Materials and Methods

Sequences analysis

All 43 CoV complete genome sequences were obtained from GenBank and GISAID (Global Initiative on Sharing All Influenza Data) (*35, 36*), and were selected to be representative of the diversity. Pan_SL-CoV_GD/P1La sequence was generated by combining Pan_SL-CoV_GD/P1L (*10*) with some additional sequences from the NCBI BioProject database PRJNA5732983 (*11, 37*) to have a maximal coverage of the complete genome sequence for analysis. A new CoV sequence from pangolin (EPI_ISL_410721) (*38*)was not inclued because that it became available after we had already completed the analyses in this study, and it was not as close to SARS-CoV-2 sequences and did not change the interpretation of our results. The whole genome sequences were first aligned using Clustal X2 (*39*). The alignments for all coding regions were manually optimized based on the amino acid sequence alignment using SeaView 5.0.1.

Recombination Analyses

SimPlot 3.5.15 (*14*) was used determine the percent identity of the query sequence to reference sequences. Potential recombinant regions among analyzed sequences were identified by sliding a 400bp-window at a 50bp-step across the alignment using the Kimura 2-parameter model. Phylogenetic trees were constructed by the maximum likelihood method using the GTR model (*40*), and their reliability was estimated from 1,000 bootstrap replicates. The positions of analyzed sequence regions were based on those in the reference SARS-CoV-2 Wuhan-Hu-1 (MN908947). Recombination regions and breakpoints were also analyzed using the LANL database (*41*) tool RIP (*16*).

Selection Analyses

Cumulative plots of the average behavior of each codon for all pairwise comparisons in the input data, for insertions and deletions (indels), synonymous (syn), and nonsynonymous (nonsyn) mutations and values of the ratios of the rate of synonymous nucleotide substitutions per synonymous site and nonsynonymous substitutions per nonsynonymous site (dN/dS, or ω) were obtained using the LANL database tool SNAP (*42*). In order to avoid counting instances where synonymous mutations were saturated, averages of all pairwise dN/dS ratios were calculated excluding pairs that yielded dS values greater than 1. Sequences were analyzed for episodic selection pressure using the mixed effects model of evolution (MEME) (*43*) from the datamonkey server (www.datamonkey.org).

Structure modeling of receptor binding

To investigate the single mutation Q498H in RBM between SARS-CoV-2 and Pan_SL-CoV_GD, Q498 in the crystal structure of S/ACE2 complex was mutated to H498 using Chimera (*44*). Local energy minimization (only H498 was allowed to move) was computed using Chimera's built-in functions. To investigate the impact of the deletion between residue 473 to 486 to the binding interface between SARS-CoV-2 and human ACE2, a homology model with the deletion was generated using I-TASSER (*45*). The top five best models provided by the server have Confidence Score (C-score) of 0.86, -2.33, -4.01, -4.17, and -4.49. The C-score was used to estimate the quality of the models, which should be between -5.0 to 2; the higher the value, the higher the confidence in the model (*45*). Based on the C-score, model 1 was used in Figure 2F. The interaction of the RBD of RaTG13 and ACE2 was modeled on PDB 6VW1, a

hybrid structure of human SARS-CoV2 (*46*) using ICM software package (*35*), and the mutational differences of the Gibbs free energy (Table S1) were calculated with the built-in algorithm.

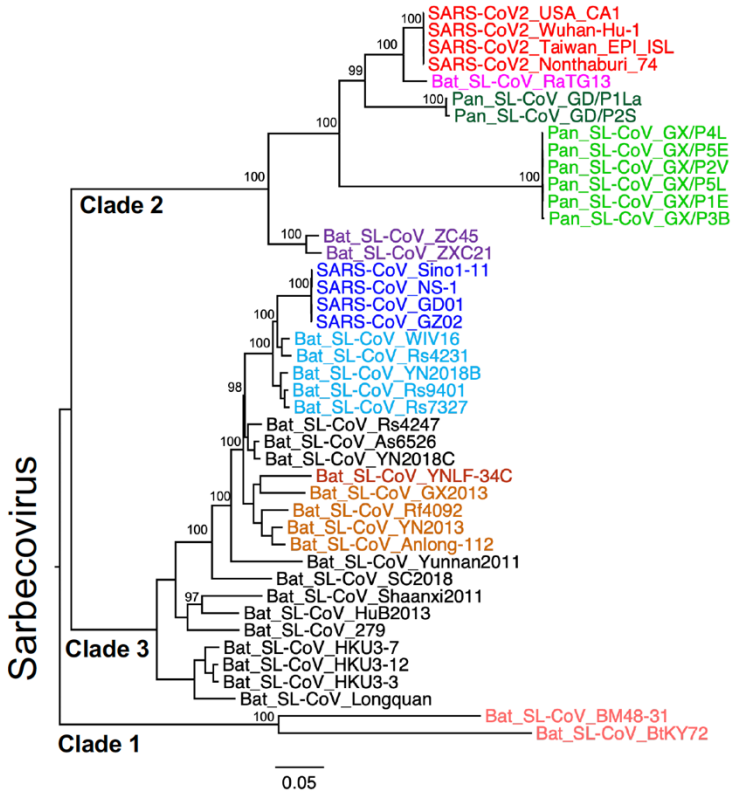

Fig. S1. Phylogenetic tree of the complete CoV genome sequences. All 43 sequences used in this study includes: 4 SARS-CoV-2 sequences (red), Bat_SL-CoV sequence RaTG13 (magenta), 2 pangolin CoV from Guangdong (Pan_SL-CoV_GD, dark green), 6 pangolin CoV from Guangxi (Pan_SL-CoV_GX , light green), and 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 (*8*), and their reliability was estimated from 1,000 bootstrap replicates.

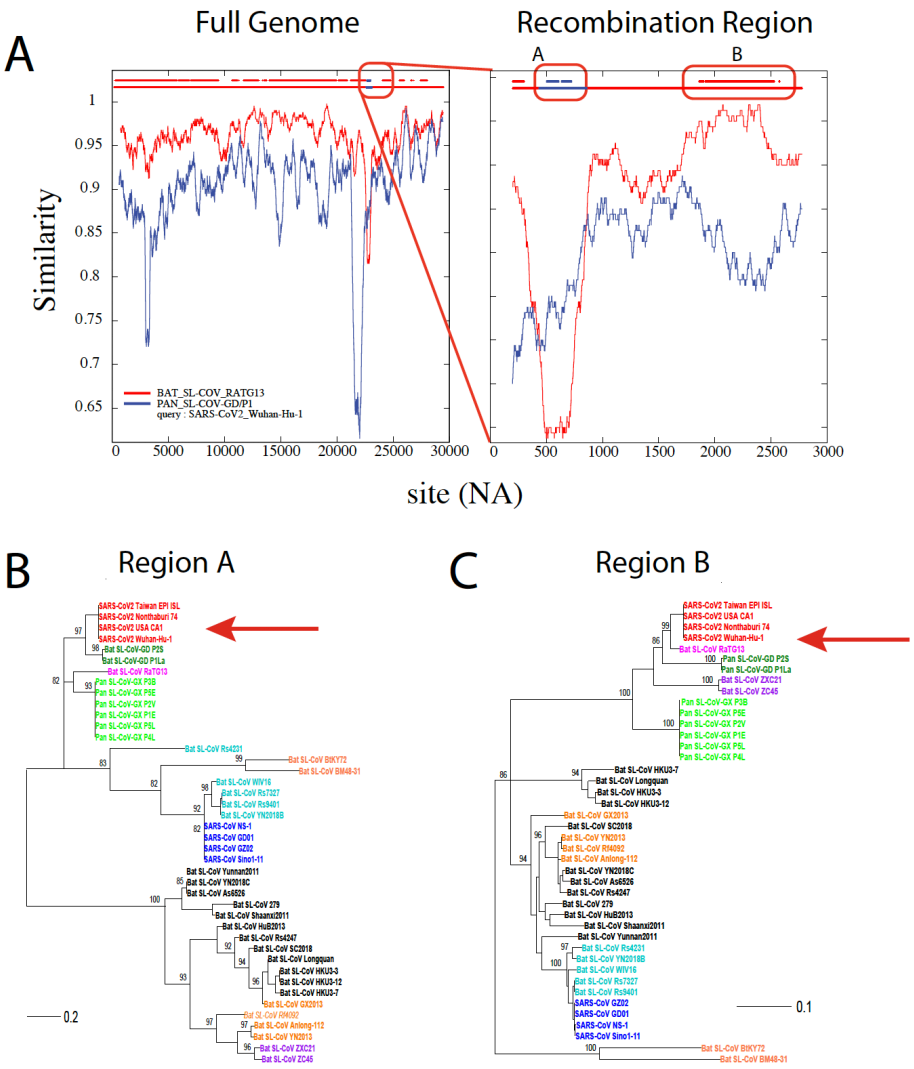

Fig. S2. Recombination analysis of the CoV-SARS-2 Wuhan-Hu-1 sequence. (A) Similarity plots comparing the Wuhan-Hu-1 sequence to the bat-Cov RaTG13 sequence (red) and the Pan_SL-CoV_GD/1PL sequence (blue). Plots were obtained using the LANL tool RIP using a window size of 400 bp. The full genome comparison 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/1PL (highlighted by the red arrows).

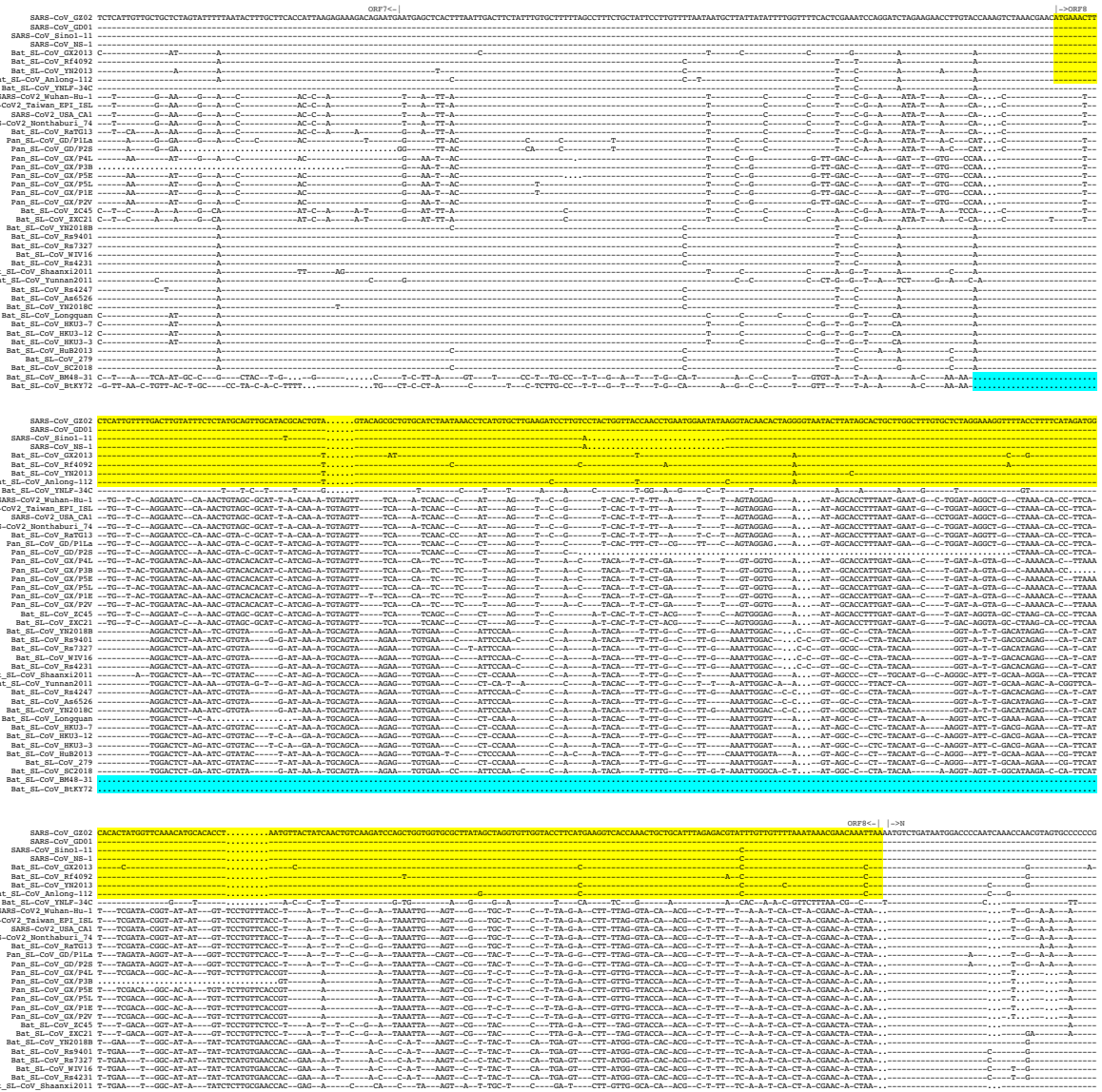

Fig. S3. Highly conserved sequences around the receptor binding motif and furin cleavage sites among SARS-CoV-2, RaTG13 and Pan_SL-CoV viruses. Alignment of amino acid sequences around receptor binding motif (RBM) and furin cleavage sites in the spike glycoprotein compared to Wuhan-Hu-1 (top sequence, 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.

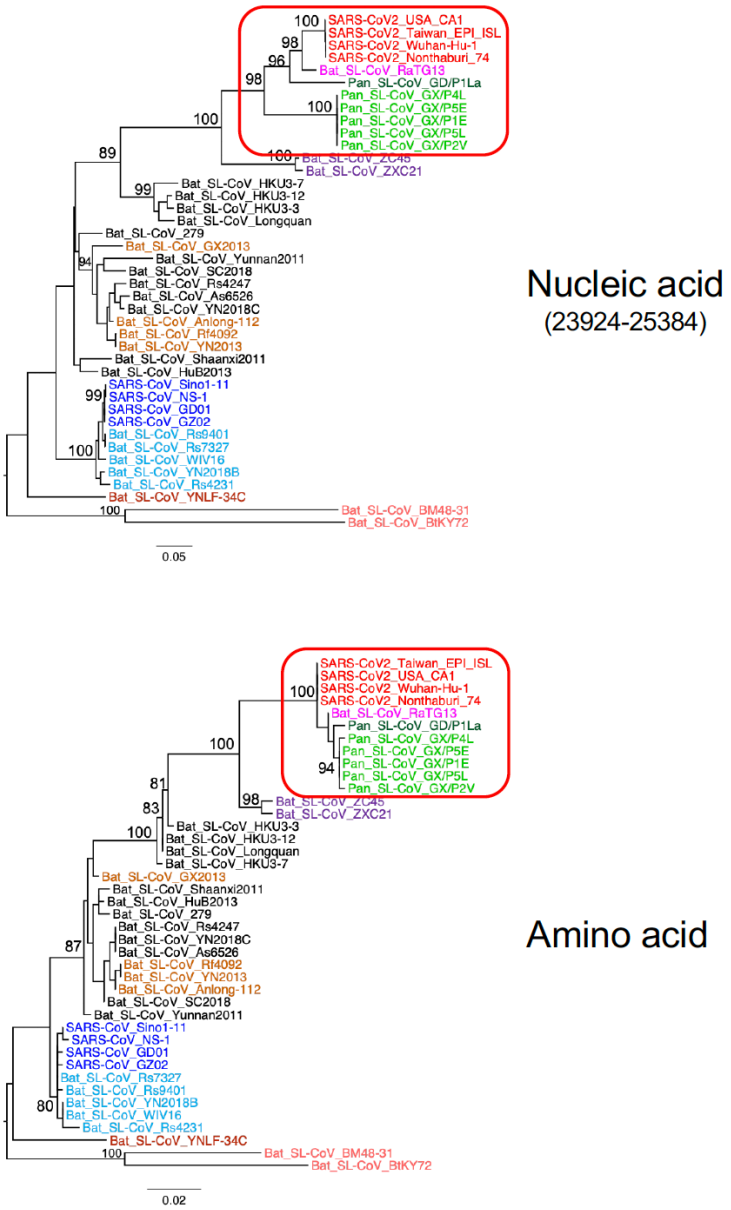

Fig. S4. Purifying selection in the 3' end region of the S gene. Purifying selection pressure on the 3' end region (na 23924-25384) of the S gene region among SARS-CoV-2, RaTG13 and Pan_SL-CoV viruses (within red boxes in phylogenetic trees) are shown by much shorter branches with amino acid sequences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

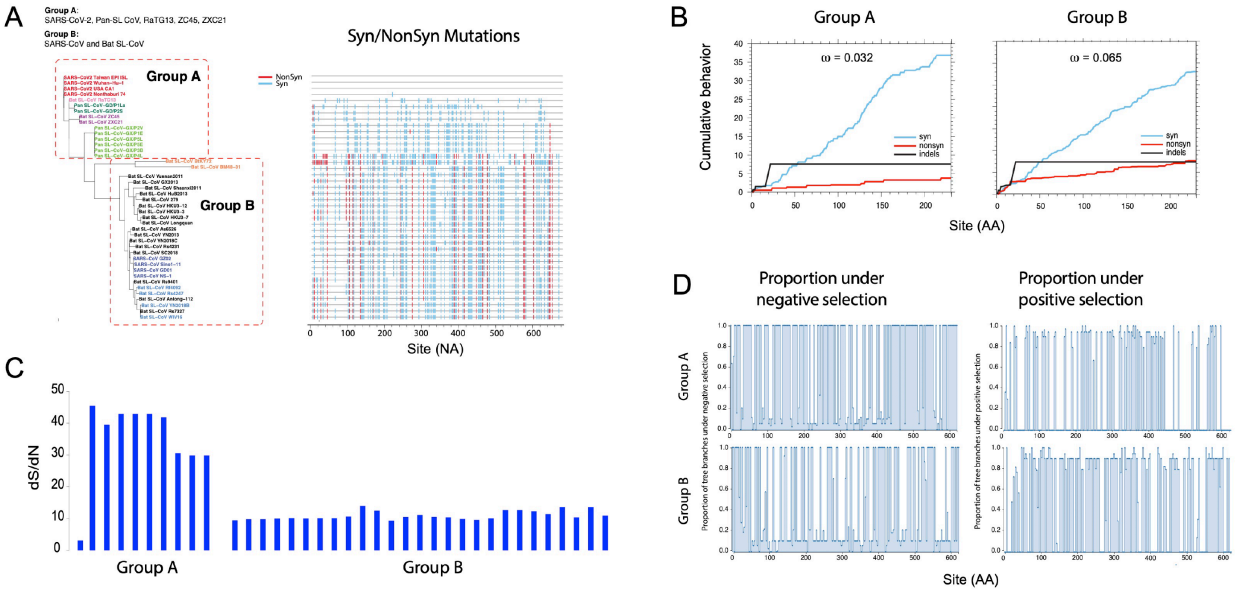

Fig. S5. Purifying selection pressure on the M gene. (A) Phylogenetic tree (left) and Highlighter plot (right) of all sequences compared to SARS-CoV-2 in the M gene. SARS-CoV-2, RaTG13, all Pan_SL-CoV and the two bat CoV (ZXC21 and ZC45) sequences are in Group A, and all other sequences in Group B, to 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. (D) Proportion of tree branches under positive and negative selection (right and left respectively) for the two sequence groups as calculated using the mixed effects model of evolution (MEME) from the datamonkey (www.datamonkey.org) server.

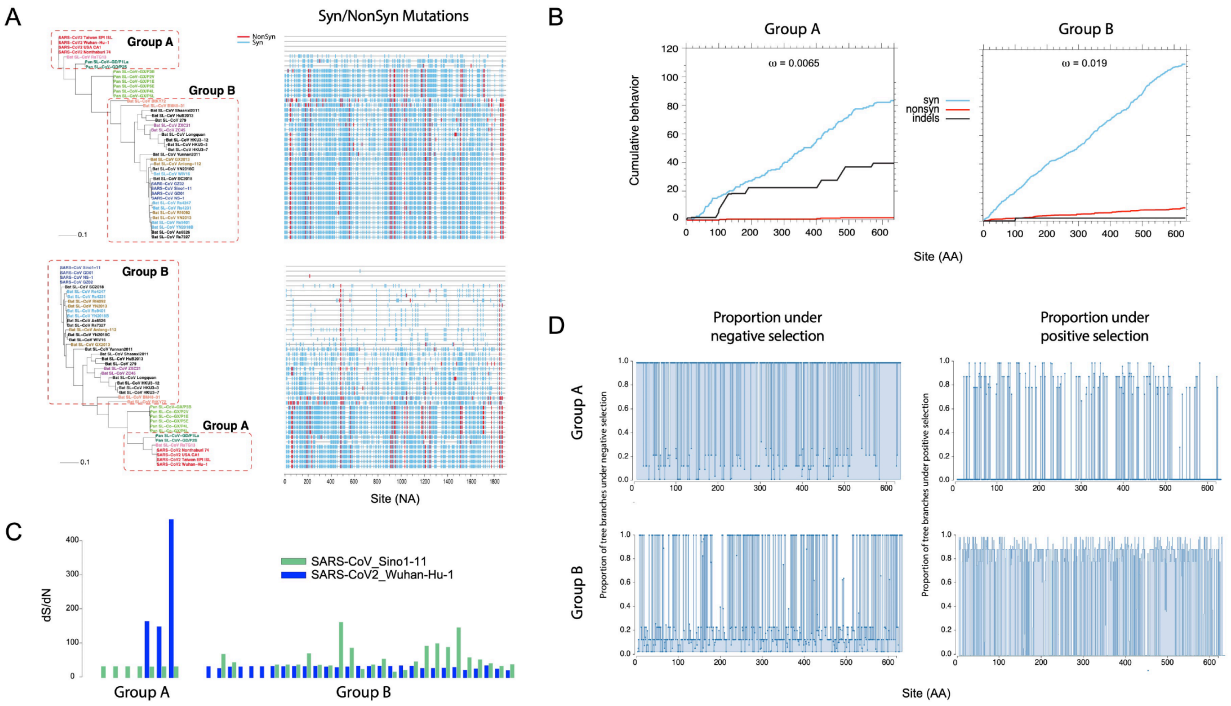

Fig. S6. Purifying selection pressure on the partial region of ORF1a. (A) Phylogenetic trees (left) and Highlighter plots (right) of sequences compared to SARS-CoV-2 (top) and to SARS-CoV (bottom) in the partial region of ORF1a. SARS-CoV-2, RaTG13 and Pan_SL-CoV from Guangdon are in Group A, and all 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. (D) Proportion of tree branches under positive and negative selection (right and left respectively) for the two groups as calculated using the mixed effects model of evolution (MEME) from the datamonkey server (www.datamonkey.org).

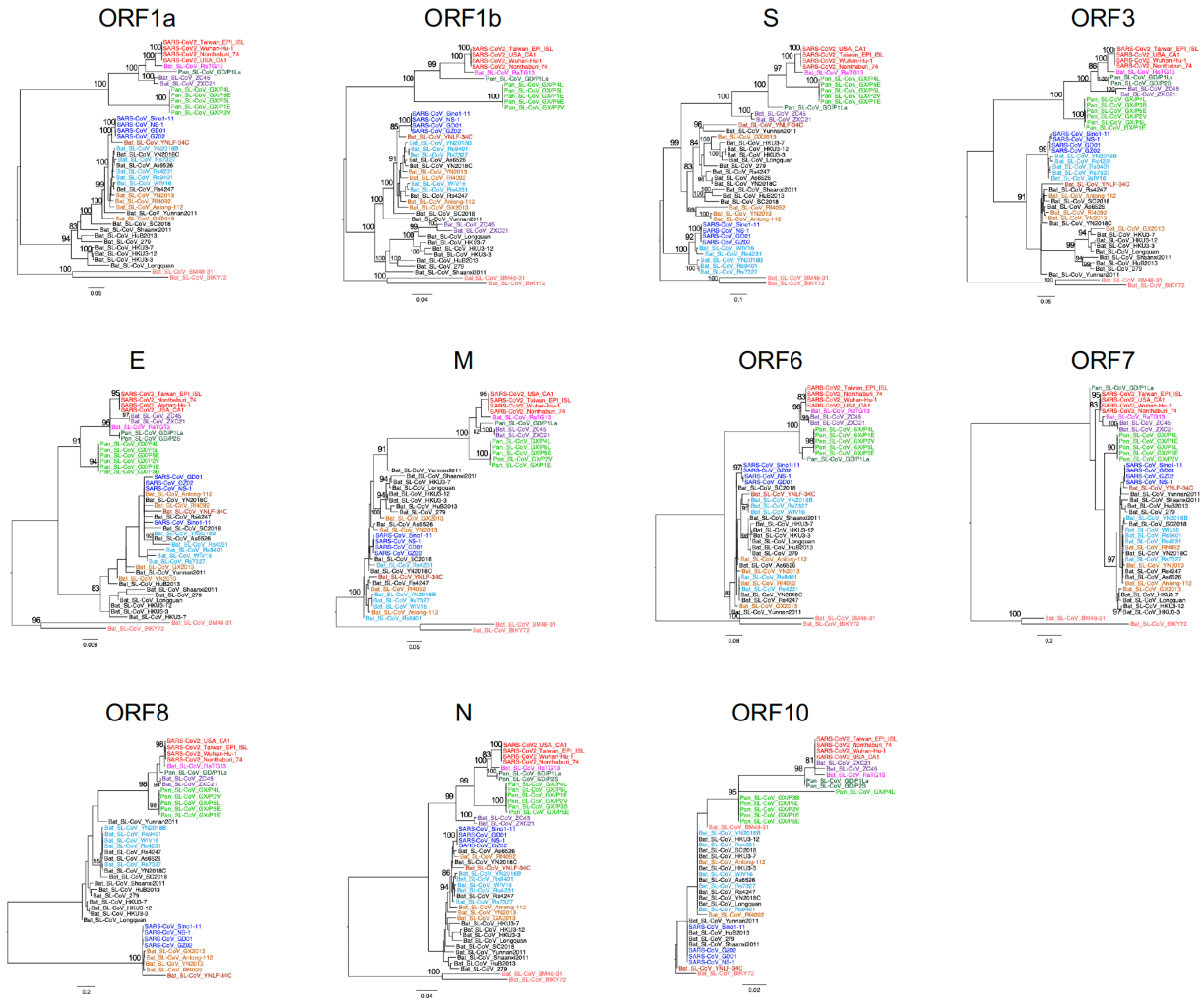

Fig. S7. Phylogenetic tree analysis of SARS-CoV-2 genes together with other CoVs. Phylogenetic trees were constructed for each coding region in the CoV genome. Sequences are colored differently based on their hosts and phylogenetic cluster: 4 SARS-CoV-2 sequences (red), Bat_SL-CoV sequence RaTG13 (magenta), 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.

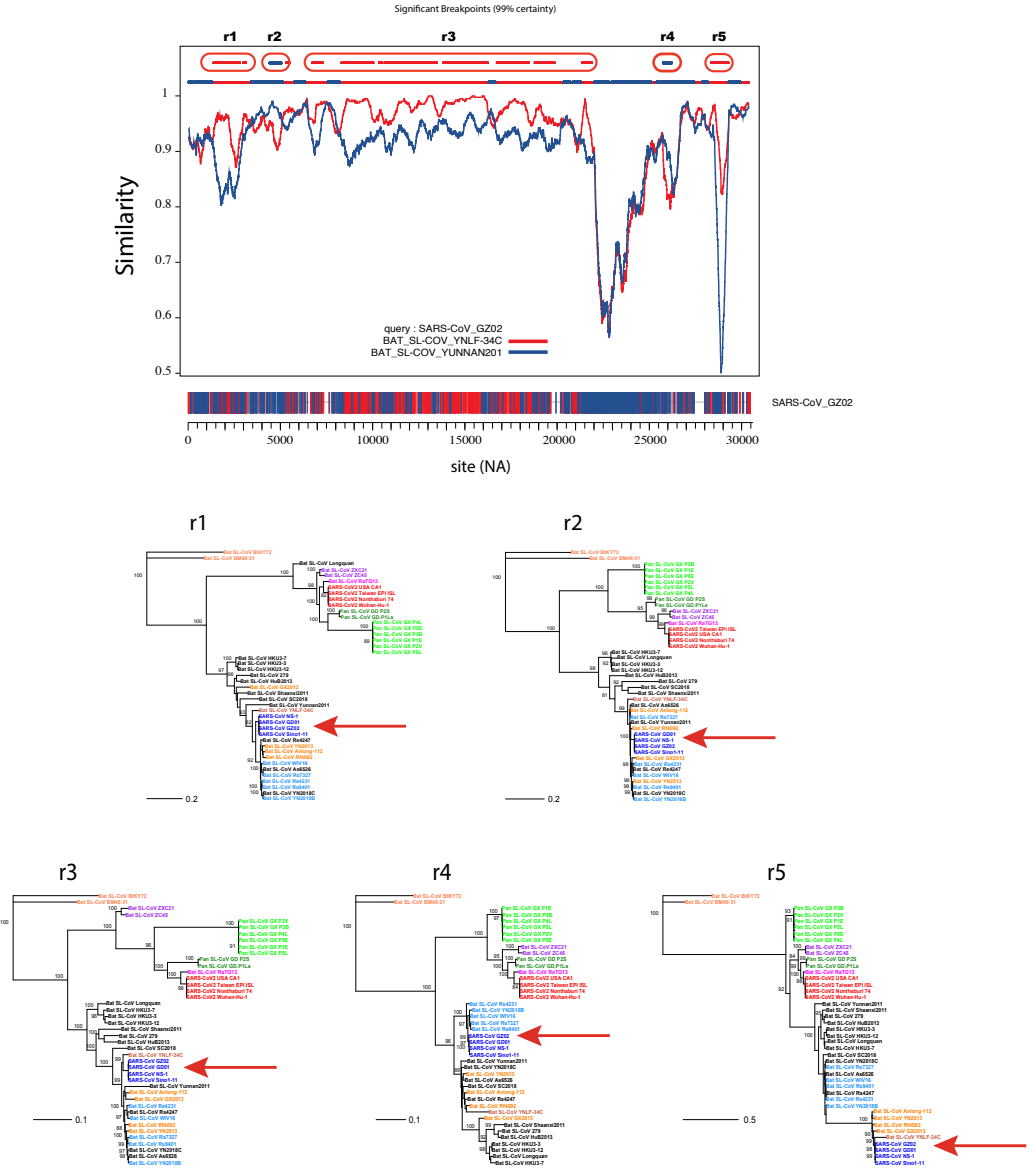

Fig. S8. Recombination analysis of SARS-CoV sequences. (A) Similarity plot comparing SARS sequence GZ02 to bat-CoV viruses YNLF-34C (red) and Yunnan2011 (blue). The plot was obtained using the recombination detection tool RIP with a window size of 400 base pairs (*10*). Top line shows break points at 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. S9. 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 SARS2 RBM	AA in SARS2	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	0.71	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	1.63	Small/less hydrophobic
12		Phe	Tyr		No contact
13	493	Gln	Tyr	3.44	Gain a h-bond/too bulky
14		Ser	Arg		No contact
15	498	Gln	Tyr	1.26	too bulky
16	501	Asn	Asp	0.31	Buried a charge
17		His	Tyr		No contact

References:

35. R. Abagyan, M. Totrov, D. Kuznetsov, ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of computational chemistry* **15**, 488-506 (1994).
36. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, (2017).
37. P. Liu *et al.*, Are pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV) ? *bioRxiv*, 2020.2002.2018.954628 (2020).
38. K. Xiao *et al.*, Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv*, 2020.2002.2017.951335 (2020).
39. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
40. S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* **52**, 696-704 (2003).
41. B. Foley *et al.*, *HIV sequence compendium 2018*. (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 18-25673, Los Alamos, New Mexico, 2018).
42. B. B. T. Korber, in *Computational Analysis of HIV Molecular Sequences,* A. G. Rodrigo, G. H. Learn, Eds. (Kluwer Academic Publishers, Dordrecht, Netherlands, 2000), chap. 4, pp. 55-72.
43. B. Murrell *et al.*, Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* **8**, e1002764-e1002764 (2012).
44. E. F. Pettersen *et al.*, UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* **25**, 1605-1612 (2004).
45. J. Yang, Y. Zhang, Protein Structure and Function Prediction Using I-TASSER. *Current protocols in bioinformatics* **52**, 5.8.1-5.8.15 (2015).
46. J. Shang *et al.*, Structural basis for receptor recognition by the novel coronavirus from Wuhan. *DOI:10.21203/rs.2.24749/v1*, (2020).