Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/sciadv.abb9153/DC1

Supplementary Materials for

Emergence of SARS-CoV-2 through recombination and strong purifying selection

Xiaojun Li, Elena E. Giorgi, Manukumar Honnayakanahalli Marichannegowda, Brian Foley, Chuan Xiao, Xiang-Peng Kong, Yue Chen, S. Gnanakaran, Bette Korber, Feng Gao*

*Corresponding author. Email: fgao@duke.edu

Published 29 May 2020, *Sci. Adv.* **6**, eabb9153 (2020)
DOI: 10.1126/sciadv.abb9153

This PDF file includes:

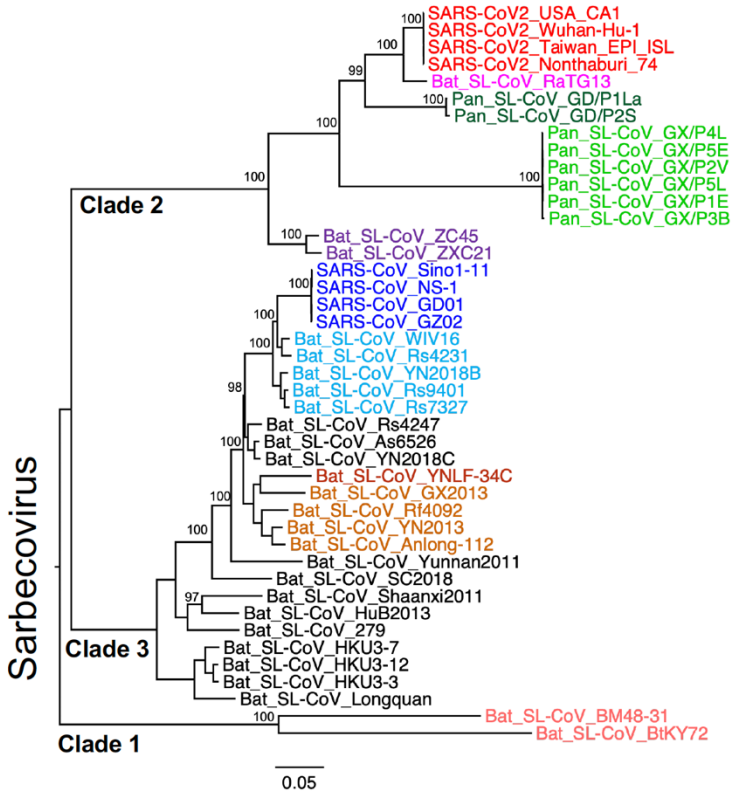

Fig. S1. Phylogenetic tree of the complete CoV genome sequences. All 43 sequences used in this study includes: 4 SARS-CoV-2 sequences (red), Bat_SL-CoV sequence RaTG13 (magenta), 2 pangolin CoV from Guangdong (Pan_SL-CoV_GD, dark green), 6 pangolin CoV from Guangxi (Pan_SL-CoV_GX , light green), and 4 SARS-CoV sequences (dark blue). The remaining Bat_SL-CoV sequences in the set are color-coded according to their phylogenetic subclusterings in the tree. Phylogenetic trees were constructed by the maximum likelihood method using the GTR model, and their reliability was estimated from 1,000 bootstrap replicates.

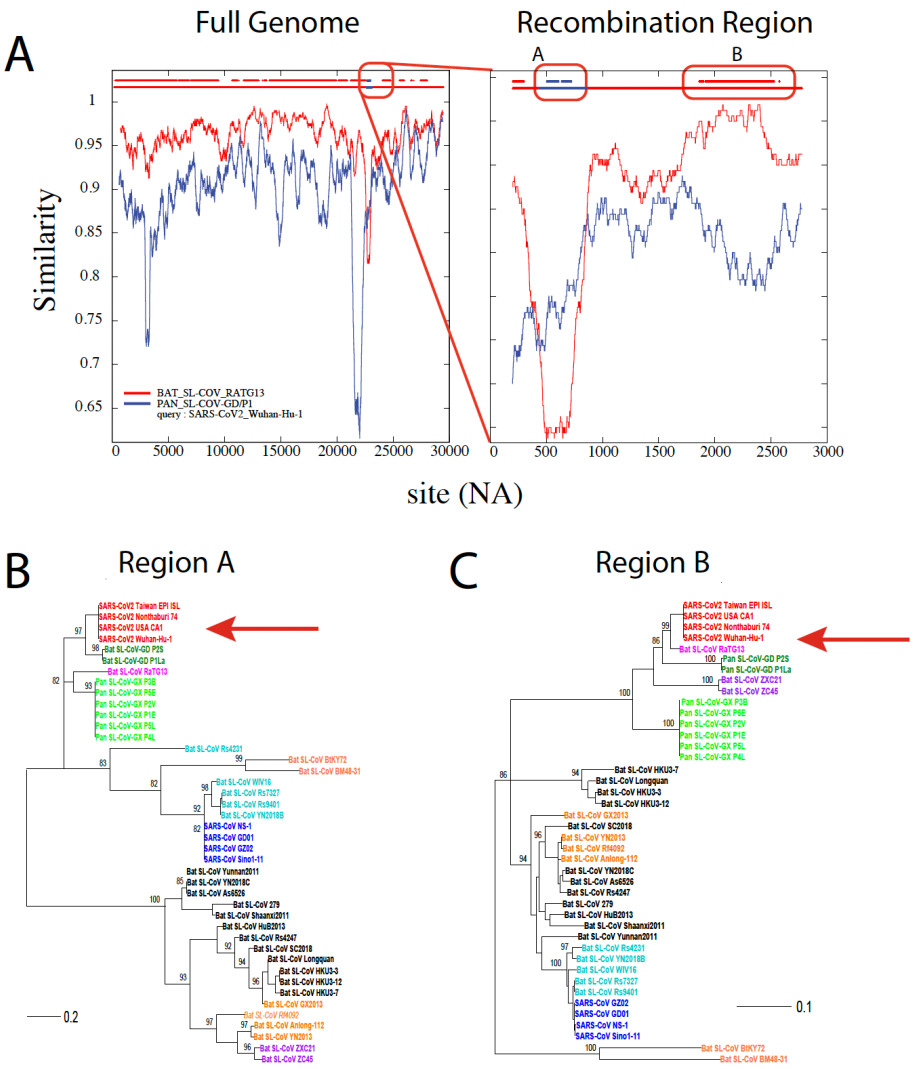

Fig. S2. Recombination analysis of the CoV-SARS-2 Wuhan-Hu-1 sequence. (A) Similarity plots comparing the Wuhan-Hu-1 sequence to the bat-Cov RaTG13 sequence (red) and the Pan_SL-CoV_GD/1PL sequence (blue). Plots were obtained using the LANL tool RIP using a window size of 400 bp. The full genome comparison is shown on the right and, on the left, a close-up of the recombination region is shown. Blue and red horizontal lines at the top of the panels show recombination breakpoints at the 99% confidence level. The blank between the A and B regions means uncertainty. (B and C) Phylogenetic trees of the individual recombination regions A and B, showing the different clusterings of the CoV-SARS-2 sequence compared to RaTG13 and Pan_SL-CoV_GD/1PLa (highlighted by the red arrows).

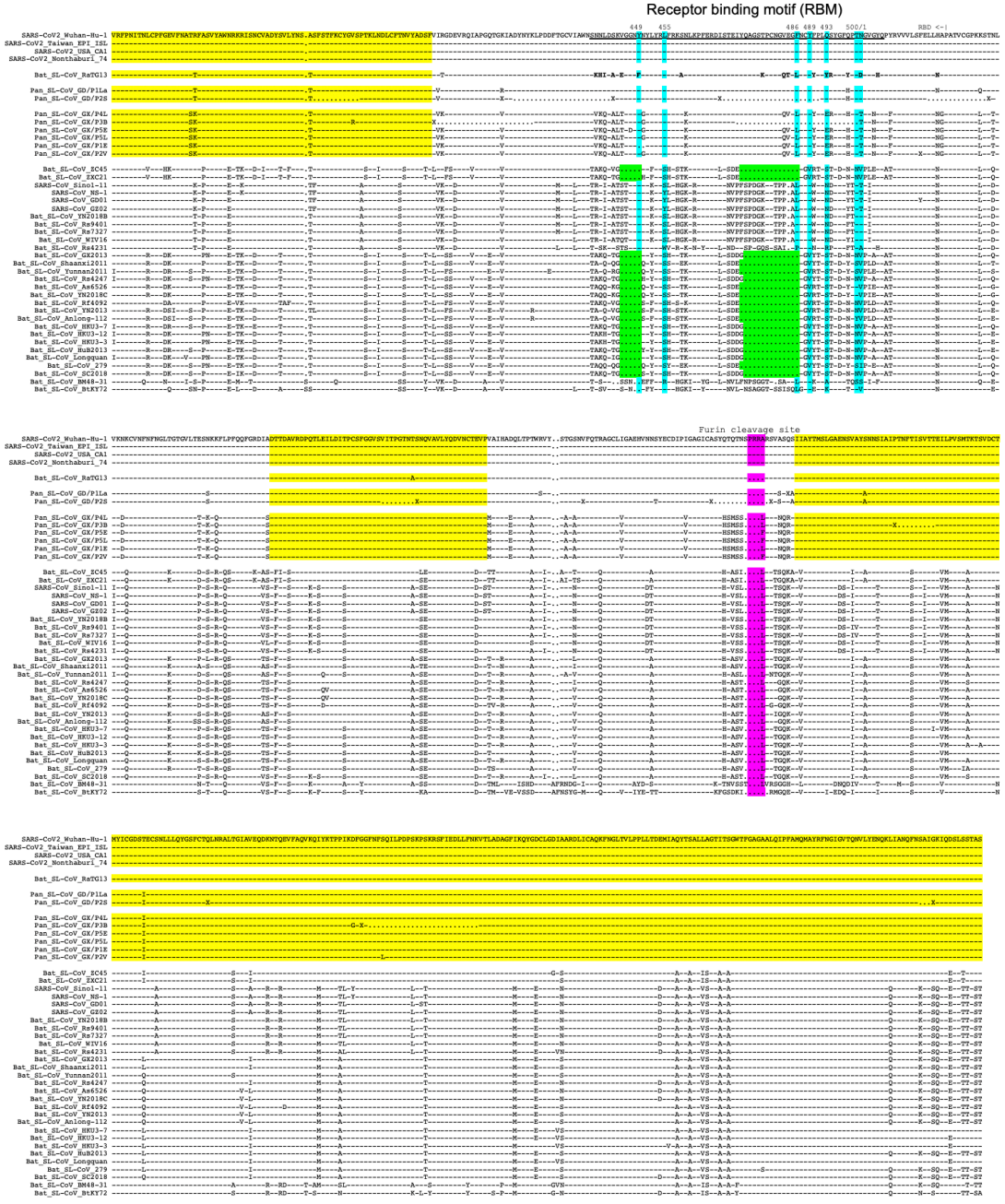

Fig. S3. Highly conserved sequences around the receptor binding motif and furin cleavage sites among SARS-CoV-2, RaTG13 and Pan_SL-CoV viruses. Alignment of amino acid sequences around receptor binding motif (RBM) and furin cleavage sites in the spike glycoprotein compared to Wuhan-Hu-1 (top sequence, na 22541-24391). Identical amino acids are shown as dashes and deletions as dots. RBM is shown at aa positions 439-508, and the furin cleavage site is highlighted in magenta. Regions with identical or nearly identical amino acid sequences among SARS-CoV-2, RaTG13 and Pan_SL-CoV viruses are highlighted in yellow. The positions of critical contact sites with ACE2 are indicated at the top of the alignment and highlighted in blue. The two large deletions in RBM are indicated in green.

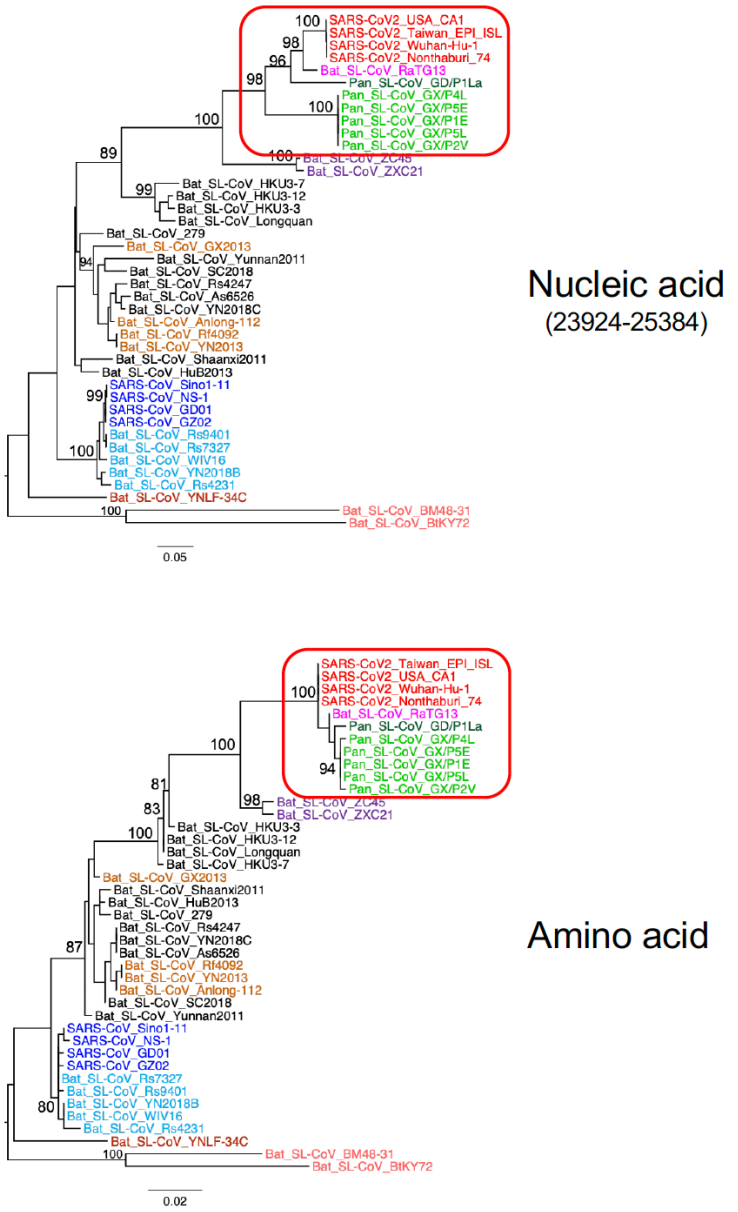

Fig. S4. Purifying selection in the 3' end region of the S gene. Purifying selection pressure on the 3' end region (na 23924-25384) of the S gene region among SARS-CoV-2, RaTG13 and Pan_SL-CoV viruses (within red boxes in phylogenetic trees) are shown by much shorter branches with amino acid sequences than with nucleic acid sequences.

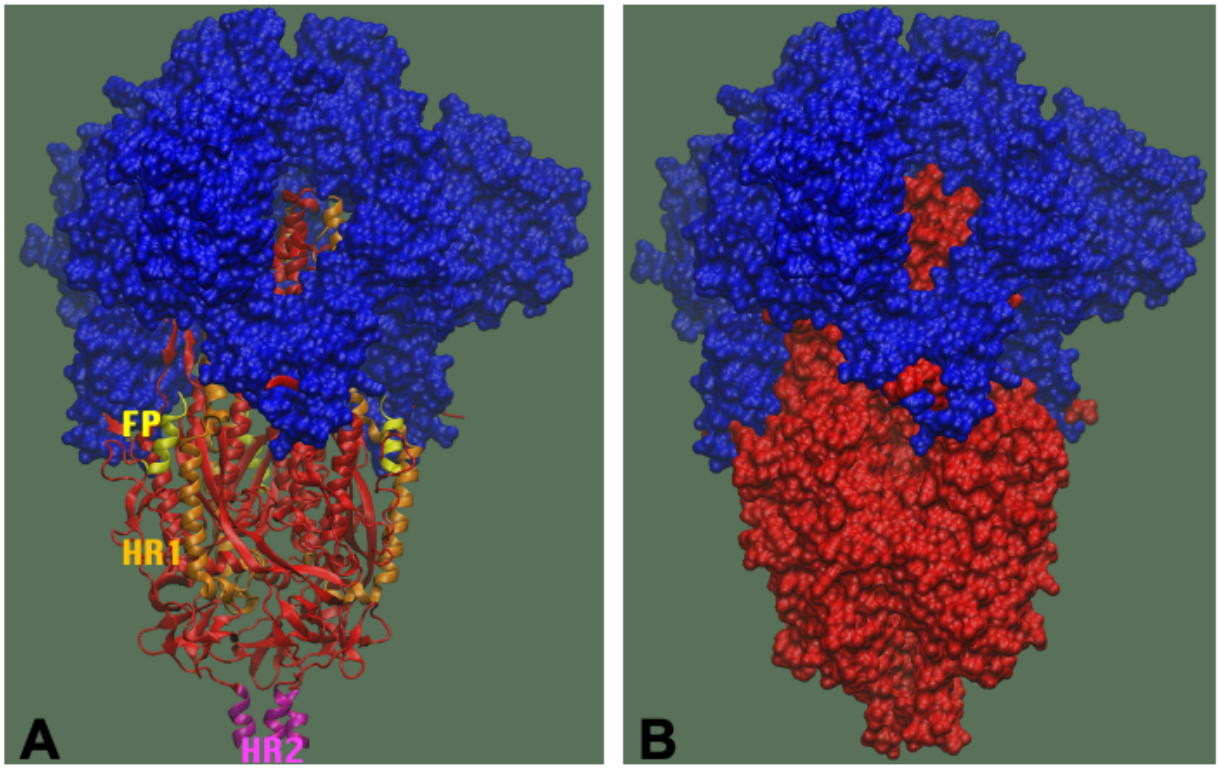

Fig. S5. The conformational architecture of Spike trimer in the closed (all RBD down) form (PDB: 6VXX). The S1 and S2 subunits are colored in blue and red, respectively. The separation of S1 and S2 subunits are made at the polybasic furin cleavage site (RRAR). (A) Cartoon like rendering (red) captures the secondary structures of the S2 subunit that take part in fusion. Fusion peptide (yellow) and HR1 (orange) are labeled. The HR2 (magenta) continues at the bottom and not considered in the cryo-EM structure of SARS CoV2. HR2 potentially adopts a symmetric coiled coil trimer similar to SARS CoV1. (B) Surface map indicates that regions of S2 subunit are covered by the S1 subunit.

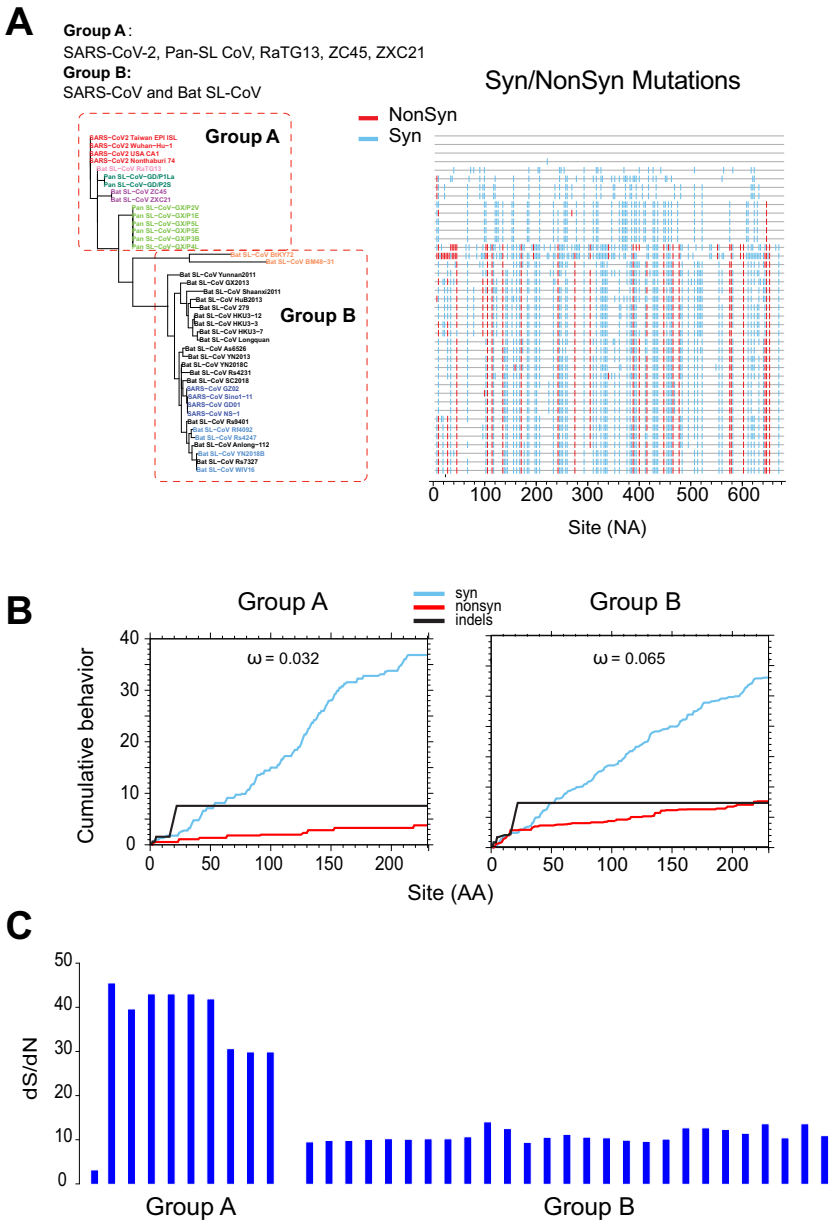

Fig. S6. Purifying selection pressure on the M gene. (A) Phylogenetic tree (left) and Highlighter plot (right) of all sequences compared to SARS-CoV-2 in the M gene. SARS-CoV-2, RaTG13, all Pan_SL-CoV and the two bat CoV (ZXC21 and ZC45) sequences are in Group A, and all other sequences in Group B, to highlight differences between the two groups. Colored tic marks are mutations compared to the top sequence (SARS-CoV-2 Wuahn-Hu-1), with synonymous as light blue and non-synonymous as red. (B) Cumulative plots of the average behavior of each codon for all pairwise comparisons in the input data, for synonymous mutations, non-synonymous mutations and indels of group A sequences (left) and group B sequences (right). Average ratios of the rate of nonsynonymous substitutions per nonsynonymous site (dN/dS, or ω) for each sequence group are reported at the top of each plot. (C) dS/dN ratios of all sequences compared to Wuhan-Hu-1.

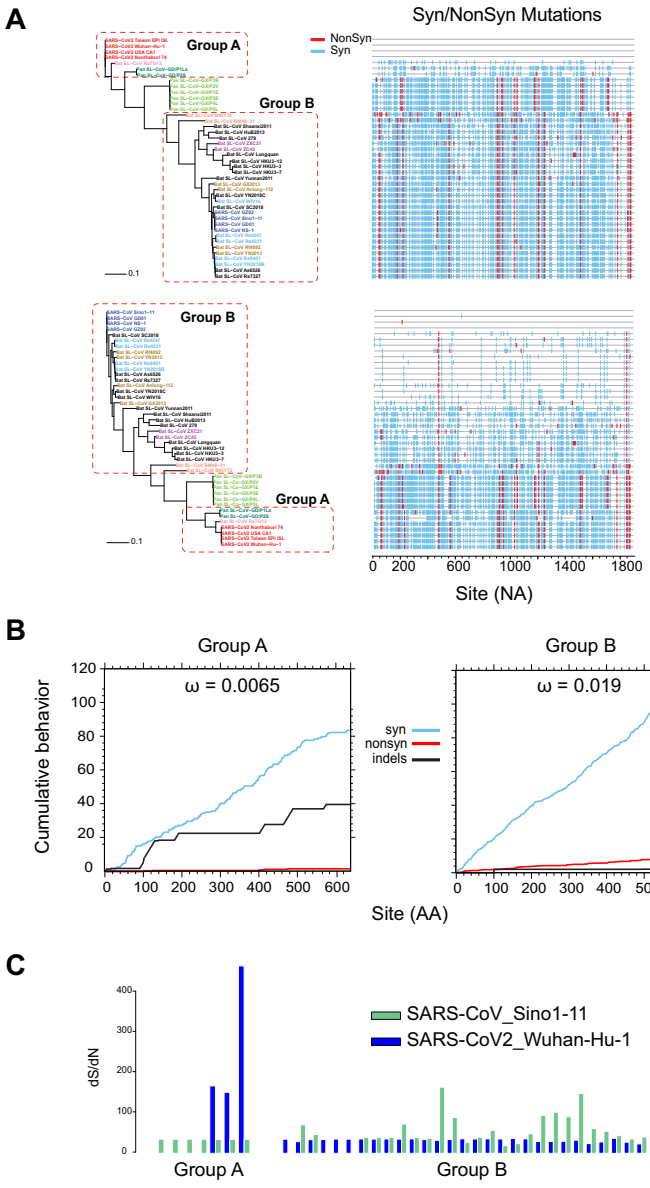

Fig. S7. Purifying selection pressure on the partial region of ORF1a. (A) Phylogenetic trees (left) and Highlighter plots (right) of sequences compared to SARS-CoV-2 (top) and to SARS-CoV (bottom) in the partial region of ORF1a. SARS-CoV-2, RaTG13 and Pan_SL-CoV from Guangdon are in Group A, and all other bat-CoV sequences in Group B, to highlight differences between the two groups. Colored tic marks are mutations compared to the top sequence (SARS-CoV-2 Wuahn-Hu-1 in the top graph and SARS-CoV Sin 1-11 in the bottom graph), with synonymous as light blue and non-synonymous as red. (B) Cumulative plots of the average behavior of each codon for all pairwise comparisons in the input data, for synonymous mutations, non-synonymous mutations and indels of group A sequences (left) and group B sequences (right). Average ratios of the rate of nonsynonymous substitutions per nonsynonymous site (dN/dS, or ω) for each sequence group are reported at the top of each plot. (C) dS/dN ratios of all sequences compared to Wuhan-Hu-1 in dark blue, and compared to SARS-CoV Sin 1-11 in green.

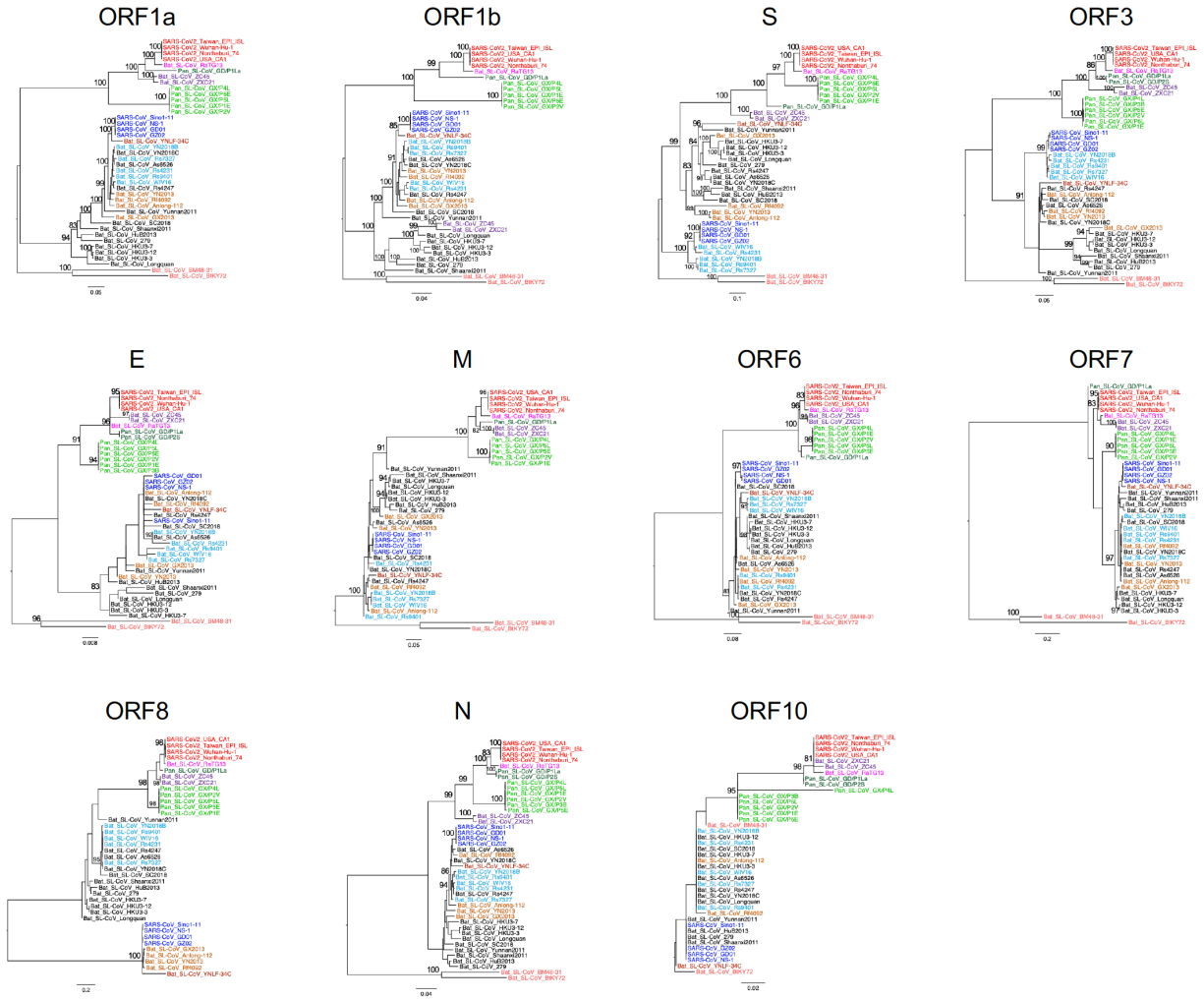

Fig. S8. Phylogenetic tree analysis of SARS-CoV-2 genes together with other CoVs. Phylogenetic trees were constructed for each coding region in the CoV genome. Sequences are colored differently based on their hosts and phylogenetic cluster: 4 SARS-CoV-2 sequences (red), Bat_SL-CoV sequence RaTG13 (magenta), 2 pangolin CoVs from Guangdong (Pan_SL-CoV_GD, dark green), 6 pangolin CoVs from Guangxi (Pan_SL-CoV_GX , light green), and 4 SARS-CoV sequences (dark blue). The remaining Bat_SL-CoV sequences in the set are color-coded according to their phylogenetic subclusterings in the tree.

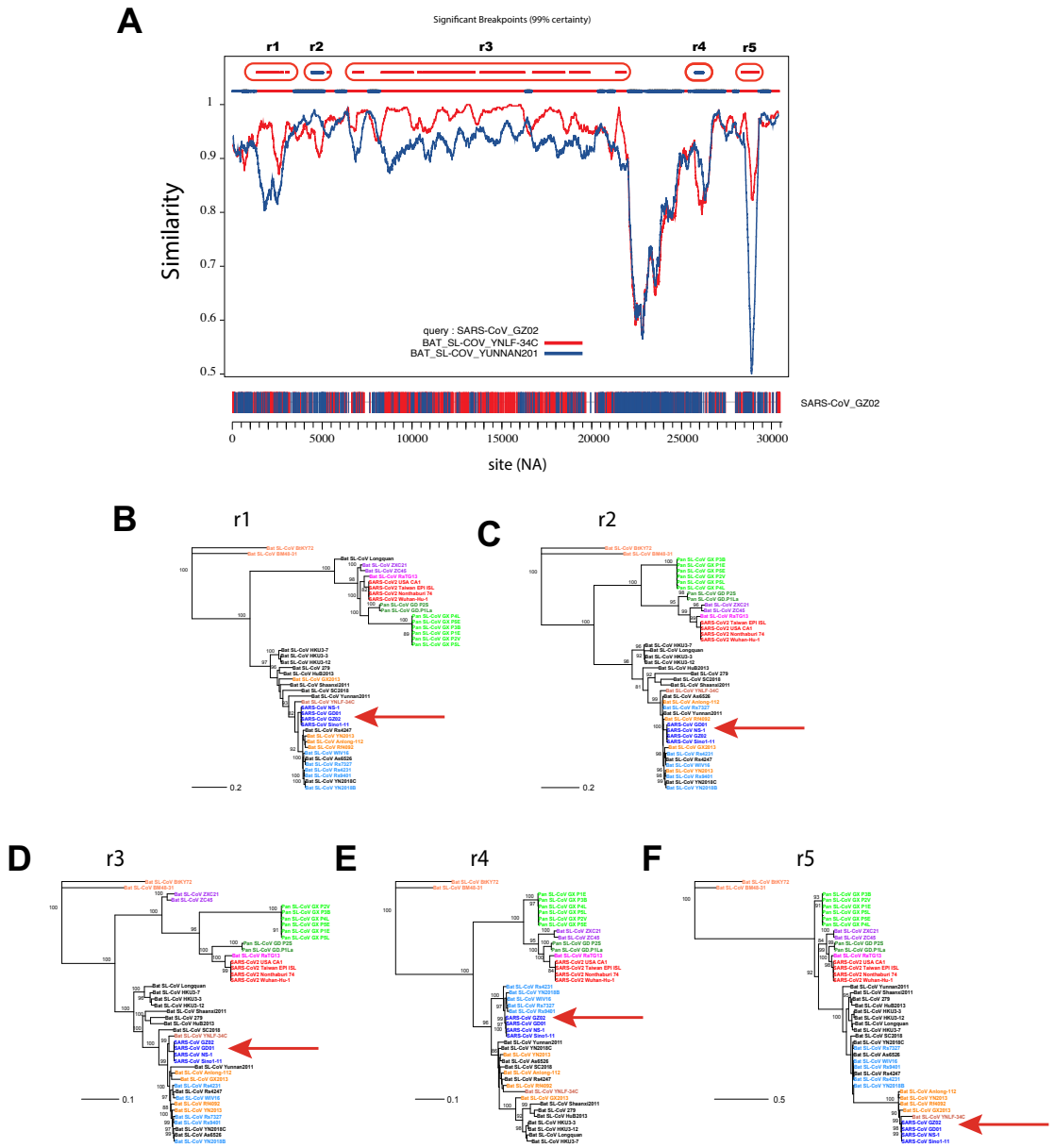

Fig. S9. Recombination analysis of SARS-CoV sequences. (A) Similarity plot comparing SARS sequence GZ02 to bat-CoV viruses YNLF-34C (red) and Yunnan2011 (blue). The plot was obtained using the recombination detection tool RIP with a window size of 400 base pairs. Top line shows break points at 99% confidence. Regions between significant break points are highlighted in the red ovals are marked r1-r5. At the bottom of the graph GZ02 is shown with nucleotide mutations colored in red if they are shared with sequence YNLF-34C, blue if they are shared with Yunnan2011. Nucleic acid unique to GZ02 are not shown. (B-F) Phylogenetic trees of the individual regions between break points, showing how the SARS sequences cluster more closely to either YNLF-34C or Yunnan2011 (red arrows). Regions between breakpoints were at the following base positions from the beginning of the genome: 1561-3303 (r1); 4621-5220 (r2); 5521-21360 (r3); 25201-25620 (r4); and 28201-29110 (r5).

Fig. S10. Nearly identical ORF8 sequences between SARS-CoV and four Bat_SL-CoV viruses. Nucleic acid sequence alignments of ORF8, partial ORF7 and the N gene compared to SARS-CoV GZ02 (top). Identical amino acids are shown as dashes and deletions as dots. Regions with nearly identical sequences among SARS-CoV and four bat_SL-CoV (GX2013, Anlong-112, Rf4092 and YN2013) viruses are highlighted in yellow. The ORF8 deletion in two highly divergent bat-SL-CoVs (BtKY72 and BM48-31) is highlighted in blue.

Table S1. Impact of amino acid substitutions in receptor binding motif

No. of mutation	Position in SARS-CoV2 RBM	AA in SARS-CoV2	AA in RaTG13	ΔΔG (kCal/Mol)	Effect for the RaTG13 mutations
1		Asn	Lys		No contact
2		Asn	His		No contact
3		Leu	Ile		No contact
4		Ser	Ala		No contact
5		Val	Glu		No contact
6	449	Tyr	Phe	1.62	Lost 1 h-bond
7		Ser	Ala		No contact
8		Thr	Lys		No contact
9		Val	Gln		No contact
10		Glu	Thr		No contact
11	486	Phe	Leu	1.47	Smaller/less hydrophobic
12		Phe	Tyr		No contact
13	493	Gln	Tyr	-0.02	more contact with residue H34 of ACE2
14		Ser	Arg		No contact
15	498	Gln	Tyr	1.78	too bulky
16	501	Asn	Asp	0.41	Buried a charge
17		His	Tyr		No contact

Table S2. Acknowledgement of sharing of SARS-CoV-2 genome sequences available at GISAID databases

Strain name	Accession ID	Virus name	Location	Collection date	Originating lab	Submitting lab	Authors
SARS-CoV2_Nonthaburi_74	EPI_ISL_403963	hCoV-19/Nonthaburi/74/2020	Asia / Thailand / Nonthaburi	2020-01-13	Bamrasnaradura Hospital	1. Department of Medical Sciences, Ministry of Public Health, Thailand 2. Thai Red Cross Emerging Infectious Diseases - Health Science Centre 3. Department of Disease Control, Ministry of Public Health, Thailand	Pilailuk,Okada; Siripaporn,Phuygun; Thanutsapa,Thanadachakul; Supaporn,Wacharapluesadee; Sittiporn,Parnmen; Warawan,Wongboot; Sunthareeya,Waicharoen; Rome,Buathong; Malinee,Chittaganpitch; Nanthawan,Mekha
SARS-CoV2_USA_CA1	EPI_ISL_406034	hCoV-19/USA/CA1/2020	North America / USA / California / Los Angeles	2020-01-23	California Department of Public Health	Pathogen Discovery, Respiratory Viruses Branch, Division of Viral Diseases, Centers for Dieases Control and Prevention	Anna Uehara, Krista Queen, Ying Tao, Yan Li, Clinton R. Paden, Jing Zhang, Xiaoyan Lu, Brian Lynch, Senthil Kumar K. Sakthivel, Brett L. Whitaker, Shifaq Kamili, Lijuan Wang, Janna' R. Murray, Susan I. Gerber, Stephen Lindstrom, Suxiang Tong
SARS-CoV2_Taiwan_EPI_ISL	EPI_ISL_406031	hCoV-19/Taiwan/2/2020	Asia / Taiwan / Kaohsiung	2020-01-23	Centers for Disease Control, R.O.C. (Taiwan)	Centers for Disease Control, R.O.C. (Taiwan)	Ji-Rong Yang, Yu-Chi Lin, Jung-Jung Mu, Ming-Tsan Liu, Shu-Ying Li
Pan_SL-CoV_GX/P1E	EPI_ISL_410539	hCoV-19/pangolin/Guangxi/P1E/2017	Asia / China / Guangxi	2017	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GX/P5E	EPI_ISL_410541	hCoV-19/pangolin/Guangxi/P5E/2017	Asia / China / Guangxi	2017	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GX/P5L	EPI_ISL_410540	hCoV-19/pangolin/Guangxi/P5L/2017	Asia / China / Guangxi	2017	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GX/P4L	EPI_ISL_410538	hCoV-19/pangolin/Guangxi/P4L/2017	Asia / China / Guangxi	2017	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GX/P3B	EPI_ISL_410543	hCoV-19/pangolin/Guangxi/P3B/2017	Asia / China / Guangxi	2017	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GX/P2V	EPI_ISL_410542	hCoV-19/pangolin/Guangxi/P2V/2017	Asia / China / Guangxi	2017	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GD/P2S	EPI_ISL_410544	hCoV-19/pangolin/Guangdong/P2S/2019	Asia / China / Guangdong	2019	Beijing Institute of Microbiology and Epidemiology	Beijing Institute of Microbiology and Epidemiology	Wu-Chun Cao; Tommy Tsan-Yuk Lam; Na Jia; Ya-Wei Zhang; Jia-Fu Jiang; Bao-Gui Jiang
Pan_SL-CoV_GD/P1L	EPI_ISL_412860	hCoV-19/pangolin/China/MP789/2019	Asia / China	2019-03-19	unknown	SCSFRI, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences (SCSFRI, CAFS)	Jiang,J.-Z., Liu,P. and Chen,J.-P.

Table S3. GenBank accession numbers of coronavirus sequences used in this study

Strain name	Accession ID	Host	Publication
SARS-CoV2_Wuhan-Hu-1	MN908947.3	Human	Wu *et al*. Nature 579 (7798), 265-269 (2020)
Bat_SL-CoV_RaTG13	MN996532.1	Rhinolophus affinis	Zhou *et al*. Nature 579 (7798), 270-273 (2020)
Bat_SL-CoV_ZC45	MG772933.1	Rhinolophus sinicus	Hu *et al*. Emerg Microbes Infect 7 (1), 154 (2018)
Bat_SL-CoV_ZXC21	MG772934.1	Rhinolophus sinicus	Hu *et al*. Emerg Microbes Infect 7 (1), 154 (2018)
SARS-CoV_Sino1-11	AY485277.1	Human	Zhang *et al*. Vaccine 23 (48-49), 5666-5669 (2005)
SARS-CoV_NS-1	AY508724.1	Human	*Journal information is not available in the GenBank record*
SARS-CoV_GD01	AY278489.2	Human	Wu *et al*. Genomics Proteomics Bioinformatics 1 (2), 131-144 (2003)
SARS-CoV_GZ02	AY390556.1	Human	*Journal information is not available in the GenBank record*
Bat_SL-CoV_YNLF-34C	KP886809.1	Rhinolophus Ferrumequinum	*Journal information is not available in the GenBank record*
Bat_SL-CoV_YN2018B	MK211376.1	Rhinolophus affinis	Han *et al*. Front Microbiol, 10: 1900 (2019)
Bat_SL-CoV_Rs9401	KY417152.1	Rhinolophus sinicus	PLoS Pathog. 13 (11), e1006698 (2017)
Bat_SL-CoV_Rs7327	KY417151.1	Rhinolophus sinicus	Hu *et al*. PLoS Pathog. 13 (11), e1006698 (2017)
Bat_SL-CoV_WIV16	KT444582.1	Rhinolophus sinicus	Yang *et al*. J. Virol. 90 (6), 3253–3256.(2016)
Bat_SL-CoV_Rs4231	KY417146.1	Rhinolophus sinicus	Hu *et al*. PLoS Pathog. 13 (11), e1006698 (2017)
Bat_SL-CoV_GX2013	KJ473815.1	Rhinolophus sinicus	Wu *et al*. J. Infect. Dis. 213 (4), 579-583 (2016) Wu *et al*. ISME J 10 (3), 609-620 (2016)
Bat_SL-CoV_Shaanxi2011	JX993987.1	Rhinolophus pusillus	Yang *et al*. Emerging Infect. Dis. 19 (6) (2013) Wu *et al*. J. Infect. Dis. 213 (4), 579-583 (2016) Wu *et al*. ISME J 10 (3), 609-620 (2016)
Bat_SL-CoV_Yunnan2011	JX993988.1	Chaerephon plicata	Yang *et al*. Emerging Infect. Dis. 19 (6) (2013) Wu *et al*. J. Infect. Dis. 213 (4), 579-583 (2016) Wu *et al*. ISME J 10 (3), 609-620 (2016)
Bat_SL-CoV_Rs4247	KY417148.1	Rhinolophus sinicus	Hu *et al*. PLoS Pathog. 13 (11), e1006698 (2017)
Bat_SL-CoV_As6526	KY417142.1	Aselliscus stoliczkanus	Hu *et al*. PLoS Pathog. 13 (11), e1006698 (2017)
Bat_SL-CoV_YN2018C	MK211377.1	Rhinolophus affinis	Han *et al*. Front Microbiol, 10: 1900 (2019)
Bat_SL-CoV_Rf4092	KY417145.1	Rhinilophus ferrumequinum	Hu *et al*. PLoS Pathog. 13 (11), e1006698 (2017)
Bat_SL-CoV_YN2013	KJ473816.1	Rhinolophus sinicus	Wu *et al*. J. Infect. Dis. 213 (4), 579-583 (2016) Wu *et al*. ISME J 10 (3), 609-620 (2016)
Bat_SL-CoV_Anlong-112	KY770859.1	Rhinolophus sinicus	Lin *et al*. Virology 507, 1-10 (2017)
Bat_SL-CoV_HKU3-7	GQ153542.1	Rhinolophus sinicus	Lau *et al*. J. Virol. 84 (6), 2808-2819 (2010)
Bat_SL-CoV_HKU3-12	GQ153547.1	Rhinolophus sinicus	Lau *et al*. J. Virol. 84 (6), 2808-2819 (2010)
Bat_SL-CoV_HKU3-3	DQ084200.1	Rhinolophus sinicus	Lau *et al*. Proc. Natl. Acad. Sci. U.S.A. 102 (39), 14040-14045 (2005)
Bat_SL-CoV_HuB2013	KJ473814.1	Rhinolophus sinicus	Wu *et al*. J. Infect. Dis. 213 (4), 579-583 (2016) Wu *et al*. ISME J 10 (3), 609-620 (2016)
Bat_SL-CoV_Longquan	KF294457.1	Rhinolophus monoceros	Lin *et al*. Virology 507, 1-10 (2017)
Bat_SL-CoV_279	DQ648857.1	Rhinolophus Macrotis	Tang *et al*. J. Virol. 80 (15), 7481-7490 (2006)
Bat_SL-CoV_SC2018	MK211374.1	Rhinolophus sp.	Han *et al*. Front Microbiol, 10: 1900 (2019)
Bat_SL-CoV_BM48-31	GU190215.1	Rhinolophus blasii	Drexler *et al*. J. Virol. 84 (21), 11336-11349 (2010)
Bat_SL-CoV_BtKY72	KY352407.1	Rhinolophus sp.	Tao *et al*. Microbiol Resour Announc 8 (28), e00548-19 (2019)