic relationships due to multiple past events of recombination (Boni et al., 2020; Forni et al., 2017; Hassanin & Rambaud, 2023). This is well illustrated in our study as the ML trees reconstructed from the whole-genome sequences, RdRp and Spike genes showed high levels of topological incongruence (Figures 3 and 4). For instance, SARS-CoV-2 was found related to different viruses if we consider bootstrap percentages (BP)  $\geq$  70%: RaTG13 + RmBANAL52 + RmaBANAL236 + RpBANAL103 (BP = 71%) based on the genome alignment; all SARSCoV2r except RacCS203 (BP = 88%) based on the RdRp gene; MjGuangxi + RaTG13 + Rp22DB159 (BP = 98%) based on the Spike gene.

To better interpret conflicting phylogenetic signals, we reported the GB barcodes of all SARS-CoV and SARS-CoV-2 nodes supported by BP  $\geq$  70% in the phylogenetic trees of Figures 3 and 4. Importantly, several of these nodes were found not supported by any genomic fragment, that is, GB barcodes without green region (GRPS = 0%), such as the sister-group relationship between SARS-CoV and RsYN2020C in the genome tree of Figure 3 (BP = 100%). Only two SARS-CoV and SARS-CoV-2 nodes in the genome tree were found to be supported by large regions representing more than one third of the alignment (GRPS > 33%): SARSCoVr (GRPS = 44%), which includes most sarbecoviruses from China and Vietnam; and a large clade uniting all Asian sarbecoviruses except SARSCoVr (GRPS = 42%).

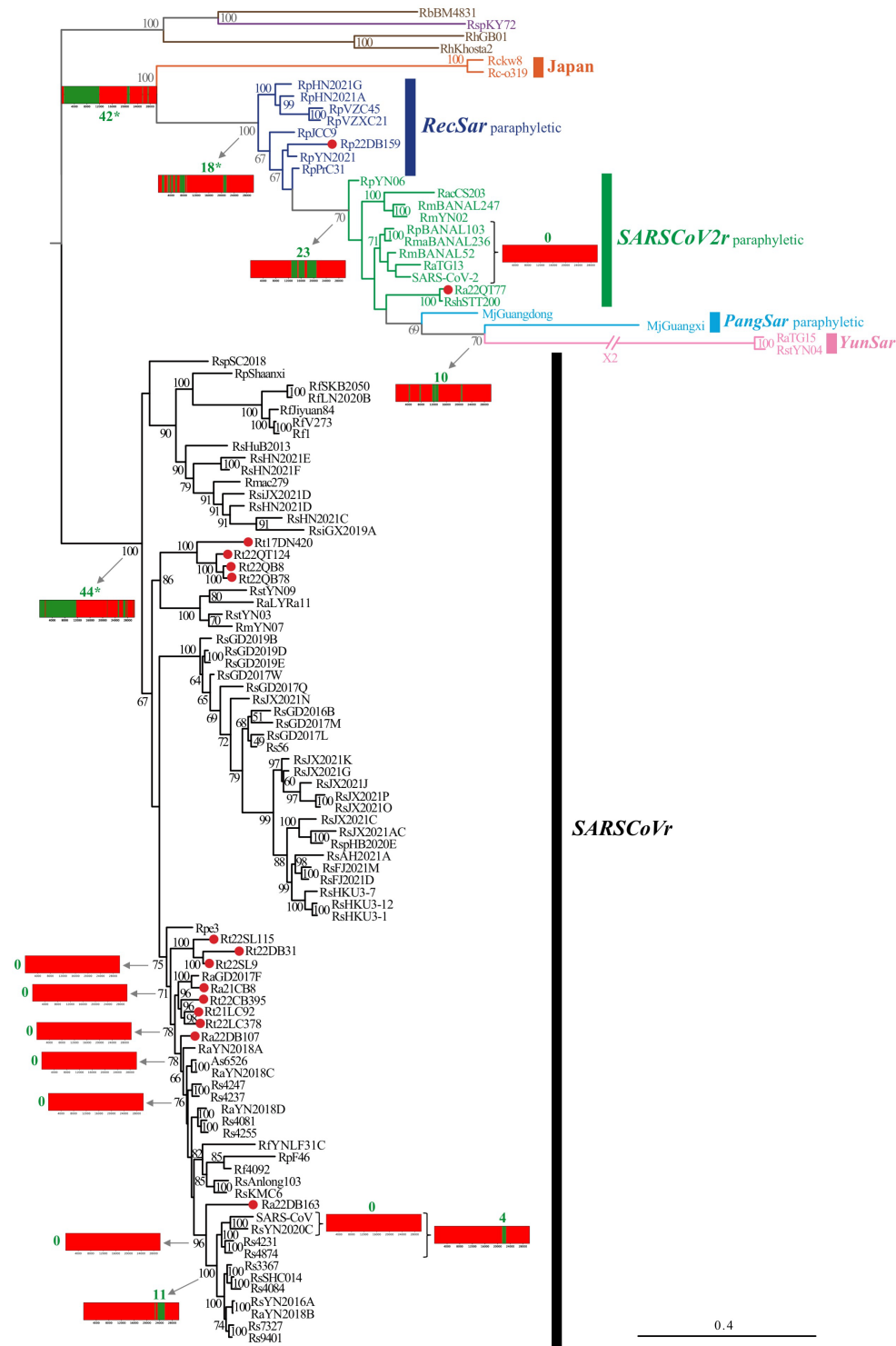
To better understand the evolution of sarbecoviruses, we also performed phylogenetic analyses based on SWB and SuperTRI methods (Table S4). This approach was designed to reveal the relationships supported by the largest proportions of the genome (Hassanin et al., 2022, 2024; Hassanin & Rambaud, 2023). Since the five SuperTRI bootstrap (SB) trees reconstructed from the SWB analyses based on different window sizes showed high node congruence (Figures S5–S9), we decided to show in Figure 5 a strict consensus tree in which all topological differences between the five SB trees were collapsed. A few nodes of the SB consensus tree showed a green or quasi-green ( $\geq$  98%) GB barcode, indicating that GRPS cover between 100% and 98% of the genome alignment. All these highly supported nodes concern sister-group relationships, that is, between the two Japanese sarbecoviruses, between the

two *YunSar* viruses, and between Ra22QT77 and RshSTT200. As expected, all these nodes were supported by BP = 100% in the ML trees based on whole-genome sequences, RdRp and Spike genes (Figures 3 and 4). Importantly, the SB consensus tree of Figure 5 showed more nodes supported by large genomic regions than the genome tree of Figure 3. Among them, there are three deep nodes: (i) SARSCoV2r (GRPS = 19%), which includes SARS-CoV-2 and 10 bat sarbecoviruses; (ii) a large clade, named SARSCoV2r sensu lato (Hassanin et al., 2022) (GRPS = 32%), which contains SARSCoV2r, the two pangolin viruses and the eight *RecSar*; and (iii) SARSCoV2r sensu lato excluding MjGuangxi (GRPS = 41%). Interestingly, the phylogenetic signal supporting SARSCoV2r was restricted to the central region of the genome (GB barcode in Figure 5), that is, the region containing the RdRp gene. In agreement with that, SARSCoV2r was found monophyletic in the RdRp tree (BP = 83%; Figure 4a). By contrast, the phylogenetic signals supporting SARSCoV2r sensu lato and SARSCoV2r sensu lato excluding MjGuangxi were found in both 5' and 3' regions of the genome (GB barcodes in Figure 5). By analysing all bipartitions involving SARS-CoV and SARS-CoV-2, we were able to detect robust signals for alternative relationships in which SARSCoVr was found as the sister-group of *RecSar* (GRPS = 12%; positions 15,251–20,050) and SARSCoV2r was found grouped with pangolin viruses and *YunSar* (GRPS = 23%; positions 12,901–20,900) (GB barcodes in Figure 5). These results confirmed that *RecSar* genomes have emerged through past recombination of SARSCoVr and SARSCoV2r: their central region is SARSCoVr-like, whereas their 5' and 3' regions are SARSCoV2r-like.

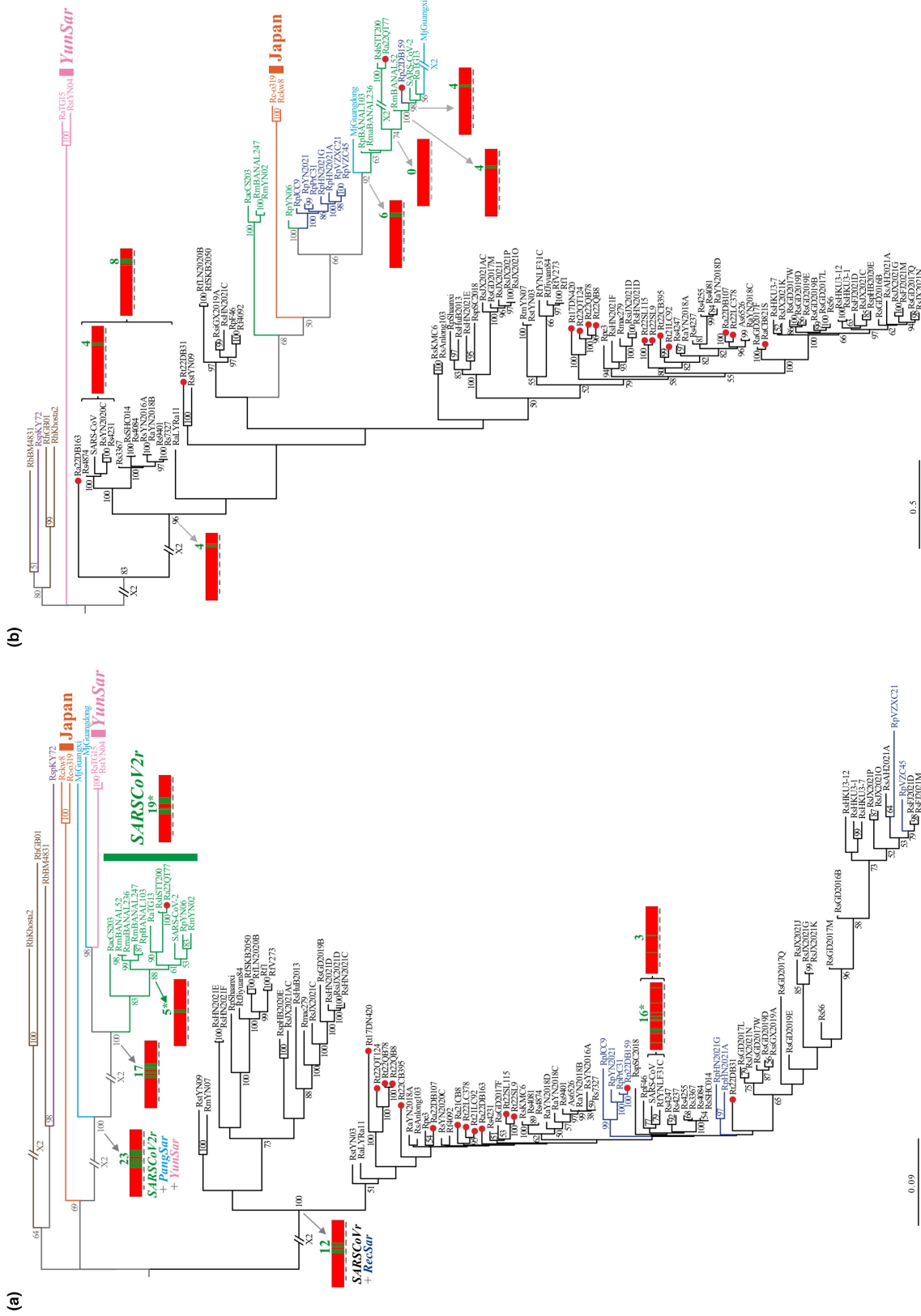
### 3.5 | Multiple phylogenetic signals detected in the spike gene of SARS-CoV-2

The Spike tree in Figure 4b provided BP = 98% for the node uniting SARS-CoV-2, Rp22DB159, RaTG13, and MjGuangxi, and BP = 100% for their grouping with RmBANAL52. However, the phylogenetic CGB barcodes of the Spike gene (Figure 6) showed a more complex situation, as different phylogenetic signals were detected along the gene. In the 5' region, we indeed found a genomic fragment supporting the grouping of SARS-CoV-2 with Rp22DB159, RaTG13, RmBANAL52 and MjGuangxi (bipartition D; positions 1–1099) and a smaller fragment in which RmBANAL52 was excluded from the group (bipartition C; positions 900–1049). However, the best signals provided support for exclusive ancestry of SARS-CoV-2 with either RaTG13 (bipartition 1; pos. 1–549) or Rp22DB159 (bipartition 2; pos. 700–999). The pangolin virus MjGuangxi was excluded from most of the best bipartitions identified in central and 3' regions of the gene. In the central region, the best signals showed evidence for the group uniting SARS-CoV-2, Rp22DB159, RmBANAL52, RmaBANAL236, and RpBANAL103 (bipartition E; pos. 1200–1599) or a larger group also containing RaTG13 (bipartition F; pos. 1750–2049). The 3'-end region provided support for different relationships: SARS-CoV-2 was found related to either RmBANAL247 + RmYN02 (bipartition B; pos. 3300–3749) or RshSTT200 (bipartition 3; pos. 3800–3900).

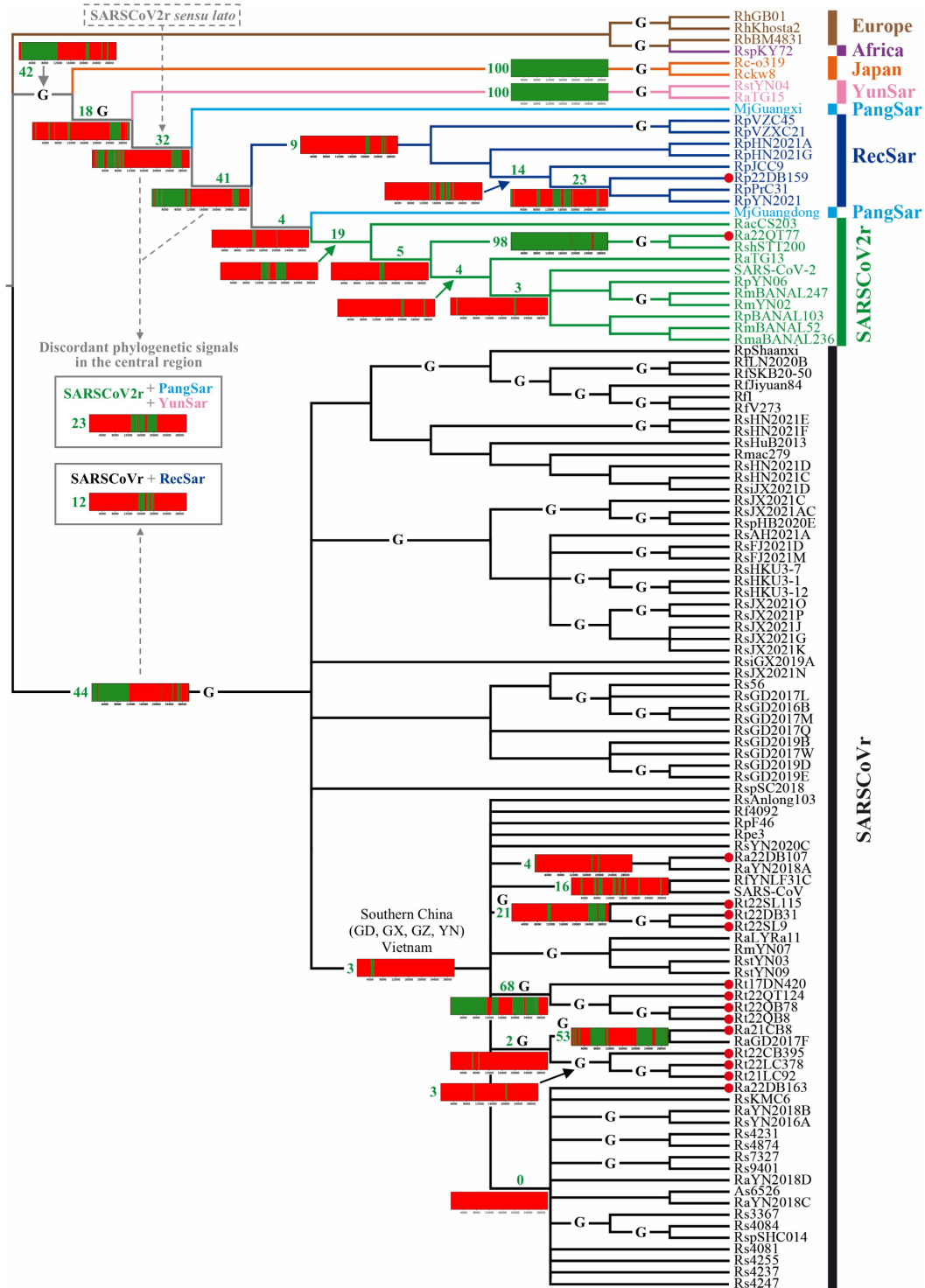




**FIGURE 3** Maximum Likelihood tree based on whole-genome sequences of sarbecoviruses (30,103 nucleotides). The tree was reconstructed with RAxML using different GTR+G models for the four partitions corresponding to the three codon positions and non-coding regions. The 15 viruses from Vietnam are highlighted with a red circle. The colours of sarbecoviruses indicate to which group of synonymous nucleotide composition they belong (written at the right; see Figure 2). Sarbecoviruses found in Europe and Africa were used as outgroup. Bootstrap percentages (BP) higher than 50% are indicated at the nodes. The genomic bootstrap (GB) barcodes were constructed for all SARS-CoV and SARS-CoV-2 nodes supported by BP  $\geq$  70%. The genomic regions containing a robust phylogenetic signal (GRPS) are highlighted in green in GB barcodes, whereas the regions with no (robust) phylogenetic signal are coloured in red. The intervals of GRPS were used to calculate the percentage of the genome alignment supporting phylogenetic relationships (value in green for each GB barcode). Values marked with an asterisk indicate nodes also found to be monophyletic in the supertree of Figure 5.



**FIGURE 4** Maximum Likelihood tree based on the RdRp (a) and Spike (b) genes of sarbecoviruses (2796 and 3900 nucleotides respectively). The trees were reconstructed with RAxML using different GTR+G models for the three codon positions. The 15 viruses from Vietnam are highlighted with a red circle. The colours of sarbecoviruses indicate to which group of synonymous nucleotide composition they belong. Bootstrap percentages (BP) higher than 50% are indicated at the nodes. The genomic regions containing a robust phylogenetic signal (GRPS) are highlighted in green in GB barcodes, whereas the regions with no (robust) phylogenetic signal are coloured in red. The intervals of the genome alignment supporting phylogenetic relationships (value in green for each GB barcode).



**FIGURE 5** Consensus tree from SWB and SuperTRI analyses based on the whole-genome alignment of sarbecoviruses. The alignment of 30,103 nucleotides was analysed using the SWB program (Hassanin et al., 2022) and five different window sizes (400, 500, 600, 1000 or 2000nt). Then, the five SWB output files were transformed into five MRP files (with LFG and SuperTRI programs; Hassanin et al., 2022; Ropiquet et al., 2009) which were then executed in PAUP 4.0a (Swofford, 2021) to construct five SuperTRI bootstrap 50% majority-rule consensus (SB) trees using weighted parsimony and 1000 bootstrap replicates. The shown tree is a strict consensus of the five SB trees based on different window sizes. The 15 viruses from Vietnam are highlighted with a red circle. The colours of sarbecoviruses indicate to which group of synonymous nucleotide composition they belong. The genomic bootstrap barcodes were constructed for all nodes involving SARS-CoV, SARS-CoV-2 and bat sarbecoviruses from Vietnam. The genomic regions containing a robust phylogenetic signal (GRPs) are highlighted in green in GB barcodes, whereas the regions with no (robust) phylogenetic signal are coloured in red. The intervals of GRPs were used to calculate the percentage of the genome alignment supporting phylogenetic relationships (value in green for each GB barcode). The 'G' letter was used to highlight nodes also found to be monophyletic in the genome tree of Figure 3.

The highest phylogenetic contributions to SARS-CoV-2 in the Spike gene (Figure 6) were found for Rp22DB159 ( $C_{T/E} = 72\%/8\%$ ), RaTG13 ( $C_{T/E} = 68\%/14\%$ ), and three viruses from Laos (RmBANAL52, RmaBANAL236 and RpBANAL103;  $C_{T/E} = 59\% - 58\%/0\%$ ).

In Figure 7, we showed the protein alignment of the receptor-binding domain (RBD) followed by the region of the furin cleavage motif in SARS-CoV-2. The alignment includes the nine animal viruses showing the highest RBD similarity with SARS-CoV-2: Rp22DB159 and RmBANAL52 (6 amino acid differences), MjGuangdong and RpBANAL103 (7 differences), RmaBANAL236 (8 differences), RaTG13 (22 differences), MjGuangxi (31 differences), RshSTT200 (35 differences) and Ra22QT77 (36 differences). The sequences of the four other SARSCoV2r (RacCS203, RmBANAL247, RmYN02 and RpYN06) and SARS-CoV were included in the alignment for comparison; they showed between 59 and 63 differences with SARS-CoV-2, including a large deletion of 14 amino acids in the receptor-binding motif (RBM) that interacts directly with ACE2 (Lan et al., 2020) (69 sites highlighted in yellow in Figure 7). In the RBM, SARS-CoV-2 was highly divergent from RaTG13 (17 differences; 25%) but showed a single difference with RmBANAL52, RpBANAL103, and MjGuangdong, and two differences with Rp22DB159 and RmaBANAL236.

In the RBD region, our analysis of CGB barcodes (Figure 6) revealed several robust phylogenetic signals, but the best signal found in the RBM region supported the grouping of SARS-CoV-2 with Rp22DB159, RmBANAL52, RmaBANAL236 and RpBANAL103 (bipartition E, 400nt). A phylogenetic analysis based on this fragment of 400nt (Figure S10) confirmed the monophyly of this group (BP=97%) and it appeared closely related to MjGuangdong (BP=72%), which is consistent with the high amino acid conservation observed in Figure 7. In contrast, all other *RecSar* and four SARSCoV2r (RacCS203, RmBANAL247, RmYN02 and RpYN06) were found more closely related to SARSCoVr (BP=100%).

Our alignment showed that the furin cleavage motif S1/S2 (amino acid sequence: RRAR) is a unique feature of SARS-CoV-2, formed by the insertion of the 12-bp motif 'CCT-CGG-CGG-GCA' (Figure 7). The furin motif is located into a large genomic region in which the best phylogenetic signal involves the grouping of SARS-CoV-2 with Rp22DB159 and six SARSCoV2r (Ra22QT77, RaTG13, RmBANAL52, RmaBANAL236, RpBANAL103 and RshSTT200; H bipartition, 500nt; Figure 6). In this region, there is no evidence for recent recombination as we did not detect discordant phylogenetic relationships. However, our method cannot detect recombination events involving genomic fragments smaller than 50nt (step parameter used during SWB analyses) and the lack of robust phylogenetic signal in this region could be also explained by unsampled viruses closely related to SARS-CoV-2.

### 3.6 | Virus and host contributions to Rp22DB159

The genome of Rp22DB159, a virus collected in north-western Vietnam close to the border with China, was found to have a Spike

gene very similar to that of SARS-CoV-2. Its phylogenetic CGB barcode and SimPlot analysis are shown in Figure 8 to better understand its multiple phylogenetic affinities. First, 73% of its genome share exclusive ancestry with five viruses, including three *RecSar*, that is, RpPrC31 ( $C_E = 32.2\%$ ), RpYN2021 (32.1%) and RpJCC9 (6.5%), and two SARSCoV2, that is, SARS-CoV-2 (1%) and RaTG13 (0.8%). Coloured arrows numbered 1–12 in Figure 8 were used to highlight that all genomic regions supporting exclusive ancestry are those showing the highest similarity to one of these five viruses. Second, the SimPlot graph confirmed that Rp22DB159 belongs to the *RecSar* lineage, as previously inferred based on SNC results and GB barcodes (Figures 2 and 5): Rp22DB159 showed high similarity with the three other *RecSar* viruses included in the analysis (RpPrC31, RpYN2021 and RpJCC9); SARS-CoV was found highly divergent from Rp22DB159 in 5' and 3' regions but very similar to Rp22DB159 and the three other *RecSar* viruses in the central region (around positions 14,000–20,000). Third, both CGB barcode and SimPlot results showed that the Spike gene (positions 21,652–25,551) of Rp22DB159 is closely related to some SARSCoV2r, including SARS-CoV-2, RaTG13, BANAL viruses, etc.

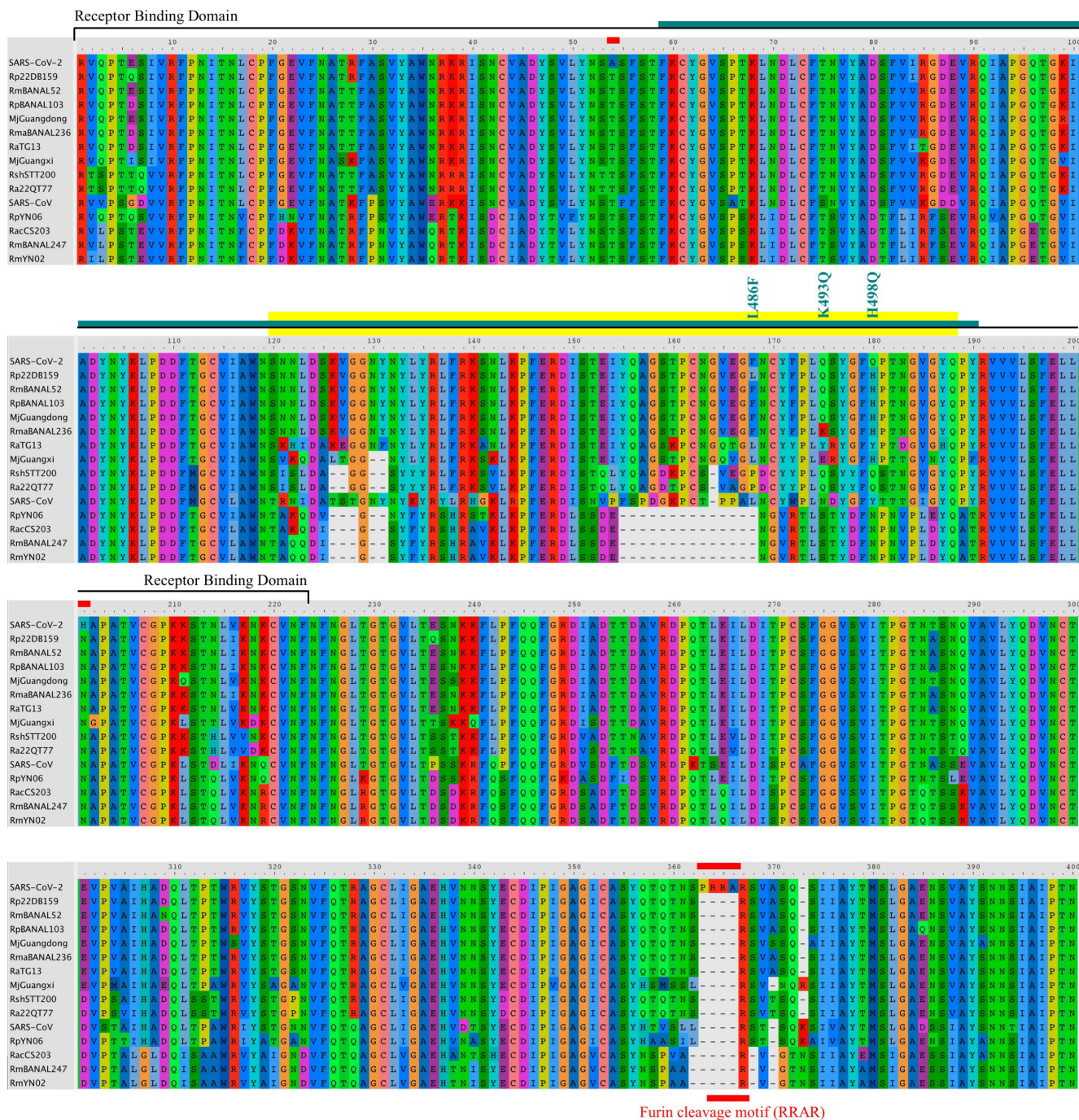
The species *R. pusillus* clearly showed the highest phylogenetic contribution to Rp22DB159 ( $C_{TH/EH} = 96\%/77\%$ ;  $C_{TH/EH} \leq 19\%/3\%$  for other species) indicating that its recent evolution has occurred mainly in this host.

### 3.7 | Phylogeographic analysis of bat SARSCoVr

Sarbecoviruses found in horseshoe bats of Vietnam showed geographical structuring in the distance tree of Figure 1. Particularly relevant is the dichotomy separating SARSCoVr from northern provinces (in blue) and those from southern provinces (in orange). Although southern Vietnamese SARSCoVr were recovered monophyletic in all phylogenetic analyses (Figures 3–5), northern Vietnamese SARSCoVr were found either paraphyletic or polyphyletic, suggesting a more complex pattern of phylogeographic evolution. To better interpret their geographical affinities, we constructed the geographical CGB barcodes of all bat SARSCoVr from Vietnam. The results showed a colour separation between viruses from northern and southern provinces (respectively surrounded by blue and orange dashed lines in Figure 9). In addition, a few viruses from northern Vietnam were found to exhibit stronger affinities with viruses from China. This is obvious for Ra22DB163, which was sampled in north-western Vietnam (location N°3): all regions of its genome can be related to viruses sampled in Yunnan ( $C_{TG} = 100\%$ ) and 37% of its genome involves exclusive ancestry with Yunnan viruses (vs.  $C_{EG} = 0\%$  for the three other geographical areas). Another example is Ra21CB8, which was collected in north-eastern Vietnam (location N°1): 97% of its genome can be related to viruses found in China except Yunnan (vs.  $C_{TG} = 46\%$  and 41% for Yunnan and northern Vietnam respectively) and 53% of its genome showed exclusive ancestry with viruses sampled in China except Yunnan (vs.  $C_{EG} = 1\%$  and 0% for Yunnan and northern Vietnam respectively).

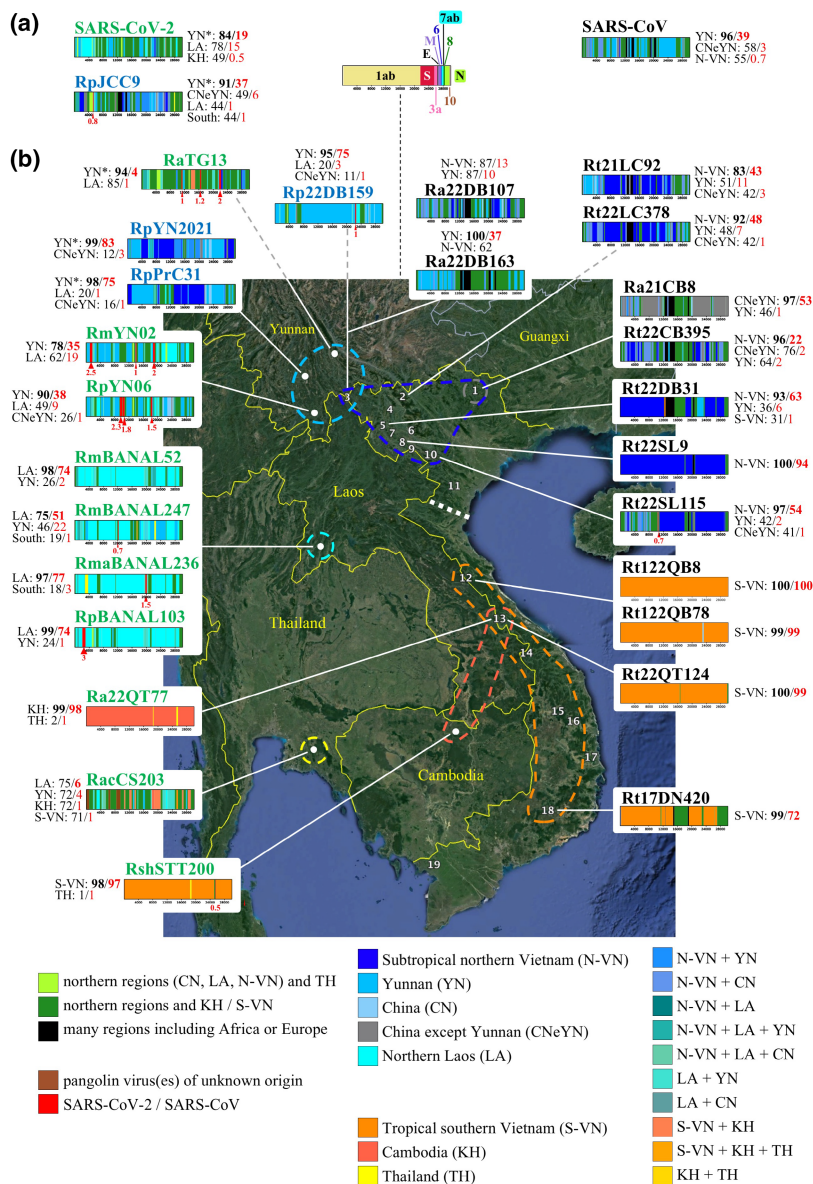












**FIGURE 9** Geographical coloured genomic bootstrap (CGB) barcodes reconstructed for sarbecoviruses of unknown origin (a) and those found in georeferenced horseshoe bats of Southeast Asia and Yunnan (b). In Figure a are shown the geographical CGB barcodes reconstructed for SARS-CoV, SARS-CoV-2 and one *RecSar* virus of unavailable origin in Yunnan (RpJCC9). We also showed the genome structure of sarbecoviruses with the following abbreviations: E: envelope gene; M: membrane gene; N: nucleocapsid gene; S: spike gene; 1ab: ORF (Open Reading Frame) 1ab; 3a: ORF3a; 6: ORF6; 7ab: ORF7ab; 8: ORF8; 10: ORF10. In Figure b are shown the geographical CGB barcodes reconstructed for sarbecoviruses found in georeferenced horseshoe bats: on the left for *SARSCoV2r* and *RecSar*; on the right for *SARSCoVr*. The best bipartitions found in the phylogenetic CGB barcodes were coded using different colours for geographical areas listed at bottom. For instance, blue was used for all intervals in which the best signals included only viruses from northern Vietnam (N-VN). Then, the contribution of each main geographical areas (N-VN; S-VN: southern Vietnam; YN: Yunnan; CNeYN: China except Yunnan; etc.) was estimated using two indicators: (i) the total contribution ( $C_{TG}$ ), which corresponds to the percentage of the genome alignment for which the best bipartitions involved at least one sarbecovirus from the geographical area (written in black above barcodes); and (ii) the exclusive contribution ( $C_{EG}$ ), which corresponds to the proportion of exclusive ancestry, that is, the percentage of the genome alignment for which the best bipartitions involved only sarbecovirus(es) from the geographical area (given in red above barcodes). The indicators found statistically significant are indicated in bold. When YN\* was used instead of YN, the calculations were made assuming that Yunnan and location N°3 in north-western Vietnam (in which Rp22DB159 was sampled) belong to the same geographical area (more details available in the main text). The  $C_{TG} < 50\%$  were not shown except those involving exclusive ancestry ( $C_{EG} > 0$ ). Red arrows below barcodes indicate the percentages of exclusive ancestry with SARS-CoV-2 (at the left) and SARS-CoV (at the right). The white dotted line shows the climate separation between subtropical northern Vietnam and tropical southern Vietnam based on the Köppen–Geiger climate classification (Peel et al., 2007). Map from Google Earth Pro (version 7.3.3.7786) US Dept of State Geographer © 2020 Google—Image Landsat/Copernicus—Data SIO, NOAA, U.S. Navy, NGA, GEBCO.

### 3.8 | Phylogeographic analysis of bat SARSCoV2r and RecSar

The geographical CGB barcodes reconstructed for SARS-CoV-2, the 10 bat SARSCoV2r and four related bat RecSar are shown in Figure 9. For Rp22DB159, which was sampled in north-western Vietnam (location N°3), close to the border with Yunnan, we found strong affinities with Yunnan viruses ( $C_{TG/EG} = 95\%/75\%$ ). We therefore conclude that Rp22DB159 from north-western Vietnam and RecSar viruses from Yunnan evolved in the same ecological region. This is also supported by the proximity of location N°3 in Vietnam and the three RecSar locations in southern Yunnan (between 110 and 140km). In agreement with that, the mitochondrial haplotype sequenced for DB159 (mt-Co1 gene) was also detected by BLAST search in *R. pusillus* found in two Yunnan locations (Table S6) and the ecological affinities between location N°3 in north-western Vietnam and Yunnan was also supported by the geographical CGB barcode of Ra22DB163, a SARSCoVr collected in location N°3, which showed higher contributions from Yunnan ( $C_{TG/EG} = 100\%/37\%$ ) than from other locations in northern Vietnam ( $C_{TG/EG} = 62\%/0\%$ ).

The geographical CGB barcode of Ra22QT77 from central Vietnam showed that it is closely related to RshSTT200 from Cambodia in most parts of its genome ( $C_{TE} = 99\%/98\%$ ). In a small region, however, its genome was found to share exclusive ancestry with RacCS203 from Thailand ( $C_E = 1\%$ ). The geographical affinities of RacCS203 remain elusive based on the available data, as very similar  $C_{TG}$  (75%–71%) were calculated for Cambodia, Laos, southern Vietnam and Yunnan. In contrast, the four BANAL viruses from Laos showed strong affinities with each other ( $C_{TG/EG} = 75\%–99\%/51\%–77\%$ ), indicating that their recent evolution has taken place more locally rather than through imports from other countries. The three SARSCoV2r from Yunnan contrasted less, although the contributions were always more important for Yunnan than for Laos: RaTG13 ( $C_{TG/EG} = 94\%/4\%$  vs. 85%/1%), RmYN02 ( $C_{TG/EG} = 78\%/35\%$  vs. 62%/19%) and RpYN06 ( $C_{TG/EG} = 90\%/38\%$  vs. 49%/9%). In addition, the geographical CGB barcode of RaTG13 appeared to contain many regions coloured in green, indicating a lack of affinities with other sarbecoviruses. A similar but lesser green pattern was also found for RacCS203 and SARS-CoV-2. These elements indicate that many genomic regions of these viruses are divergent from all sarbecoviruses currently known. For SARS-CoV-2 however, the two geographical indicators provided very high support for an origin in the zone covering northern Laos, southern Yunnan and north-western Vietnam ( $C_{TG/EG} = 100\%/46\%$ ) and more support (significance level: .05 but not .01) for an origin in Yunnan ( $C_{TG/EG} = 84\%/19\%$ ) than Laos ( $C_{TG/EG} = 78\%/15\%$ ).

## 4 | DISCUSSION

### 4.1 | Divergent evolution of bat sarbecoviruses in subtropical northern Vietnam and tropical southern Vietnam

Although several bat coronaviruses have already been described in Vietnam, none of these were found to belong to the subgenus

*Sarbecovirus* (Berto et al., 2018; Latinne et al., 2023; Phan et al., 2018). However, the ecological niche previously inferred for SARSCoV2r suggested that these viruses could occur in two regions of Vietnam (Hassanin, Tu, et al., 2021): the north-western region and the borders with Laos and Cambodia in southern Vietnam. These predictions were partly confirmed here as we detected two sarbecoviruses related to SARS-CoV-2 in Vietnam. On one hand, Ra22QT77 was found in four *R. affinis* bats collected in central Vietnam (Figure 1). Phylogenetically, it belongs to SARSCoV2r (Figure 5) and its SNC is very similar to that of SARS-CoV-2 (Figure 2). Most parts of its genome, including the Spike gene, appeared very similar to RshSTT200 (Figures 5 and 9), a virus sampled in northern Cambodia in 2010 (Delaune et al., 2021). On the other hand, Rp22DB159 was discovered in *R. pusillus* close to the border with Yunnan. This is the first RecSar virus found outside China. Although most parts of its genome were closely related to RecSar from Yunnan (RpJCC9, RpPrC31, and RpYN2021) (Figure 8), its Spike gene appeared very similar to SARS-CoV-2 (Figures 6 and 7).

The geographical distribution of SARSCoVr viruses here identified in Vietnam (Figure 9) fits well with the ecological niche previously inferred for SARSCoVr (Hassanin, Tu, et al., 2021), although the detection of Rt17DN420 in Dak Nong province indicates a more southerly distribution. The southern limit seems to coincide with the southern mountains of the Annamite Range. Most SARSCoVr viruses were sampled in *R. thomasi*, a bat species occurring in southern China, Vietnam, Laos, northern Thailand and eastern Myanmar (IUCN, 2022). A few SARSCoVr were also detected in *R. affinis*, suggesting cross-species transmission between *R. thomasi* and *R. affinis*. However, four points suggest that *R. thomasi* is the main reservoir host of SARSCoVr in Vietnam: (i) the geographical distribution of *R. thomasi* in Vietnam (IUCN, 2022) coincides well with that of SARSCoVr; (ii) the overall prevalence of SARSCoVr is much more important in *R. thomasi* (46%) than *R. affinis* (6%); (iii) in location N°13 (central Vietnam), we sequenced the same SARSCoVr in both *Rhinolophus* species, but with more positives in *R. thomasi* (7 vs. 1), suggesting a recent transmission from *R. thomasi* to *R. affinis*; the hypothesis was also corroborated by the tree of Figure 1, as Ra22QT27, extracted from *R. affinis*, was placed within a group of SARSCoVr sampled in *R. thomasi*; and (iv) the diversity of SARSCoVr was higher in *R. thomasi* than *R. affinis* (11 and 4 viruses respectively).

Our analysis of mitochondrial haplotypes for *R. thomasi* showed that Vietnamese bats carrying SARSCoVr fall into two geographical clades (Figure S11): (i) the northern clade includes bats from northern Vietnam (Cao Bang, Dien Bien, Lao Cai and Son La province), north-eastern Laos (Houaphan province) and south-eastern China (Guangxi, Guizhou and Hunan provinces); and (ii) the southern clade contains bats from three latitudinal lineages, that is, the first includes bats from Vientiane in northern Laos, the second comprises bats from the Annamite Range in central Vietnam (Quang Binh and Quang Tri provinces) and central and southern Laos (Champasak and Khammouan provinces), and the third is represented by the single bat from southern Vietnam (Dak Nong province). A north-south phylogeographic pattern has been previously observed for several other bat taxa distributed in Vietnam, such as *Kerivoula*, *Scotophilus*

and *Tylonycteris* species (Tu et al., 2017, 2018, 2021). The main geographical separation fits well with the Köppen–Geiger climate classification (Peel et al., 2007), which divides Vietnam into two latitudinal regions: the northern third has a more temperate climate (from China to Nghe An province), generally referred as a humid subtropical climate, whereas the southern two-thirds has a tropical wet climate (from Ha Tinh province in the north to the Mekong River delta in the south) (Figure 9). In subtropical northern Vietnam, winter lasts usually 2 or 3 months and temperatures can fall below freezing in mountainous areas. In tropical southern Vietnam, the temperatures of coldest months rarely drop below 18°C (Van et al., 2000). This suggests that northern bats need to hibernate in winter, whereas their southern counterparts do not.

The latitudinal structure of genetic diversity in *R. thomasi* suggests that bats of this species have a limited dispersal capacity. Interestingly, the SARSCoVr isolated from *R. thomasi* showed the same phylogeographic pattern, with three latitudinal lineages detected in northern, central and southern Vietnam (Figures 5 and 9). To our knowledge, this is the first study showing evidence of latitudinal coevolution between sarbecoviruses and their bat hosts. In addition, the three latitudinal lineages of SARSCoVr detected in *R. thomasi* showed different SNCs (Figure 2) and we found an inverse linear correlation between the latitude and level of C-to-U mutations at synonymous third codon positions (Figure S3). What bat host-dependent mechanism(s) may decrease the levels of cytosine deamination in SARSCoVr genomes with latitude? As previously discussed in Hassanin et al. (2022), several arguments suggest a key role of bat hibernation: (i) bat hibernation may be correlated with latitude because it depends on winter length (Dunbar & Brigham, 2010); (ii) bat hibernation is likely to impact the SNC of *Sarbecovirus* genomes via two possible mechanisms: viral replication may be significantly reduced; and the concentrations of free nucleotides available in bat cells may be modified due to reduction and remodelling of many metabolic pathways (Andrews, 2007); (iii) C-to-U mutations are more common in sarbecoviruses of tropical bats than those of temperate bats; this bias was confirmed in our study as we found higher percentages of U and lower percentages of C in *Sarbecovirus* genomes extracted from bats of tropical latitudes than in their closest relatives extracted from bats of higher latitudes (subtropical and temperate climates) (Table S5): RspKY72 from Kenya versus European sarbecoviruses; and SARSCoV2r of Southeast Asia versus Japan sarbecoviruses and *YunSar* endemic to Yunnan. The hibernation hypothesis could be further tested in Vietnam by comparing seasonal variations in cellular nucleotide pools in *R. thomasi* bats from different locations along a latitudinal gradient. The role of bat APOBEC cytosine deaminases (Kim et al., 2022) should also be explored.

## 4.2 | Emergence of SARS-CoV in horseshoe bats of northern Yunnan

Our phylogeographic analyses provides strong evidence that the progenitor of SARS-CoV originated in *Rhinolophus* species living in

Yunnan: the geographical contribution of Yunnan was significantly higher than other geographical areas ( $C_{TG/EG} = 96\%/39\%$ ; Figure 9a); and all parts of the Spike gene of SARS-CoV provided robust relationships with SARSCoVr sampled exclusively in Yunnan from *R. sinicus* ( $C_{TH/EH} = 90\%/71\%$ ) and *R. affinis* ( $C_{TH/EH} = 37\%/10\%$ ). In addition, the highest virus contributions came from northern Yunnan (Figure S12): RfYNLF/31C in Lufeng (Lau et al., 2015) and Rs4874, Rs7327, Rs9401, Rs3367 and RsSHC014 close to Kunming city (Ge et al., 2013; Hu et al., 2017) (Figure S12). In agreement with an origin in northern Yunnan, the two host species showing the highest contributions to SARS-CoV, *R. (ferrumequinum) nippon* ( $C_{TH/EH} = 73\%/17\%$ ) and *R. sinicus* ( $C_{TH/EH} = 71\%/11\%$ ) (Figure S12), are both distributed in northern Yunnan but not in southern Yunnan (IUCN, 2022).

## 4.3 | Spike genes related to SARS-CoV-2 circulate in horseshoe bats of northern Indochina subtropical forests

Several studies have revealed an exceptional diversity of sarbecoviruses in southern Yunnan: three divergent *Sarbecovirus* lineages were sampled in *Rhinolophus* bats collected in Mengla county, that is, SARSCoVr, SARSCoV2r and *YunSar* (Zhou et al., 2021); and several recombinant viruses between SARSCoVr and SARSCoV2r (*RecSar*) were detected in *R. pusillus* near Pu'er City (Li et al., 2021; Wu et al., 2023). Despite limited sampling effort in location N°3 of north-western Vietnam, we detected there two *RecSar* (Rp22DB159 and RpDB167) and two SARSCoVr, including one exhibiting a very divergent Spike gene (Ra22DB163). The highlands of southern Yunnan and north-western Vietnam belong to the same ecoregion, namely northern Indochina subtropical forests (Eco region ID number: 256; Dinerstein et al., 2017), which also includes the highlands of north-eastern Myanmar and northern Laos. Various bat species live there: some of them have more ecological affinities with Chinese subtropical forests (e.g. *Rhinolophus rex*), whereas others have more affinities with Southeast Asian tropical forests (e.g. *R. malayanus*) and some species are relatively ubiquitous and occur in both subtropical and tropical latitudes (e.g. *R. affinis* and *R. pusillus*) (IUCN, 2022). As several *Rhinolophus* species often occupy the same caves, they can occasionally exchange viruses, suggesting that co-infection with divergent sarbecoviruses can generate, by genomic recombination, chimeric viruses potentially better adapted to a rapidly evolving environment than their parental strains. We argue here that it was the case for Rp22DB159 from north-western Vietnam. Although most regions of its genome showed exclusive ancestry with three *RecSar* from Yunnan (RpJCC9, RpPrC31 and RpYN2021), its Spike gene was found closely related to SARS-CoV-2 and four bat SARSCoV2r (RaTG13, RmBANAL52, RmaBANAL236, and RpBANAL103) (Figures 6–8). We therefore concluded that the Spike gene of Rp22DB159 was acquired through a recent recombination of SARSCoV2r and *RecSar*, most probably in a *R. pusillus* circulating in southern Yunnan and north-western Vietnam. Two points



suggest that recombination of divergent sarbecoviruses might be more frequent in *R. pusillus* than other horseshoe bat species: co-infection with four distinct coronaviruses, including two sarbecoviruses, was detected in a faecal sample (DB167) from *R. pusillus* in north-western Vietnam; and the eight *RecSar* currently identified were all sampled in *R. pusillus* (Hu et al., 2018; Li et al., 2021; Wu et al., 2023). Higher rates of recombination can accelerate the adaptive evolution of sarbecoviruses, thereby facilitating interspecies transmission (Simon-Lorieri & Holmes, 2011). In relation to this, we propose that the SARSCoV2r X *RecSar* recombination in the progenitor of Rp22DB159 created an advantageous genetic combination. The spike RBD of Rp22DB159 is very similar to that of SARS-CoV-2, differing in only six amino acids (Figure 7), including two residues interacting with human ACE2 in the RBM (Lan et al., 2020), L486F and H498Q. Based on our analyses (Figures 6, 7 and S8), it can be inferred that the common ancestor of SARS-CoV-2, Rp22DB159, RmBANAL52, RmaBANAL236, and RpBANAL103 was characterised by F486 and H498, and that a F486L mutation has arisen in the progenitor of Rp22DB159 while a H498Q mutation has occurred in the progenitor of SARS-CoV-2. Importantly, both F486L and Q498H mutations were detected in SARS-CoV-2 during the COVID-19 pandemic, and neither was found to have a negative effect on the binding affinity of RBD to human ACE2 (Huang et al., 2021; Zhou et al., 2022). In agreement with experimental studies (Nie et al., 2021; Wacharapluesadee et al., 2021; Zhang et al., 2021), we suggest that the five animal spike proteins with an RBD very similar to SARS-CoV-2 (Rp22DB159, RmBANAL52, RmaBANAL236, RpBANAL103 and MjGuangdong) are all able to efficiently bind to the human ACE2 receptor and mediate entry and replication in human cells (despite the absence of the furin cleavage motif). This is not the case for viruses exhibiting a more divergent spike RBD/RBM, such as RaTG13 (Zhang et al., 2021). The SARS-CoV-2-like RBD appears to be generalist since it was capable of infecting host species representing at least five mammalian orders, including Chiroptera (*R. malayanus*, *R. marshalli* and *R. pusillus*; Temmam et al., 2022; this study) and Pholidota (*Manis javanica*; Lam et al., 2020; Liu et al., 2019) before the COVID-19 pandemic, and then Primates (humans), Carnivora (e.g. farmed mink; Oude Munnink et al., 2021), and Cetartiodactyla (e.g. white-tailed deer; Hale et al., 2022).

#### 4.4 | Emergence of SARS-CoV-2 in horseshoe bats of northern Indochina subtropical forests

Based on currently available data, our phylogeographic analysis provides strong evidence that the progenitors of SARS-CoV-2 evolved in *Rhinolophus* species circulating in the area covering Yunnan, north-western Vietnam and northern Laos (Figure S13): the highest contributions were found for SARSCoV2r collected in northern Laos (RpBANAL103, RmBANAL52 and RmaBANAL236;  $C_T = 72\% - 65\%$ ) and Yunnan (RmYN02, RpYN06 and RaTG13;  $C_T = 64\% - 60\%$ ); SARS-CoV-2 was found to share exclusive ancestry with all

the three SARSCoV2r from Yunnan, that is, RpYN06 ( $C_E = 5.8\%$ ), RmYN02 (5.5%) and RaTG13 (4.2%), three of the four SARSCoV2r from Laos, that is, RmBANAL247 (3.2%), RmaBANAL236 (1.7%) and RpBANAL103 (0.7%), and two *RecSar* viruses, that is, Rp22DB159 from north-western Vietnam (1%) and RpJCC9 from Yunnan (0.8%). In addition, most analyses provided more support for an origin in Yunnan and bordering regions in north-western Vietnam (i.e. Location N°3) than Laos: both geographical indicators were higher for Yunnan than Laos based on the whole-genome alignment ( $C_{TG/EG} = 84\%/19\%$  vs.  $78\%/15\%$ ; Figure 9) and the Spike gene only ( $C_{TG/EG} = 97\%/27\%$  vs.  $71\%/0\%$ ; Figure S14); the Spike gene of SARS-CoV-2 showed more phylogenetic affinities with Rp22DB159 and RaTG13 than BANAL viruses (Figures 4b, 6 and 7); exclusive ancestry was more important with RmYN02, RpYN06 and RaTG13 from Yunnan than with BANAL viruses from Laos. However, three SARSCoV2r collected in northern Laos (RpBANAL103, RmBANAL52 and RmaBANAL236) showed higher overall phylogenetic contributions than the three SARSCoV2r described from southern Yunnan (RmYN02, RpYN06 and RaTG13) (Figure S13). This point suggests that additional localities in northern Southeast Asia and southern Yunnan should be explored to more accurately identify the geographical origin of SARS-CoV-2. Based on our results, we consider that viruses closely related to SARS-CoV-2 are likely to circulate in *R. pusillus* and *R. malayanus* (which are the two species showing the highest contributions;  $C_{TH/EH} = 84\%/10.8\%$  and  $82\%/7.6\%$  respectively; Figure S13) along the borders between Yunnan and Southeast Asia countries, including the Xishuangbanna region in southern Yunnan, the Phongsaly province in northern Laos and the Shan province of eastern Myanmar.

#### AUTHOR CONTRIBUTIONS

A.H., V.T.T. and S.W. designed and supervised the research. V.T.T., L.Q.N., P.V.P., C.T.H. and T.A.T. collected samples in 2021 and 2022 (with A.H. in June 2022). V.T.T. and G.K. collected samples in 2017. A.H. and V.T.T. made most RNA extractions. S.W. made most PCRs and MiSeq sequencing runs. M.P. and E.S.L. made NextSeq sequencing. T.G. and G.E.T. extracted RNA from samples collected in 2017, made PCRs and assembled Rt17DN420. A.H. made genome assembly of all other samples. A.H. performed phylogenetic analyses and constructed GB and CGB barcodes. V.T.T. and A.H. performed mt-Co1 analyses. A.H. wrote the paper. All authors edited the paper.

#### ACKNOWLEDGEMENTS

We acknowledge the Provincial People's Committees of numerous provinces, the Forest Protection Department of the Ministry of Agriculture and Rural Development, the local manager boards of different protected areas, and the Institute of Ecology and Biological Resources (IEBR, Hanoi, Vietnam) for field research authorisations. We thank Marion Goulet, Morgane Levert (Eau de Paris), Brigitta Zana, Ágota Ábrahám and Zsófia Lanszki (National Laboratory of Virology, University of Pécs) for technical help during laboratory work, Nguyen Van Sinh, Nguyen Quang Truong and

Le Hung Anh (IEBR) for administrative supports and Neil Furey for English-language reviewing. A.H. thanks Vincent Maréchal for introducing L.M. and S.W. We thank the three anonymous reviewers for their helpful comments on the first version of the manuscript. The research was funded by the 'Agence nationale de la recherche' (AAP RA-COVID-19, grant number ANR-21-CO12-0002). Field surveys were also funded by the Vietnam Academy of Science and Technology (Project No: QTHU01.01/22-23 and KHCBTD.02/22-24). The research on Rt17DN420 received support from the National Research, Development, and Innovation Fund of Hungary (NKFIH FK137778, RRF-2.3.1-21-2022-00010), and the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (BO/00825/21).

### CONFLICT OF INTEREST STATEMENT

The authors report that there are no competing interests to declare.

### OPEN RESEARCH BADGES



This article has earned Open Data, Open Materials and Preregistered Research Design badges. Data, materials and the preregistered design and analysis plan are available at <https://osf.io/k5ehf>.

### DATA AVAILABILITY STATEMENT

The meta-transcriptomic sequencing reads generated in this study have been deposited in the SRA database under accession codes PRJNA1025946 and PRJNA1027129. The viral genome sequences have been deposited in GenBank with the accession codes OR233291-OR233328.

### BENEFIT SHARING STATEMENT

Benefits from this research accrue from the sharing of our data and results on public databases. The datasets generated and analysed during the current study, including the whole-genome alignment, SWB CSV files (5), BBC CSV files (5), SuperTRI files (2925 Log files used as inputs for SuperTRI, 5 MRP files and 5 SB trees), and CGB files (31 BBC Excel files, 47 CSV files and all associated PNG files) are available in the Open Science Framework (OSF) platform at <https://osf.io/k5ehf>. The SuperTRI program is available at <http://www.normalesup.org/~bli/Programs/programs.html>. The SWB, BBC, CGB and LFG programs are available at [https://github.com/OpaleRambaud/GB\\_barcodes\\_project](https://github.com/OpaleRambaud/GB_barcodes_project).

### ORCID

Alexandre Hassanin  <https://orcid.org/0000-0002-4905-8540>

### REFERENCES

Alkhovsky, S., Lenshin, S., Romashin, A., Vishnevskaya, T., Vyshemirsky, O., Bulycheva, Y., Lvov, D., & Gitelman, A. (2022). SARS-like coronaviruses in horseshoe bats (*Rhinolophus* spp.) in Russia, 2020. *Viruses*, 14(1), 113. <https://doi.org/10.3390/v14010113>

- Andrews, M. T. (2007). Advances in molecular biology of hibernation in mammals. *BioEssays*, 29(5), 431–440. <https://doi.org/10.1002/bies.20560>
- Berto, A., Anh, P. H., Carrique-Mas, J. J., Simmonds, P., Van Cuong, N., Tue, N. T., Van Dung, N., Woolhouse, M. E., Smith, I., Marsh, G. A., Bryant, J. E., Thwaites, G. E., Baker, S., Rabaa, M. A., & VIZIONS Consortium. (2018). Detection of potentially novel paramyxovirus and coronavirus viral RNA in bats and rats in the Mekong Delta region of southern Viet Nam. *Zoonoses and Public Health*, 65(1), 30–42. <https://doi.org/10.1111/zph.12362>
- Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., Rambaut, A., & Robertson, D. L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology*, 5(11), 1408–1417. <https://doi.org/10.1038/s41564-020-0771-4>
- Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M. L., Mulders, D. G., Haagmans, B. L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J. L., Ellis, J., Zambon, M., ... Drosten, C. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveillance: Bulletin Europeen sur les maladies transmissibles = European Communicable Disease Bulletin*, 25(3), 2000045. <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>
- Crook, J. M., Murphy, I., Carter, D. P., Pullan, S. T., Carroll, M., Vipond, R., Cunningham, A. A., & Bell, D. (2021). Metagenomic identification of a new sarbecovirus from horseshoe bats in Europe. *Scientific Reports*, 11(1), 14723. <https://doi.org/10.1038/s41598-021-94011-z>
- Delaune, D., Hul, V., Karlsson, E. A., Hassanin, A., Ou, T. P., Baidaliuk, A., Gámbaro, F., Prot, M., Tu, V. T., Chea, S., Keatts, L., Mazet, J., Johnson, C. K., Buchy, P., Dussart, P., Goldstein, T., Simon-Lorière, E., & Duong, V. (2021). A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nature Communications*, 12(1), 6563. <https://doi.org/10.1038/s41467-021-26809-4>
- Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E. C., Jones, B., Barber, C. V., Hayes, R., Kormos, C., Martin, V., Crist, E., ... Saleem, M. (2017). An ecoregion-based approach to protecting half the terrestrial realm. *Bioscience*, 67(6), 534–545. <https://doi.org/10.1093/biosci/bix014>
- DNA Pipelines R&D, Farr, B., Rajan, D., Betteridge, E., Shirley, L., Quail, M., Park, N., Redshaw, N., Bronner, I. F., Aigrain, L., Goodwin, S., Thurston, S., Lensing, S., Bonfield, J., James, K., Salmon, N., Beaver, C., Nelson, R., Jackson, D. K., ... Johnston, I. (2020). COVID-19 ARTIC v3 Illumina library construction and sequencing protocol V.5. <https://doi.org/10.17504/protocols.io.bibtckann>
- 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., Herler, 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. *Journal of Virology*, 84(21), 11336–11349. <https://doi.org/10.1128/JVI.00650-10>
- Dunbar, M. B., & Brigham, R. M. (2010). Thermoregulatory variation among populations of bats along a latitudinal gradient. *Journal of Comparative Physiology B*, 180(6), 885–893. <https://doi.org/10.1007/s00360-010-0457-y>
- Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular evolution of human coronavirus genomes. *Trends in Microbiology*, 25(1), 35–48. <https://doi.org/10.1016/j.tim.2016.09.001>
- Francis, C. M. (2019). *Field guide to the mammals of South East-Asia* (2nd ed., p. 416). Bloomsbury Wildlife.



- 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>
- 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 & Infections*, 10(1), 1507–1514. <https://doi.org/10.1080/22221751.2021.1956373>
- Hale, V. L., Dennis, P. M., McBride, D. S., Nolting, J. M., Madden, C., Huey, D., Ehrlich, M., Grieser, J., Winston, J., Lombardi, D., Gibson, S., Saif, L., Killian, M. L., Lantz, K., Tell, R. M., Torchetti, M., Robbe-Austerman, S., Nelson, M. I., Faith, S. A., & Bowman, A. S. (2022). SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature*, 602(7897), 481–486. <https://doi.org/10.1038/s41586-021-04353-x>
- Han, Y., Du, J., Su, H., Zhang, J., Zhu, G., Zhang, S., Wu, Z., & Jin, Q. (2019). Identification of diverse bat alphacoronaviruses and betacoronaviruses in China provides new insights into the evolution and origin of coronavirus-related diseases. *Frontiers in Microbiology*, 10, 1900. <https://doi.org/10.3389/fmicb.2019.01900>
- Hassanin, A. (2022). Variation in synonymous nucleotide composition among genomes of sarbecoviruses and consequences for the origin of COVID-19. *Gene*, 835, 146641. <https://doi.org/10.1016/j.gene.2022.146641>
- Hassanin, A., Grandcolas, P., & Veron, G. (2021). Covid-19: Natural or anthropic origin? *Mammalia*, 85(1), 1–7. <https://doi.org/10.1515/mammalia-2020-0044>
- Hassanin, A., & Rambaud, O. (2023). Retracing phylogenetic, host and geographic origins of coronaviruses with coloured genomic bootstrap barcodes: SARS-CoV and SARS-CoV-2 as case studies. *Viruses*, 15(2), 406. <https://doi.org/10.3390/v15020406>
- 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>
- Hassanin, A., Tu, V. T., Curaudeau, M., & Corsora, G. (2021). Inferring the ecological niche of bat viruses closely related to SARS-CoV-2 using phylogeographic analyses of *Rhinolophus* species. *Scientific Reports*, 11(1), 14276. <https://doi.org/10.1038/s41598-021-93738-z>
- Hassanin, A., Tu, V. T., Pham, P. V., Ngon, L. Q., Chabane, T., Moulin, L., & Wurtzer, S. (2024). Bat rhinacoviruses related to swine acute diarrhoea syndrome coronavirus evolve under strong host and geographic constraints in China and Vietnam. *Viruses*, 16, 1114. <https://doi.org/10.3390/v16071114>
- 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. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathogens*, 13(11), e1006698. <https://doi.org/10.1371/journal.ppat.1006698>
- Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., Yang, L., 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 & Infections*, 7(1), 154. <https://doi.org/10.1038/s41426-018-0155-5>
- Huang, K., Zhang, Y., Hui, X., Zhao, Y., Gong, W., Wang, T., Zhang, S., Yang, Y., Deng, F., Zhang, Q., Chen, X., Yang, Y., Sun, X., Chen, H., Tao, Y. J., Zou, Z., & Jin, M. (2021). Q493K and Q498H substitutions in spike promote adaptation of SARS-CoV-2 in mice. *eBioMedicine*, 67, 103381. <https://doi.org/10.1016/j.ebiom.2021.103381>
- IUCN. (2022). *The IUCN red list of threatened species. Version 2022-2*. Retrieved June 10, 2023, from <https://www.iucnredlist.org>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–780. <https://doi.org/10.1093/molbev/mst010>
- Kim, K., Calabrese, P., Wang, S., Qin, C., Rao, Y., Feng, P., & Chen, X. S. (2022). The roles of APOBEC-mediated RNA editing in SARS-CoV-2 mutations, replication and fitness. *Scientific Reports*, 12(1), 14972. <https://doi.org/10.1038/s41598-022-19067-x>
- Kruskop, S. V. (2013). *Bats of Vietnam: Checklist and an identification manual. Izdanie vtoroe, pererabotannoe i dopolnennoe* (p. 299). Tovarishchestvo nauchnykh izdaniy KMK.
- Lam, T. T., Jia, N., Zhang, Y. W., Shum, M. H., Jiang, J. F., Zhu, H. C., Tong, Y. G., Shi, Y. X., Ni, X. B., Liao, Y. S., Li, W. J., Jiang, B. G., Wei, W., Yuan, T. T., Zheng, K., Cui, X. M., Li, J., Pei, G. Q., Qiang, X., ... 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>
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807), 215–220. <https://doi.org/10.1038/s41586-020-2180-5>
- Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22), 3276–3278. <https://doi.org/10.1093/bioinformatics/btu531>
- Latinne, A., Nga, N. T. T., Long, N. V., Ngoc, P. T. B., Thuy, H. B., Predict Consortium, Long, N. V., Long, P. T., Phuong, N. T., Quang, L. T. V., Tung, N., Nam, V. S., Duoc, V. T., Thinh, N. D., Schoepp, R., Ricks, K., Inui, K., Padungtod, P., Johnson, C. K., ... Fine, A. E. (2023). One health surveillance highlights circulation of viruses with zoonotic potential in bats, pigs, and humans in Viet Nam. *Viruses*, 15(3), 790. <https://doi.org/10.3390/v15030790>
- Lau, S. K., Feng, Y., Chen, H., Luk, H. K., Yang, W. H., Li, K. S., Zhang, Y. Z., Huang, Y., Song, Z. Z., Chow, W. N., Fan, R. Y., Ahmed, S. S., Yeung, H. C., Lam, C. S., Cai, J. P., Wong, S. S., Chan, J. F., Yuen, K. Y., Zhang, H. L., & Woo, P. C. (2015). Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. *Journal of Virology*, 89(20), 10532–10547. <https://doi.org/10.1128/JVI.01048-15>
- 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. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39), 14040–14045. <https://doi.org/10.1073/pnas.0506735102>
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multi-variate analysis. *Journal of Statistical Software*, 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*, 310(5748), 676–679. <https://doi.org/10.1126/science.1118391>
- Liu, P., Chen, W., & Chen, J. P. (2019). Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses*, 11(11), 979. <https://doi.org/10.3390/v11110979>
- Liu, W. J., Liu, P., Lei, W., Jia, Z., He, X., Shi, W., Tan, Y., Zou, S., Wong, G., Wang, J., Wang, F., Wang, G., Qin, K., Gao, R., Zhang, J., Li, M., Xiao, W., Guo, Y., Xu, Z., ... Wu, G. (2023). Surveillance of SARS-CoV-2 at the Huanan seafood market. *Nature*, 631, 402–408. <https://doi.org/10.1038/s41586-023-06043-2>

- Murakami, S., Kitamura, T., Matsugo, H., Kamiki, H., Oyabu, K., Sekine, W., Takenaka-Uema, A., Sakai-Tagawa, Y., Kawaoka, Y., & Horimoto, T. (2022). Isolation of bat Sarbecoviruses, Japan. *Emerging Infectious Diseases*, 28(12), 2500–2503. <https://doi.org/10.3201/eid2812.220801>
- Nie, J., Li, Q., Zhang, L., Cao, Y., Zhang, Y., Li, T., Wu, J., Liu, S., Zhang, M., Zhao, C., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Li, X., Liu, J., Liang, H., Jiang, T., ... Wang, Y. (2021). Functional comparison of SARS-CoV-2 with closely related pangolin and bat coronaviruses. *Cell Discovery*, 7(1), 21. <https://doi.org/10.1038/s41421-021-00256-3>
- 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., ... Koopmans, M. P. G. (2021). Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science*, 371(6525), 172–177. <https://doi.org/10.1126/science.abe5901>
- Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, 11, 1633–1644. <https://doi.org/10.5194/hess-11-1633-2007>
- Phan, M. V. T., Ngo Tri, T., Hong Anh, P., Baker, S., Kellam, P., & Cotten, M. (2018). Identification and characterization of coronaviridae genomes from Vietnamese bats and rats based on conserved protein domains. *Virus Evolution*, 4(2), vey035. <https://doi.org/10.1093/ve/vey035>
- Posada, D., Crandall, K. A., & Holmes, E. C. (2002). Recombination in evolutionary genomics. *Annual Review of Genetics*, 36, 75–97. <https://doi.org/10.1146/annurev.genet.36.040202.111115>
- Ropiquet, A., Li, B., & Hassanin, A. (2009). SuperTRI: A new approach based on branch support analyses of multiple independent data sets for assessing reliability of phylogenetic inferences. *Comptes Rendus Biologies*, 332(9), 832–847. <https://doi.org/10.1016/j.crvbi.2009.05.001>
- Samson, S., Lord, É., & Makarenkov, V. (2022). SimPlot++: A python application for representing sequence similarity and detecting recombination. *Bioinformatics*, 38(11), 3118–3120. <https://doi.org/10.1093/bioinformatics/btac287>
- Simmons, N. B., & Cirranello, A. L. (2020). *Bats of the world: A taxonomic and geographic database*. <https://batnames.org/explore.html>
- Simon-Loriere, E., & Holmes, E. C. (2011). Why do RNA viruses recombine? *Nature Reviews Microbiology*, 9(8), 617–626. <https://doi.org/10.1038/nrmicro2614>
- Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Swofford, D. L. (2021). *PAUP\*. Phylogenetic analysis using parsimony (\* and other methods)*. Version 4. Sinauer Associates.
- Tao, Y., & Tong, S. (2019). Complete genome sequence of a severe acute respiratory syndrome-related coronavirus from Kenyan bats. *Microbiology Resource Announcements*, 8(28), e00548-19. <https://doi.org/10.1128/MRA.00548-19>
- Temmam, S., Montagutelli, X., Herate, C., Donati, F., Regnault, B., Attia, M., Baquero Salazar, E., Chretien, D., Conquet, L., Jouvion, G., Pipoli Da Fonseca, J., Cokelaer, T., Amara, F., Relouzat, F., Naninck, T., Lemaitre, J., Derreudre-Bosquet, N., Pascal, Q., Bonomi, M., ... Eloit, M. (2023). SARS-CoV-2-related bat virus behavior in human-relevant models sheds light on the origin of COVID-19. *EMBO Reports*, 24(4), e56055. <https://doi.org/10.15252/embr.202256055>
- Temmam, S., Vongphayloth, K., Baquero, E., Munier, S., Bonomi, M., Regnault, B., Douangboubpha, B., Karami, Y., Chretien, D., Sanamxay, D., Xayaphet, V., Paphaphanh, P., Lacoste, V., Somlor, S., Lakeomany, K., Phommavanh, N., Pérot, P., Dehan, O., Amara, F., ... Eloit, M. (2022). Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature*, 604(7905), 330–336. <https://doi.org/10.1038/s41586-022-04532-4>
- Trujillo, M., Cheung, K., Gao, A., Hoxie, I., Kannoly, S., Kubota, N., San, K. M., Smyth, D. S., & Dennehy, J. J. (2021). Protocol for safe, affordable, and reproducible isolation and quantitation of SARS-CoV-2 RNA from wastewater. *PLoS ONE*, 16(9), e0257454. <https://doi.org/10.1371/journal.pone.0257454>
- Tu, V. T., Csorba, G., Ruedi, M., Furey, N. M., Son, N. T., Thong, V. D., Bonillo, C., & Hassanin, A. (2017). Comparative phylogeography of bamboo bats of the genus *Tylonycteris* (Chiroptera, Vespertilionidae) in Southeast Asia. *European Journal of Taxonomy*, 274, 1–38. <https://doi.org/10.5852/ejt.2017.274>
- Tu, V. T., Görföl, T., Csorba, G., Arai, S., Kikuchi, F., Fukui, D., Koyabu, D., Furey, N. M., Bawm, S., Lin, K. S., Alviola, P., Hang, C. T., Son, N. T., Tuan, T. A., & Hassanin, A. (2021). Integrative taxonomy and biogeography of Asian yellow house bats (Vespertilionidae: *Scotophilus*) in the Indomalayan region. *Journal of Zoological Systematics and Evolutionary Research*, 59, 772–795. <https://doi.org/10.1111/jzs.12448>
- Tu, V. T., Hassanin, A., Furey, N. M., Son, N. T., & Csorba, G. (2018). Four species in one: Multigene analyses reveal phylogenetic patterns within Hardwicke's woolly bat, *Kerivoula hardwickii*-complex (Chiroptera, Vespertilionidae) in Asia. *Hystrix, the Italian Journal of Mammalogy*, 29(1), 111–121. <https://doi.org/10.4404/hystrix-00017-2017>
- Van, K. V., Hien, N. T., Loc, P. K., & Hiep, N. T. (2000). *Bioclimatic diagrams of Vietnam* (p. 126). Vietnam National University Publishing House.
- Wacharapluesadee, S., Tan, C. W., Maneeorn, P., Duengkae, P., Zhu, F., Joyjinda, Y., Kaewpoom, T., Chia, W. N., Ampoot, W., Lim, B. L., Worachotsueptrakun, K., Chen, V. C., Sirichan, N., Ruchisrisarod, C., Rodpan, A., Noradechanon, K., Phaichana, T., Jantarat, N., Thongnumchaima, B., ... Wang, L. F. (2021). Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nature Communications*, 12(1), 972. <https://doi.org/10.1038/s41467-021-21240-1>
- Worobey, M., Levy, J. I., Malpica Serrano, L., Crits-Christoph, A., Pekar, J. E., Goldstein, S. A., Rasmussen, A. L., Kraemer, M. U. G., Newman, C., Koopmans, M. P. G., Suchard, M. A., Wertheim, J. O., Lemey, P., Robertson, D. L., Garry, R. F., Holmes, E. C., Rambaut, A., & Andersen, K. G. (2022). The Huanan seafood wholesale market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science*, 377(6609), 951–959. <https://doi.org/10.1126/science.abp8715>
- Wu, Z., Han, Y., Wang, Y., Liu, B., Zhao, L., Zhang, J., Su, H., Zhao, W., Liu, L., Bai, S., Dong, J., Sun, L., Zhu, Y., Zhou, S., Song, Y., Sui, H., Yang, J., Wang, J., Zhang, S., ... Jin, Q. (2023). A comprehensive survey of bat sarbecoviruses across China in relation to the origins of SARS-CoV and SARS-CoV-2. *National Science Review*, 10, nwac213. <https://doi.org/10.1093/nsr/nwac213>
- 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>
- Wurtzer, S., Marechal, V., Mouchel, J. M., Maday, Y., Teyssou, R., Richard, E., Almayrac, J. L., & Moulin, L. (2020). Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in waste water, greater Paris, France, 5 March to 23 April 2020. *Euro Surveillance: Bulletin European Sur les Maladies Transmissibles = European Communicable Disease Bulletin*, 25(50), 2000776. <https://doi.org/10.2807/1560-7917.ES.2020.25.50.2000776>
- Wurtzer, S., Waldman, P., Ferrier-Rembert, A., Frenois-Veyrat, G., Mouchel, J. M., Boni, M., Maday, Y., OBEPINE Consortium, Marechal, V., & Moulin, L. (2021). Several forms of SARS-CoV-2 RNA can be detected in wastewaters: Implication for wastewater-based

- epidemiology and risk assessment. *Water Research*, 198, 117183. <https://doi.org/10.1016/j.watres.2021.117183>
- Zhang, S., Qiao, S., Yu, J., Zeng, J., Shan, S., Tian, L., Lan, J., Zhang, L., & Wang, X. (2021). Bat and pangolin coronavirus spike glycoprotein structures provide insights into SARS-CoV-2 evolution. *Nature Communications*, 12(1), 1607. <https://doi.org/10.1038/s41467-021-21767-3>
- 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*, 184(17), 4380–4391.e14. <https://doi.org/10.1016/j.cell.2021.06.008>
- Zhou, J., Peacock, T. P., Brown, J. C., Goldhill, D. H., Elrefaey, A. M. E., Penrice-Randal, R., Cowton, V. M., De Lorenzo, G., Furnon, W., Harvey, W. T., Kugathasan, R., Frise, R., Baillon, L., Lassaunière, R., Thakur, N., Gallo, G., Goldswain, H., Donovan-Banfield, I., Dong, X., ... Barclay, W. S. (2022). Mutations that adapt SARS-CoV-2 to mink or ferret do not increase fitness in the human airway. *Cell Reports*, 38(6), 110344. <https://doi.org/10.1016/j.celrep.2022.110344>
- 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., ... Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270–273. <https://doi.org/10.1038/s41586-020-2012-7>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Hassanin, A., Tu, V. T., Görföl, T., Ngon, L. Q., Pham, P. V., Hang, C. T., Tuan, T. A., Prot, M., Simon-Lorière, E., Kemenesi, G., Tóth, G. E., Moulin, L., & Wurtzer, S. (2024). Phylogeography of horseshoe bat sarbecoviruses in Vietnam and neighbouring countries. Implications for the origins of SARS-CoV and SARS-CoV-2. *Molecular Ecology*, 33, e17486. <https://doi.org/10.1111/mec.17486>