scientific reports

OPEN

Metagenomic identification of a new sarbecovirus from horseshoe bats in Europe

Jack M. Crook[1,5,6], Ivana Murphy[2,6], Daniel P. Carter[1], Steven T. Pullan[1], Miles Carroll[1,3], Richard Vipond[1], Andrew A. Cunningham[4,6]✉ & Diana Bell[2,6]

The source of the COVID-19 pandemic is unknown, but the natural host of the progenitor sarbecovirus is thought to be Asian horseshoe (rhinolophid) bats. We identified and sequenced a novel sarbecovirus (RhGB01) from a British horseshoe bat, at the western extreme of the rhinolophid range. Our results extend both the geographic and species ranges of sarbecoviruses and suggest their presence throughout the horseshoe bat distribution. Within the spike protein receptor binding domain, but excluding the receptor binding motif, RhGB01 has a 77% (SARS-CoV-2) and 81% (SARS-CoV) amino acid homology. While apparently lacking hACE2 binding ability, and hence unlikely to be zoonotic without mutation, RhGB01 presents opportunity for SARS-CoV-2 and other sarbecovirus homologous recombination. Our findings highlight that the natural distribution of sarbecoviruses and opportunities for recombination through intermediate host co-infection are underestimated. Preventing transmission of SARS-CoV-2 to bats is critical with the current global mass vaccination campaign against this virus.

The sources of the current COVID-19 pandemic and of the 2003 Severe Acute Respiratory Syndrome (SARS) epidemic are unknown[1]. Currently, the natural hosts of both SARS-CoV and SARS-CoV-2 (family Coronaviridae; subgenus *Sarbecovirus*)[2], the causative agents of SARS and COVID-19 respectively, are thought to be horseshoe bats (Rhinolophidae), with the zoonotic spillover process involving one or more intermediate hosts, during which time viral mutation, recombination and/or amplification could have occurred[3–7]. Phylogenetic analyses of novel horseshoe bat sarbecoviruses in China have shown these to be most closely related to both SARS-CoV and to SARS-CoV-2[3,4,7]. Recently, a SARS-CoV-2-like virus was also reported from a species of horseshoe bat in Thailand[8].

The range of horseshoe bats extends across much of the Old World, but most sampling for coronaviruses has been conducted in East and South East Asia, where around 50 SARS-related coronaviruses (SARSr-CoVs) have been detected across ten species of bat, with 48 of these being from nine species of horseshoe bat[8–11]. Here we expand the investigation of SARSr-CoVs to the lesser horseshoe bat (*Rhinolophus hipposideros*) in the UK, which is at the western extreme of the range of the Rhinolophidae.

Methods

Sample collection. Faecal samples were collected from lesser horseshoe bats *R. hipposideros* during routine annual population monitoring surveys at three sites in August and September 2020 following approval by the University of East Anglia Ethics Committee and adhering to UK Government COVID-19 safety regulations in place at the time. Samples were either obtained from individual bats at the time of capture or from under-roost sampling. In Somerset (n = 20 bats) and Monmouthshire, Wales (n = 7) bats were captured using harp traps or mist nets placed near roosts or in woodland under government license. A faecal pellet was collected from each of 23 bats held individually in sterile holding bags and the other four samples were collected as anal swabs using rayon-tipped dry swabs (MW100; Medical Wire & Equipment), which were taken from bats that did not defaecate when captured. All bats were released at the site of capture immediately after sample collection. In Glouces-

[1]National Infection Service, Public Health England, Porton Down, Salisbury, UK. [2]Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK. [3]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, Oxford University, Oxford, UK. [4]Institute of Zoology, Zoological Society of London, London NW1 4RY, UK. [5]NIHR Health Protection Unit in Emerging and Zoonotic Infections, Department of Clinical Infection, Microbiology and Immunology, University of Liverpool, Liverpool L69 7TX, UK. [6]These authors contributed equally: Jack M. Crook, Ivana Murphy, Andrew A. Cunningham and Diana Bell. ✉email: a.cunningham@ioz.ac.uk

tershire, 26 faecal pellets were collected by placing a sterile sheet underneath a lesser horseshoe roost and any droppings that landed directly onto the sheet within 30 min were collected. Each faecal sample was placed into an individual sterile tube containing 2 ml RNAlater, refrigerated overnight and stored frozen prior to analysis.

Genomic sequencing. For metagenomic analysis, samples were homogenised by vortexing and spiked with 10^6 genome copies per ml of Hazara virus as an internal control. A 140 µl aliquot of each sample was extracted using the QIAamp Viral RNA extraction kit. Extracts were DNase treated, reverse transcribed and randomly amplified using a Sequence-Independent Single-Primer Amplification (SISPA) based method described in detail previously[12]. Illumina sequencing used the Nextera XT protocol with 2 × 150-bp paired-end sequencing on a MiSeq. Nanopore library preparation was as described previously[12] and sequencing was performed on an Oxford nanopore GridION with base calling via Guppy. Nanopore reads were trimmed using NanoFilt[13] to remove 25 bp SISPA primer sequences from the start and end of each read. Raw data and the assembly sequence are deposited at NCBI under BioProject PRNJA706167.

Genomic analyses. Read-level taxonomic classification for each sample was performed using Kraken2 against the RefSeq database (2.0.8-beta)[14]. *De novo* genome assembly was performed using SPAdes (3.15.1) for both Illumina and Illumina/nanopore hybrid assemblies[15]. Contigs of interest were identified using BLASTn[16]. Illumina reads were mapped to the assembled contigs of interest using BWA-MEM[17] and nanopore reads using Minimap2[18]. Read depth values were generated using SAMtools (1.10)[19].

The assembled genome was aligned with selected reference genomes (NC_014470.1, KJ473814.1, NC_045512.2) in MEGA X (10.2.4) using MUSCLE alignment[20,21]. Nucleotide and codon alignments were generated for each gene with visual depictions of alignment and pairwise alignment scores generated in JALVIEW (2.11.1.3)[22].

Aligned nucleotide sequences from 21 sarbecoviruses obtained from GenBank were used to generate maximum likelihood trees using IQTREE (2.0.3)[23]. The best fitting nucleotide substitution model was used as selected by ModelFinder for each individual phylogeny[24]. Nodes were evaluated using UFBoot with 1,000 bootstrap approximations[25]. Phylogenetic tree visualisation was carried out in FigTree (5.7)[26].

Homology modelling. SWISS-MODEL was used for homology modelling of the tertiary structure of the receptor binding domain (RBD) in RhGB01[27]. The input target sequence was residues 324–515 in the spike protein sequence. SWISS-MODEL performs a template search of BLAST and HHblits databases for pre-existing characterised structures with homology. The most suitable template was selected according to the highest Global Model Quality Estimate (GMQE) and lowest Quaternary Structure Quality Estimate (QMEAN). GMQE expresses the accuracy of the model built as a number between 0 and 1. QMEAN uses statistical potentials of mean force to generate global and per residue estimates. Modelling of the complete S protein amino acid sequence did not yield reliable models (GMQE < 0.65), whereas modelling of the RBD amino acid sequence yielded a reliable model (GMQE > 0.7, QMEAN – 2.18). The template structure selected was the crystal structure of SARS-CoV RBD complexed with a neutralising antibody (PDB ID 2DD8).

The predicted 3D structure was visualised in Visual Molecular Dynamics[28] using sequence alignment and STAMP structural alignments for superimpositions. Structural identity is represented by a Qres (Q) score where Q is measure of structural similarity measuring the fraction of Cα atoms that superimpose. Validation of the homology model was performed using ProCheck[29,30] and ERRAT2[31] on the structural analysis and verification server (SAVES). ProCheck generates a Ramachandran plot which classifies the torsion angles of each residue in the predicted 3D structure into 3 groups to assess its validity. ERRAT2 uses a database of 96 high-resolution protein structures to classify the 6 types of atomic interactions (carbon–carbon, carbon–nitrogen, carbon–oxygen, nitrogen-nitrogen, nitrogen–oxygen, oxygen–oxygen) into correct or incorrectly determined protein regions based on the atomic interactions. This provides an overall quality factor for non-bonded atomic interactions expressed as the percentage of a protein for which the calculated values sits below the set rejection limits, with a score > 50 indicating a high-quality model[31].

Results

Metagenomic analysis revealed an unclassified betacoronavirus in a single sample with genome organisation consistent with *Sarbecovirus*. An initial screening of the 53 samples identified one sample from Gloucestershire with > 650 reads classified to the *Coronavirinae* family, with the positive control spike of Hazara virus detected in 49/53 samples. The percentage of reads classified in total in each sample ranged from 28 to 96% and in both the extraction and library preparation negative controls, no other significant level of reads classified to RNA viruses. Hazara virus was detected in the extraction negative, but not in the library preparation negative.

Classification of the reads in the positive sample identified 0.41% of reads (2550/614,996) as being viral in origin. Of these viral reads, 68% of reads (1668) were classified at species level to bat betacoronavirus BM48-31/BGR/2008 (GenBank reference NC_014470.1)[32]. *De novo* assembly of the Illumina reads generated multiple contigs with homology to members of the subgenus *Sarbecovirus* as assessed via BLASTn; the largest single contig being ~ 7 kb. To investigate, further additional sequence data were generated using Oxford nanopore technology. Using the 562,461 nanopore reads with an average length of ~ 600 kb as scaffold, a hybrid assembly generated a single 29 kb contig similar to sarbecovirus genomes in both size and gene organisation (Fig. 1A,B). To increase confidence in the assembly we performed further Illumina sequencing. With the increased depth, Illumina data alone were assembled into a contig of 21 kb, which was again further assembled to a 29 kb contig with the inclusion of the nanopore data in a hybrid assembly. Mapping all raw reads to this contig shows that

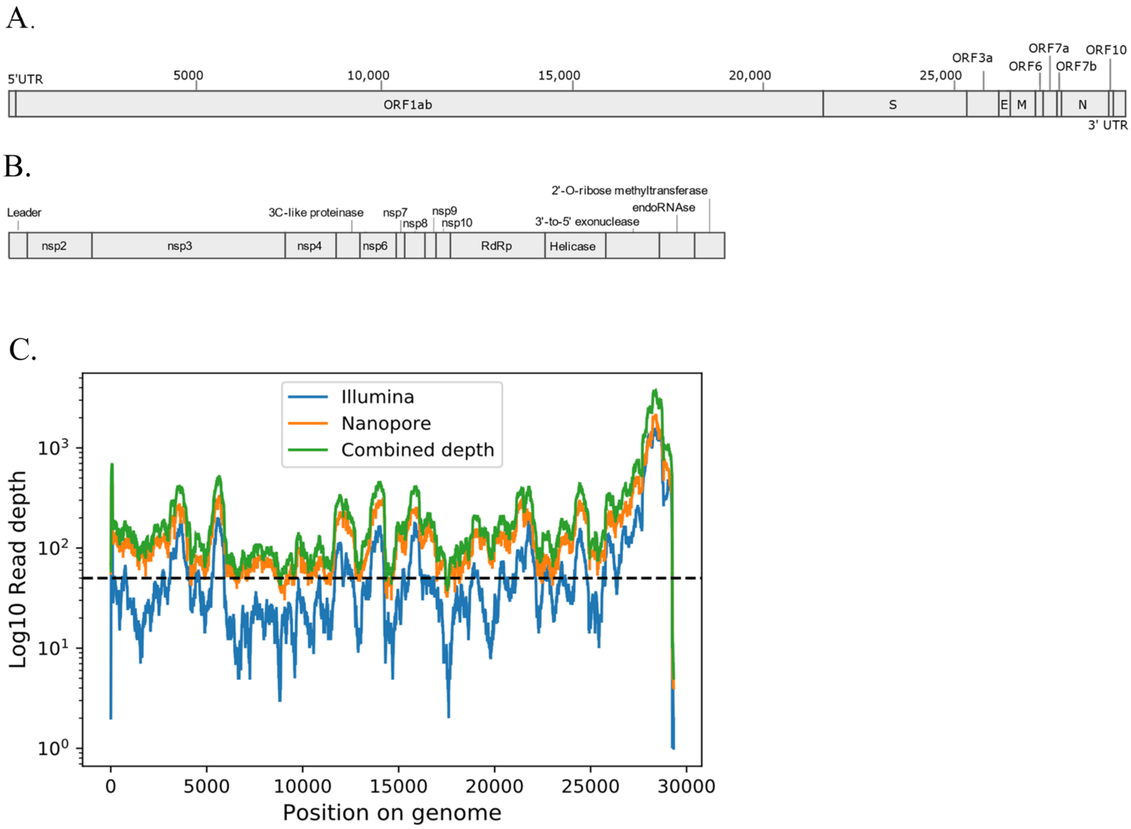


Figure 1. The genomic structure of RhGB01 within the (**A**) entire genome, and (**B**) non-structural proteins. (**C**) Read depth across the genome of RhGB01. Read depth is shown per base across the entire genome from the alignment of Illumina (blue), Nanopore (orange) and combined raw reads (green). The lengths of the genomic features in RhGB01 are 5′UTR (279 bp), ORF1ab (21 kb), S (3.7 kb), ORF3a (813 bp), E (231 bp), M (669 bp), ORF6 (189 bp), ORF7ab (465 bp), N (1.2 kb), ORF10 (78 bp). Total genome length is 29,324 bp.

0.97% and 2.37% map to the contig for Illumina and Nanopore respectively. In total, mean read depth along the assembly is ~ 25 × for Illumina data and 20 × for nanopore data. Combined depth coverage across the assembly is ~ 50 ×, confidently supporting the presence of this virus, in the positive sample (Fig. 1C).

BLASTn analysis of the GenBank nr/nt database shows the assembly shares the highest nucleotide identity with a bat betacoronavirus, BtRs-betaCoV/Hub2013 (KJ473814.1), with 81.01% identity across 85% of the assembly. By comparison it shares 79.78% nucleotide identity across 85% of the assembly with SARS-CoV-2 (isolate SARS-CoV-2/human/USA/FL-CDC-STM-000005640/2021, MW586221.1). We named the virus identified as RhGB01 (*Rhinolophus hipposideros*, Great Britain 01) representing the first detection of a sarbecovirus from *R. hipposideros* in Great Britain.

The genomic structure of RhGB01 mostly mirrors that of other sarbecoviruses with genes encoding non-structural proteins (nsp) in the 5′ region of the genome housed within ORF1ab, and genes encoding structural and accessory proteins at the 3′ region of the genome (Fig. 1A,B)[33]. RhGB01 contains 10 coding genes, whereas SARS-CoV-2, SARS and SARSr-CoV-2 viruses contain 11 coding genes, with the addition of ORF8. In RhGB01, ORF8 and 20 bases in the 5′ region of the ORF7b transcript are absent, comparable to BM48-31/BGR/2008, the closest related virus as determined by phylogenetic analysis (Fig. 2; Supplementary Fig. 1).

Phylogenetic analysis. Maximum likelihood phylogenies of the spike glycoprotein and RdRp (nsp12) nucleotide sequences demonstrate that RhGB01 clusters in a monophyletic clade with BM48-31/BGR/2008, a sarbecovirus isolated from a Blasius's horseshoe bat (*Rhinolophus blasii*) in 2008 in Bulgaria (Fig. 2). This clustering is maintained within phylogenies inferred from nucleotide sequences for all structural and accessory proteins (Supplementary Fig. 1a) and most non-structural proteins (Supplementary Fig. 1b). Phylogenies inferred from the 3C-like proteinase, nsp9, helicase and endoRNAse nucleotide sequences demonstrate alterations in the clustering of RhGB01 into the SARS-CoV clade, into a clade with Rc-o139 and into its own clade, respectively (Supplementary Fig. 1b). In all phylogenies, RhGB01 is distinct from clades containing the human pathogenic betacoronaviruses SARS-CoV and SARS-CoV-2 but, of these, is more closely related to SARSr-CoVs.

The predicted structure of the receptor binding domain of RhGB01 demonstrates structural differences at key residues compared to SARS-CoV and SARS-CoV-2. The major human cellular entry receptor for both SARS-CoV and SARS-CoV-2 is Angiotensin-converting enzyme 2 (hACE2). This bind-

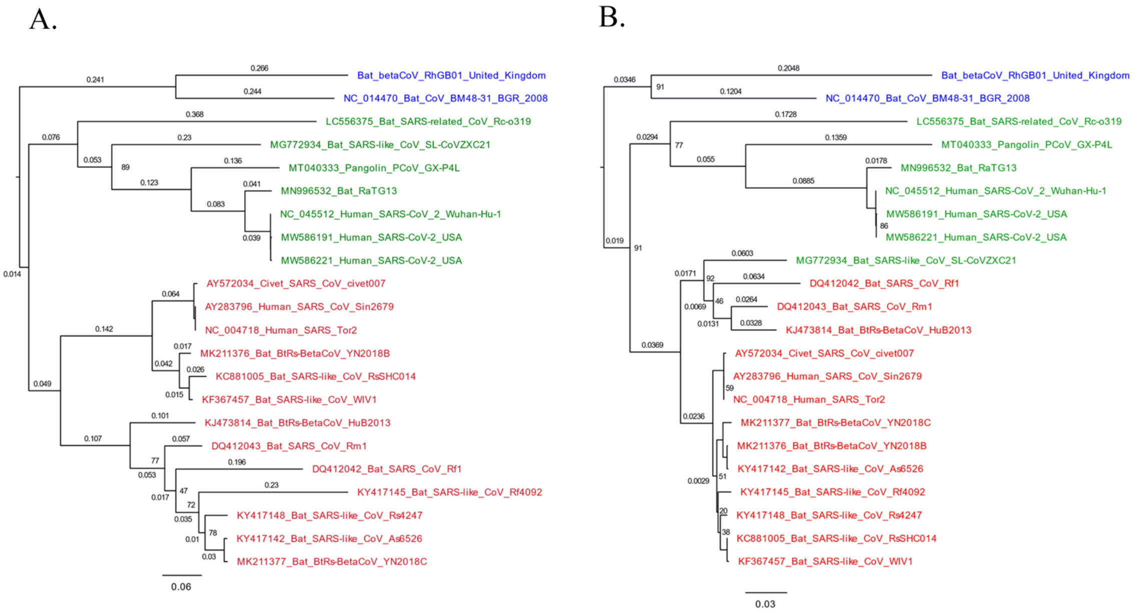


Figure 2. Maximum likelihood phylogenies of the nucleotide sequences for the (**A**) spike glycoprotein and (**B**) RdRp in SARS-CoV, SARS-CoV-2, and related viruses with 1000 bootstrap approximations. Each phylogeny is midpoint rooted and, for visualisation purposes, branch lengths (substitutions per site) in closely related clades and bootstrap supports ≥ 95 are removed. Green taxa are SARS-CoV-2 and related viruses, red taxa are SARS-CoV and related viruses, blue taxa are RhGB01 and BM48-31/GBR-2008.

ing ability is conferred by a receptor binding motif (RBM) within the RBD of the spike glycoprotein. RhGB01 shares amino acid identity of 68% and 67% across the RBD with SARS-CoV and SARS-CoV-2 respectively and just 43% and 48% within the RBM (Fig. 3A). By comparison across the RBD, the closest SARSr-CoV-2 viruses from bat and pangolin hosts share 89% and 86% amino acid homology to SARS-CoV-2 and 75.77% to SARS-CoV (Supplementary Fig. 2). Within the RBM SARSr-CoV-2 viruses from bat and pangolin hosts both share 75% amino acid homology to SARS-CoV-2 and 50% and 49% to SARS-CoV (Fig. 3B). RhGB01 also shows little homology to the RBM of Middle East Respiratory Syndrome (MERS) virus (Fig. 3A). The RhGB01 spike amino acid sequence contains motifs comparable to host transmembrane serine protease 2 cleavage site (TMPRSS2) seen in both SARS-CoV-2 and SARS-CoV in the S2' target site but lacks the additional furin cleavage site specific to SARS-CoV-2 at the S1/S2 intersection (Fig. 3B).

Homology modelling.

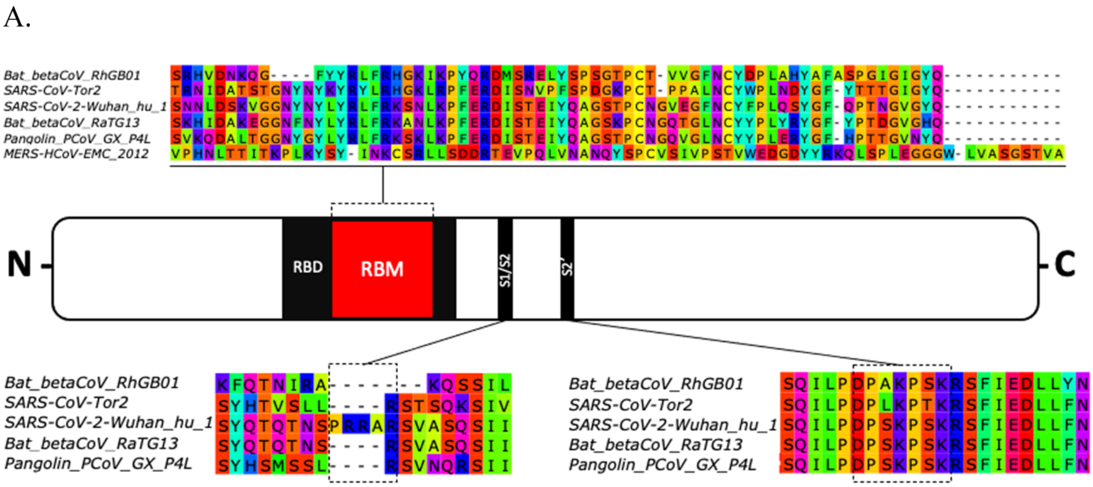
The predicted 3D structure of the RBD contains one α-helix, four 3_{10} helices, three β-bridges and three β-pleated sheets (Supplementary Fig. 3a). It is important to note that the majority of experimentally validated spike glycoprotein and RBD structures are biased towards pathogenic viruses in complex with hACE2 or neutralising antibodies and the accuracy could be improved if the database included more zoonotic spike protein structures.

The predicted 3D structure was validated using ERRAT2 and a Ramachandran plot. The overall quality factor predicted by ERRAT2 is 88.09, supporting the quality of the model. The residues rejected at the 95% and 99% confidence levels are characterised in the literature as contact residues for SARS-CoV to hACE2, highlighting differences within this region compared to the SARS-CoV RBD structure (Supplementary Fig. 3b). A Ramachandran plot predicted that 86% of the residues were in the most favourable region, 13.4% in additional allowed regions and 0.6% in generously allowed regions (Supplementary Fig. 3c).

Structural conservation within the RBD of sarbecoviruses, including RhGB01, is predicted as the RBD is required for viral entry into host cells via the ACE2 receptor. The predicted RBD structure of RhGB01 is more conserved to both SARS-CoV and SARS-CoV-2 than sequence alignment alone indicates, but still differs in regions containing key contact residues for binding hACE2[34–42] (Fig. 4).

Discussion

Here we discovered a novel sarbecovirus (RhGB01), the first to be described in the UK, after sampling just 53 lesser horseshoe bat faecal samples. It is possible that the infection prevalence was even higher and RT-PCR, which would improve sensitivity of detection, should be utilised to screen British bat samples in the future now a sarbecovirus has been detected in the UK. While other sarbecoviruses have been identified in rhinolophid bats in other European countries by polymerase chain reaction and partial gene sequence analyses, RhGB01 is only the second from Europe to be fully sequenced[43], and the first from a lesser horseshoe bat. The only other full sequence betacoronavirus from a European horseshoe bat is BM48-31/BGR/2008 from *R. blasii*. Our results, therefore, extend the geographic and species ranges of SARSr-CoVs and suggest that sarbecoviruses are likely


A.
Bat_betaCoV_RhGB01 SRHVDNKQG----FYYRLFRHGKIKPYQRDMSRELYSPSGTPCT-VVGFNCYDPLAHYAFASPGIGIGYQ----------
SARS-CoV-Tor2 TRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGF-YTTTGIGYQ----------
SARS-CoV-2-Wuhan_hu_1 SNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGF-QPTNGVGYQ----------
Bat_betaCoV_RaTG13 SKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGF-YPTDGVGHQ----------
Pangolin_PCoV_GX_P4L SVKQDALTGGNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGF-HPTTGVNYQ----------
MERS-HCoV-EMC_2012 VPHNLTTITKPLKYSY-INKCSRLLSDDRTEVPQLVNANQYSPCVSIVPSTVWEDGDYYRKQLSPLEGGGW-LVASGSTVA
N
RBD
RBM
S1/S2
S2'
C
Bat_betaCoV_RhGB01 KFQTNIRA-------KQSSIL
SARS-CoV-Tor2 SYHTVSLL----RSTSQKSIV
SARS-CoV-2-Wuhan_hu_1 SYQTQTNSPRRARSVASQSII
Bat_betaCoV_RaTG13 SYQTQTNS----RSVASQSII
Pangolin_PCoV_GX_P4L SYHSMSSL----RSVNQRSII
Bat_betaCoV_RhGB01 SQILPDPAKPSKRSFIEDLLYN
SARS-CoV-Tor2 SQILPDPLKPTKRSFIEDLLFN
SARS-CoV-2-Wuhan_hu_1 SQILPDPSKPSKRSFIEDLLFN
Bat_betaCoV_RaTG13 SQILPDPSKPSKRSFIEDLLFN
Pangolin_PCoV_GX_P4L SQILPDPSKPSKRSFIEDLLFN

B.

	Receptor binding motif		Receptor binding domain	
	SARS-CoV	SARS-CoV-2	SARS-CoV	SARS-CoV-2
RhGB01	43.48	48.57	68.21	67.35
RaTG13	50.72	75.36	75.77	89.18
PCoV-GX-P4L	49.28	75.36	75.77	86.80

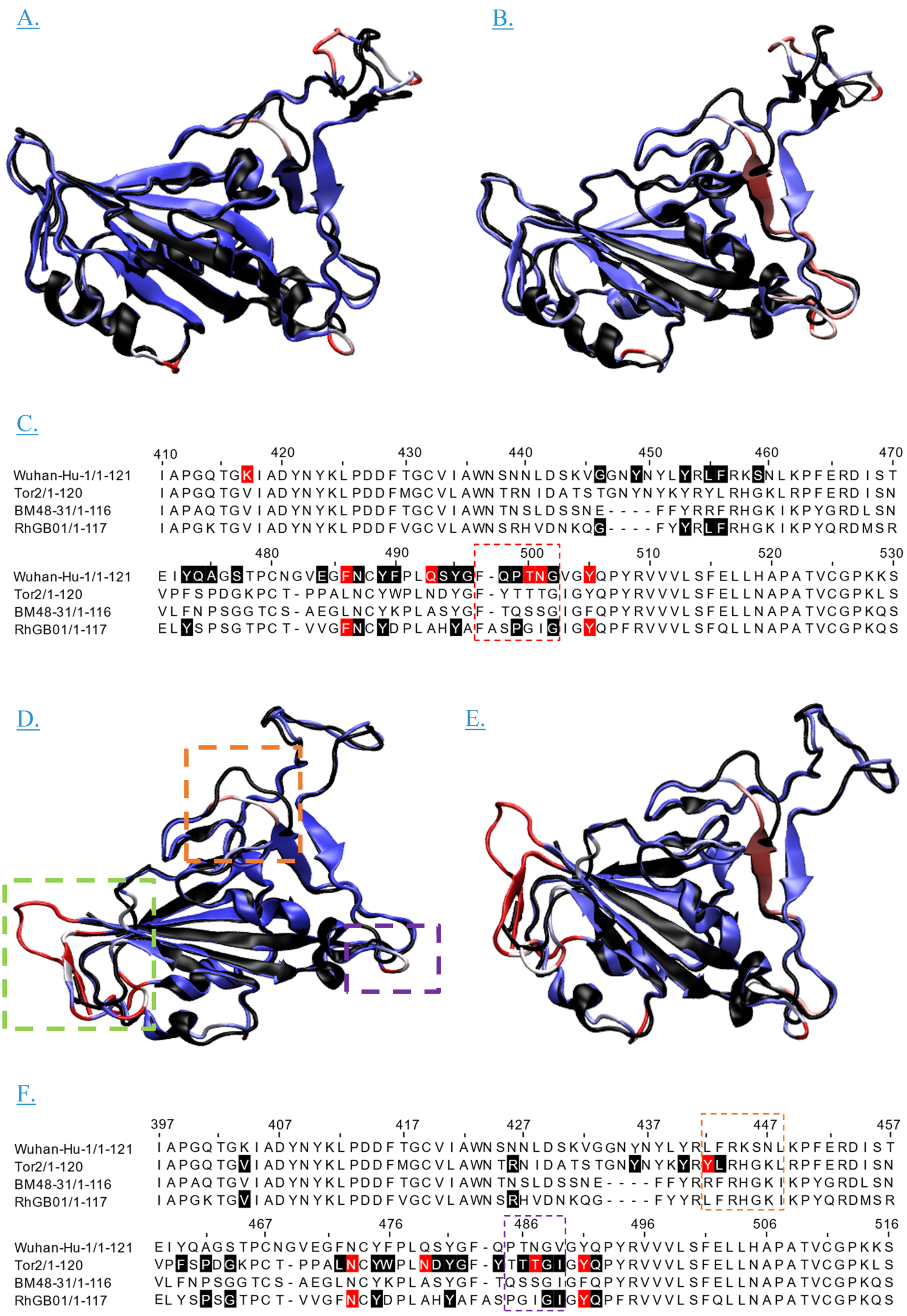
Figure 3. A schematic representation of the entire S protein with the RBM and cleavage sites highlighted and compared between RhGB01, Pangolin and Bat derived related viruses, SARS-CoV and SARS-CoV-2 and percentage identity values for the RBD and RBM regions. Amino acids are coloured according to the Taylor colour scheme. (**A**) RBM comparison demonstrates higher percentages of homology within sarbecoviruses from other zoonotic hosts. The RBM of MERS-HCoV-EMC-2012, subgenus *Merbecovirus,* demonstrates little amino acid homology to RBM from the *Sarbecovirus* subgenus. The furin cleavage site (S1/S2) is only present in SARS-CoV-2, distinct from the TMPRSS2 cleavage motif (S2′), which is more conserved within the sarbecoviruses. (**B**) Percentage identity scores for the RBM and RBD to SARS-CoV and SARS-CoV-2. RaTG13 and PCoV represent the most closely SARSr-CoV-2 virus from zoonotic hosts.

to be present throughout the range of the Rhinolophidae, which are distributed from Australia and Japan to Europe and Africa.

The range of the lesser horseshoe bat extends from Western Europe to Central Asia, overlapping with those of other rhinolophid species, including the greater horseshoe bat (*R. ferrumequinum*), which ranges from Western Europe to Japan[44,45]. Where they co-exist, the species can be syntopic allowing opportunity for cross-species virus transfer. Prior to our results, the observed and predicted (cut off ≥ 0.9821) number of coronaviruses in the greater horseshoe bat were 13 and 19 respectively, and in the lesser horseshoe bat these figures were zero and three respectively[46]. This suggests that the complement of *Sarbecovirus* species in horseshoe bats is greater than predicted so far, with the possibility of virus sharing across species and large geographic areas.

Genomic alignments between RhGB01 and related sarbecoviruses highlight key genomic differences between RhGB01 and known zoonotic sarbecoviruses. Host specificity is dependent on the ability of a virus to attach to host receptors and enter host cells; a binding process facilitated by contact residues contained within the RBM[37]. RhGB01 demonstrates low amino acid homology to SARS-CoV and SARS-CoV-2 in the RBM compared to that between SARS-CoV, SARS-CoV-2 and the closely related bat (RaTG13) and pangolin (PCoV_GX_P4L) sequences identified in Asia[48,49]. The low level of homology, lack of contact residues and structural differences compared to the RBD of SARS-CoV and SARS-CoV-2 most likely indicate a lack of ability to bind hACE2 and, hence, RhGB01 is unlikely to be zoonotic without mutation. To experimentally validate the absence of hACE2 or other human cell receptor binding and to identify binding abilities to other mammalian ACE2 receptors, in vitro binding assays are required.

Aside from the variation observed in amino acid homology within the RBM, RhGB01 also exhibits variation within the furin cleavage site and ORF8 when compared to zoonotic sarbecoviruses. The absence of the furin cleavage site indicates the absence of enhanced efficiency of host cell entry observed with SARS-CoV-2[50]. However, RhGB01 does retain a similar motif responsible for cleavage in the S2' region by host transmembrane serine protease 2, also required for spike protein proteolytic priming for ACE2 attachment[47]. SARS-CoV-2 variants with a functional ORF8 are associated with greater pathogenicity, thought to be due to downregulation of major histocompatibility complex class 1 (MHC 1), and thus a reduction in antigen presentation to CD8$^+$ T



```
                  410       420       430       440       450       460       470
Wuhan-Hu-1/1-121  IAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDIST
Tor2/1-120        IAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISN
BM48-31/1-116     IAPAQTGVIADYNYKLPDDFTGCVIAWNTNSLDSSNE----FFYRRFRHGKIKPYGRDLSN
RhGB01/1-117      IAPGKTGVIADYNYKLPDDFVGCVLAWNSRHVDNKQG----FYYRLFRHGKIKPYQRDMSR

                     480       490       500       510       520       530
Wuhan-Hu-1/1-121  EIYQAGSTPCNGVEGFNCYFPLQSYGF-QPTNGVGYQPYRVVVLSFELLHAPATVCGPKKS
Tor2/1-120        VPFSPDGKPCT-PPALNCYWPLNDYGF-YTTTGIGYQPYRVVVLSFELLNAPATVCGPKLS
BM48-31/1-116     VLFNPSGGTCS-AEGLNCYKPLASYGF-TQSSGIGFQPYRVVVLSFELLNAPATVCGPKQS
RhGB01/1-117      ELYSPSGTPCT-VVGFNCYDPLAHYAFASPGIGIGYQPFRVVVLSFQLLNAPATVCGPKQS
```

```
                  397       407       417       427       437       447       457
Wuhan-Hu-1/1-121  IAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDIST
Tor2/1-120        IAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISN
BM48-31/1-116     IAPAQTGVIADYNYKLPDDFTGCVIAWNTNSLDSSNE----FFYRRFRHGKIKPYGRDLSN
RhGB01/1-117      IAPGKTGVIADYNYKLPDDFVGCVLAWNSRHVDNKQG----FYYRLFRHGKIKPYQRDMSR

                     467       476       486       496       506       516
Wuhan-Hu-1/1-121  EIYQAGSTPCNGVEGFNCYFPLQSYGF-QPTNGVGYQPYRVVVLSFELLHAPATVCGPKKS
Tor2/1-120        VPFSPDGKPCT-PPALNCYWPLNDYGF-YTTTGIGYQPYRVVVLSFELLNAPATVCGPKLS
BM48-31/1-116     VLFNPSGGTCS-AEGLNCYKPLASYGF-TQSSGIGFQPYRVVVLSFELLNAPATVCGPKQS
RhGB01/1-117      ELYSPSGTPCT-VVGFNCYDPLAHYAFASPGIGIGYQPFRVVVLSFQLLNAPATVCGPKQS
```

◂**Figure 4.** Amino acid sequence and STAMP structural alignments of the RBD of SARS-CoV and SARS-CoV-2 compared to a predicted 3D structure of the RBD of RhGB01 highlighting structural differences in regions housing hACE2 contact residues for SARS-CoV and SARS-CoV-2. Contact residues are described in the literature (black), with residues in red the most identified residues. Regions coloured in purple have high structural similarity (> 0.7) and residues with a low structural similarity coloured in red/white (< 0.3). BM48-31 is included as comparison to a second Sarbecovirus from a bat source. (**A**) STAMP alignment of SARS-CoV-2 RBD (PDB ID 6M0J) in the conformation of being complexed with hACE2 aligned with the RhGB01 RBD. The areas where structural similarity is low are seen to contain contact residues defined in the literature. (**B**) Sequence alignment of RhGB01 and SARS-CoV-2. The sequence used is identical to (**C**). (**C**) Sequence alignment of the RBM with 28 residues up and downstream, to facilitate the inclusion of K417. The coordinates provided refer the SARS-CoV-2 (NC_45512.2) residue position within the spike protein alignment. (**D**) STAMP alignment of SARS-CoV RBD (PBD ID 2DD8) aligned with the predicted RhGB01 RBD. The region highlighted in green are residues upstream and downstream of the structural alignment which do not align. (**E**) Sequence alignment of RhGB01 and SARS-CoV. (**F**) Sequence alignment of RBM with SARS-CoV contact residues highlighted, and the coordinates provided refer to the residue position for SARS-CoV (NC_004718.3) within the spike protein alignment. 28 residues are included up and downstream of the RBM to include V404.

lymphocytes which facilitates prolonged infection[51,52]. The absence of ORF8 from the genome of RhGB01 suggests that this virus lacks these immune evasion properties.

It has been postulated that both SARS-CoV and SARS-CoV-2 evolved through mutation, possibly involving homologous recombination, during passage through at least one intermediate host; probably civets (viverrids) or mustelids for SARS-CoV[53,54] and possibly a species of pangolin for SARS-CoV-2[49]. In this way, the progenitor virus from the natural host (a species of horseshoe bat) gained genetic adaptations to allow successful infection of, and transmission between, human beings. Where there is opportunity for homologous recombination of sarbecoviruses through co-infection, there is the possibility of novel zoonotic emergence. Thus, co-infection of horseshoe bats with their natural suites of coronaviruses and with SARS-CoV-2 could lead to the development of novel zoonotic emergence. While there is a need to increase surveillance for coronaviruses in horseshoe bats across their range, and also in other bat species, especially those syntopic with, or closely related to, horseshoe bats (e.g. the Old World leaf-nosed bats, family Hipposideridae), it is also important that steps are taken to minimise opportunities of virus transmission between novel hosts.

In Europe, unlike in Asia, direct contact between people and bats most commonly occurs when the animals are captured by bat researchers or when sick animals are taken in by bat rescuers and wildlife rehabilitation centres. While the risk of reverse spill over of SARS-CoV-2 from researchers to bats and onward spread within bat populations has been shown to be medium to high[55], it is the caring of sick or injured bats, in particular, that provides most opportunity for long-term close contact and virus transfer in either direction. Although the IUCN Bat Specialist Group has produced guidelines to minimise this risk[56], the degree to which these are known or followed is unclear. Our findings highlight that the natural distribution of sarbecoviruses and opportunities for recombination through intermediate host co-infection are underestimated. Preventing transmission of SARS-CoV-2 to horseshoe bats, with the risk this presents of further mutation, is of particular significance with the current roll out of a global mass vaccination campaign against this virus.

Received: 24 March 2021; Accepted: 29 June 2021
Published online: 19 July 2021

References

1. WHO. *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*. https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19). Accessed 24 Feb 2021.
2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).
3. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
4. Li, W. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science* **310**, 676–679 (2005).
5. Lau, S. K. P. *et al.* Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. USA* **102**, 14040–14045 (2005).
6. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
7. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
8. Wacharapluesadee, S. *et al.* Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat. Commun.* **12**, 972 (2021).
9. Luk, H. K. H., Li, X., Fung, J., Lau, S. K. P. & Woo, P. C. Y. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect. Genet. Evol.* **71**, 21–30 (2019).
10. Murakami, S. *et al.* Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2, Japan. *Emerg. Infect. Dis.* **26**, 3025–3029 (2020).
11. Hul, V. *et al.* A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *BioRxiv.* https://doi.org/10.1101/2021.01.26.428212 (2021).
12. Lewandowski, K. *et al.* Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples. *J. Clin. Microbiol.* **58**, 19 (2019).
13. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
14. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

15. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
17. Li, H. *Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM*. http://arxiv.org/abs/1303.3997 (2013).
18. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
19. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
20. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
21. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
22. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2: A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
23. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
24. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
25. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
26. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
27. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, 296–303 (2018).
28. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph* **14**, 33–38 (1996).
29. Laskowski, R., Macarthur, M. W., Moss, D. S. & Thornton, J. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
30. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486 (1996).
31. Colovos, C. & Yeates, T. O. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511–1519 (1993).
32. Drexler, J. F. *et al.* Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. *J. Virol.* **84**, 11336–11349 (2010).
33. Nakagawa, K., Lokugamage, K. G. & Makino, S. Viral and cellular mRNA translation in coronavirus-infected cells. *Adv. Virus Res.* **96**, 165–192 (2016).
34. Yan, H. *et al.* ACE2 receptor usage reveals variation in susceptibility to SARS-CoV and SARS-CoV-2 infection among bat species. *Nat. Ecol. Evol.* **5**, 600–608 (2021).
35. Zhai, X. *et al.* Comparison of severe acute respiratory syndrome coronavirus 2 spike protein binding to ACE2 receptors from human, pets, farm animals, and putative intermediate hosts. *J. Virol.* **94**, 15 (2020).
36. Rodriguez, J. H. & Gupta, A. Contact residue contributions to interaction energies between SARS-CoV-1 spike proteins and human ACE2 receptors. *Sci. Rep.* **11**, 1156 (2021).
37. Ali, A. & Vijayan, R. Dynamics of the ACE2-SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. *Sci. Rep.* **10**, 14214 (2020).
38. Fraguas Bringas, C. & Booth, D. Identification of a SARS-like bat coronavirus that shares structural features with the spike glycoprotein receptor-binding domain of SARS-CoV-2. *Access Microbiol.* **2**, 000166 (2020).
39. Fang, S. *et al.* Key residues influencing binding affinities of 2019-nCoV with ACE2 in different species. *Brief Bioinform.* **22**, 963–975 (2021).
40. Delgado, J. M. *et al.* Molecular basis for higher affinity of SARS-CoV-2 spike RBD for human ACE2 receptor. *Proteins* https://doi.org/10.1002/prot.26086 (2021).
41. Preziuso, S. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) exhibits high predicted binding affinity to ACE2 from lagomorphs (rabbits and pikas). *Animals* **10**, 1460 (2020).
42. Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448 (2020).
43. ArGouilh, M. *et al.* SARS-CoV related Betacoronavirus and diverse Alphacoronavirus members found in western old-world. *Virology* **517**, 88–97 (2018).
44. Taylor, P. *Rhinolophus hipposideros. The IUCN Red List of Threatened Species 2016: e.T19518A21972794*. https://doi.org/10.2305/IUCN.UK.2016-2.RLTS.T19518A21972794.en. Accessed 24 Feb 2021.
45. Piraccini, R. *Rhinolophus ferrumequinum. The IUCN Red List of Threatened Species* 2016: e.T19517A21973253. https://doi.org/10.2305/IUCN.UK.2016-2.RLTS.T19517A21973253.en. Accessed 24 Feb 2021.
46. Wardeh, M., Baylis, M. & Blagrove, M. S. C. Predicting mammalian hosts in which novel coronaviruses can be generated. *Nat. Commun.* **12**, 780 (2021).
47. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
48. Zhou, H. *et al.* A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Curr. Biol.* **30**, 2196-2203.e3 (2020).
49. Lam, T.T.-Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285 (2020).
50. Papa, G. *et al.* Furin cleavage of SARS-CoV-2 Spike promotes but is not essential for infection and cell-cell fusion. *PLOS Pathog.* **17**, e1009246 (2021).
51. Young, B. E. *et al.* Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: An observational cohort study. *Lancet* **396**, 603–611 (2020).
52. Zhang, Y. *et al.* The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. *Proc. Natl. Acad. Sci. U.S.A.* **118**(23), e2024202118 (2021).
53. Guan, Y. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278 (2003).
54. Bell, D., Roberton, S. & Hunter, P. Animal origins of SARS Coronavirus: Possible links with the international trade in small carnivores. *Philos. Trans. R. Soc. Lond. B* **359**, 1107–1114 (2004).
55. Common, S. M., Shadbolt, T., Walsh, K. & Sainsbury, A. W. The risk from SARS-CoV-2 to bat species in England and mitigation options for conservation field workers. *Transbound Emerg. Dis.* https://doi.org/10.1111/tbed.14035 (2021).
56. Jolliffe, T. *et al. IUCN SSC Bat Specialist Group (BSG) recommendations to reduce the risk of transmission of SARS-CoV-2 from humans to bats in bat rescue and rehabilitation centers: MAP: Minimize, Assess, Protect*. (2020). https://www.iucnbsg.org/bsg-publications.html

Acknowledgements

We thank Daniel Whitby, Animal Ecology and Wildlife Consultants Limited, for logistical support. This project was funded by Public Health England, the University of East Anglia and Research England. The views expressed in this publication are those of the authors and not the views of the National Health Service, the Department of Health and Social Care, Public Health England or any other institution.

Author contributions

I.M., D.B. & A.A.C. devised the study. I.M., J.C., and D.C. conducted the study, J.C. & S.P. analysed the sequence data. M.C., R.V., D.B. & A.A.C. obtained the funding. A.A.C., J.C., S.P., I.M. & D.B. wrote the manuscript, and all authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-94011-z.

Correspondence and requests for materials should be addressed to A.A.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.