Appendices

S1 Supplementary Digital Material

The supplementary digital material is available at Zenodo (https://zenodo.org), DOI: 10.5281/zenodo.3740770 (https://doi.org/10.5281/zenodo.3740770). In all files containing molecular information from GISAID's EpiCoV database, we masked GISAID's data, viz. each nucleotide was replaced by missing data ("?" or "N"), in compliance with that database's policies.

S1a Selected terminals

The accession numbers of the 2,006 terminals (76 COVID-2019 from GISAID and 1,930 sequences from NCBI's RefSeq and GenBank databases) used in this study are listed in "**terminals.csv**." Sequence metadata (with different tabs that contain notes on terminal names and host information) is in "**metadata.xlsx**."

All sequences here are unique, and no sequence is a substring of another complete genome on the database. Also, selected sequences are longer than 26 Kbp and have less than 0.1% of character states that are different from A, C, T, or G (e.g., missing data and gaps). Finally, we were able to predict the partitions ORF1ab, M, S, and N for all sequences herein.

S1b Data matrix

The final DNA matrix in "**matrix.ss**" comprises 38,274 characters divided into four partitions, representing the genes ORF1ab (translated by ribosomal frameshifting), S (spike glycoprotein trimer), M (membrane protein), and N (nucleoprotein).

The same matrix is also available in NEXUS format ("**matrix.nex**"), and the partitions and selected models are descriped in the NEXUS file "**partitions.nex**."

S1c Tree search

The template for the script used to perform different tree search replicates on TNT is named "**treeSearch.RUN**." This script was executed ten times, changing the replicate number accordingly. A total of 100 rounds of tree fusing were executed using all trees found this way (see "**fuse.RUN**"). Consensus trees were produced with "**consensus.RUN**." Trees with branch lengths were produced with "**branchLength.RUN**." Bootstrap calculations were performed with "**bootstrap.RUN**." The calculation of Goodman-Bremer support values was based on the macro "**Bremer.RUN**".

S1d Recombination analyses

The parameters used for whole-genome alignment and recombination detection among the complete genomes of the SARS-CoV-2 reference sequence (RefSeq accession number NC_045512.2), a bat-hosted COV RaTG13 (GISAID accession number EPI_ISL_41402131), a representative of the Pan_SL-CoV_GD clade (GISAID accession number EPI_ISL_410721), and two other bat-hosted SARS-like viruses (GenBank accession numbers MG772933.1 and MG772934.1), as well as the main results, are provided in a single PDF file ("**recombination.pdf**").

S1e Graphical abstract

The graphical abstract below summarizes our main results. See full image in file "**graphicalAbstract.pdf**" (Figure S1.1).

S1f Phylogenetic trees from parsimony analyses

The NEXUS file "**parsimony.nex**" contains the best heuristic results from the parsimony analyses (six trees), the tree with branch lengths, the tree with bootstrap values, the tree with Goodman-Bremer support values, the tree with REP values, and the strict consensus tree. The file also contains a tree with merged data (e.g., node numbers, clade frequencies, branch lengths).

S1g Bootstrap values and clade sizes from parsimony analysis

Boostrap values among all nodes varied from 0 to 100% (mean = 65.74%, median = 80%, and mode = 100%). Boostrap values on the consensus tree varied from 1 to 100% (mean = 75.17%, median = 90%, and mode = 100%). For scatter plots and histograms showing the variation of boostrap values in relation to clade size, see file "**bootstrap.png**" (Figure S1.2).

S1h Complete consensus tree from parsimony analyses

A high-resolution version of the consensus tree from the best six heuristic results from tree search performed under the parsimony criterion is in file "**parsimony.pdf**" Branch lengths are proportional to the number of transformations and branch colors correspond to bootstrap values (see legend in the figure).

S1i Host shifts

The spreadsheet in "**hosts.csv**" contains the minimum and the maximum number of each type of host transformations. The complete consensus tree with the YBYRÁ's categorization of host transformations is available in "**hosts.pdf**."

S1j TreeTime analyses

Analyses with TreeTime v0.7.5 (available at `https://github.com/neherlab/treetime`) following instructions from its documentation (revision f1c83c30, available at `https://treetime.readthedocs.io`). We included the results of the following analyses:

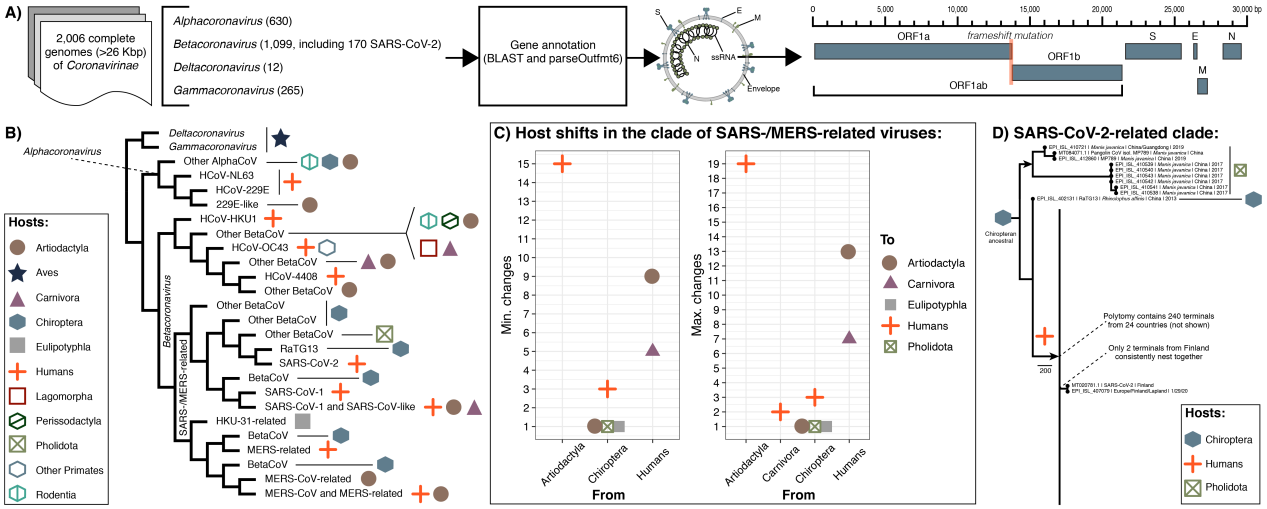

Fig. S1.1. (A) We analyzed a total of 2,006 complete genomes (¿26 kbp) of *Orthocoronavirinae* viruses (630 *Alphacoronavirus*, 1,099 *Betacoronavirus*, 12 *Deltacoronavirus*, and 265 *Gammacoronavirus*). We annotated four partitions (polyprotein ORF1ab, spike glycoprotein S, E protein, membrane protein M, and nucleoprotein N) by filtering and expanding BLAST results with a homemade Python package (parseOutfmt6). (B) The strict consensus of the best heuristic results from parsimony analysis was mostly congruent with the maximum likelihood tree. We recovered the monophyly of all genera. We optimized host transformations (ten orders; we separated humans from other Primates) and found evidence supporting the hypotheses that viruses from bats were responsible for the human infections related to SARS-CoV, SARS-CoV-2, and MERS-CoV. (C) The data plots show the minimum and the maximum number of host shifts on the clade of SARS- and MERS-related coronaviruses, including SARS-CoV-2. The number of shifts from Artiodactyla to humans is mostly inflated in the MERS-related clade after an ancestral virus from Chiroptera infected humans and Artiodactyla. Host shifts from Chiroptera to humans occurred three times on this clade, with humans infecting Carnivora hosts at least five times and humans infecting camels at least nine times. Shifts from Chiroptera to Eulipotyphla or from Chiroptera to Pholidota occurred only once in events that are not associated with human infections. (D) Independent analysis of SARS-CoV-2-related clade. The figure shows that the infection of Pholidota from a Chiroptera ancestral is independent of the human infection by SARS-CoV-2. The 240 unique sequences from SARS-CoV-2 (out of 341 SARS-CoV-2 samples), although not identical to each other, do not contain phylogenetically informative SNPs. The lack of informative sites results in a large polytomy in which only two sequences from patients from Finland appear to be consistently nested together.

• The spreadsheet in "**treetime.csv**" contains the main results from TreeTime analysis, including estimated mutation rates and the minimum and maximum estimated dates for the selected virus clades. It also gives each of the virus' earliest publications and their respective DOIs. Finally, this spreadsheet has the details about the earliest genetic sequences submitted to NCBI's databases for each of the virus it lists.

• Host shift calculation using the "mugration" model: the compressed folder "**mugration.zip**" contains the GTR model calculations ("**GTR.txt**"), confidence values per node and state ("**confidence.csv**"), and the annotated tree data showing all host shifts ("**annotated_tree.nex**" and "**annotated_tree.pdf**").

• Mutation rates: the compressed folder "**mutation_rates.zip**" contains details about selected clades, including branch lengths ("**clade_data.csv**"). It also contains host and collection dates for terminals ("**terminal_data.csv**") and root-to-tip regression analyses ("**root-to-tip-regressions.csv**" and "**root-to-tip-regressions.pdf**")

S1k Recombination detection analysis

The spreadsheet in "**summaryFromRdp5_505terminals.xlsx**" contains the results of the recombination detection analysis of a 505 terminals dataset. The results in there were used to test the sensitivity of phylogenetic analysis to the removal of putative recombinant sequences.

S1l Maximum likelihood trees

The maximum likelihood tree (log-likelihood: -2,240,329.5917) is available in "**likelihood.nex**." Node labels show the support values formatted as SH-aLRT support and bootstrap values. The branch lengths are proportional to the average number of nucleotide substitutions per nucleotide site.

Unconstrained maximum likelihood trees for each partition are in "**ml_gene_trees.nex**."

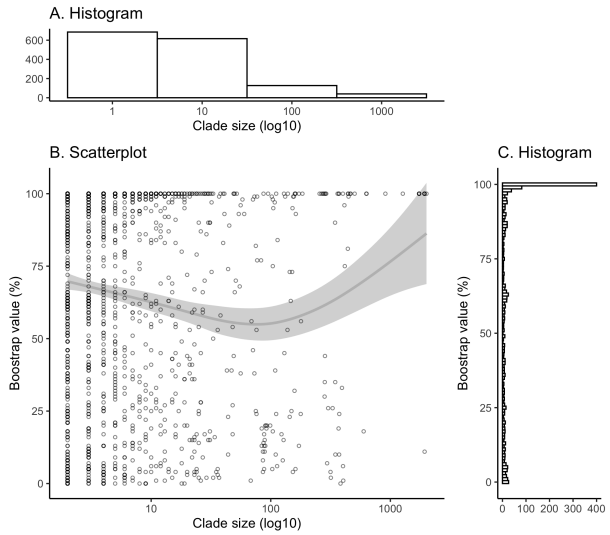

Fig. S1.2. This figure is available in file "bootstrap.png." (A) Histogram of clade sizes (i.e, number of terminals per clade), on log_{10} scale. (B) Scatterplot showing the correlation between clade size (on log_{10} scale) and clade frequencies calculated using the bootstrap metric (percentage). The shaded grey line represents the smoothed conditional means calculated in R v3.6.3 with "geom_smooth()" (from ggplot2 v3.3.0) using "method = gam" and formula y ~ s(x, bs = "cs"). (C) Histogram of bootstrap values.

S1m Subsets for sensitivity analysis

The matrices, partition schemes, best heuristic solutions, and strict consensus trees from the datasets of 505 and 315 terminals used to test the sensitivity to putative recombinant genomes are in the NEXUS files "**dataset505terminals.nex**" and "**dataset315terminals.nex**", respectively.

S1n Phylogenetic analyses of the SARS-CoV-2 related clade

The NEXUS file "**sarscov2.nex**" contains the alignment matrix and partition scheme used in the independent phylogenetic analyses of the SARS-CoV-2 clade.

The best heuristic solutions (8,900 steps each) and strict consensus tree from parsimony analyses are available in "**sarscov2_parsimony.nex**."

The maximum likelihood tree (likelihood score equal to -67,779.744) is in "**sarscov2_ml.nex**." Node labels show the support values formatted as SH-aLRT support and bootstrap values. The branch lengths are proportional to the average number of nucleotide substitutions per nucleotide site.

S1o Alignment comparisons in the SARS-CoV-2-related clade

The file "**sarscov2_aligns.xlsx**" contain summary stats of the alignment comparisons between the SARS-CoV-2 reference sequence (NCBI's RefSeq accession number NC_045512.2) and related viruses found in humans, bats, and pangolin hosts.

S1p Alignment comparisons of the repeat binding motif of the spike glycoprotein

The file "**rbm.xlsx**" contains details on the comparisons between the receptor-binding motif (RBM) of the spike glycoprotein of SARS-CoV-2 (NCBI's RefSeq accession number NC_045512) and other viruses infecting humans, bats, and pangolins in the SARS-CoV-2-related clade. This Excel spreadsheet has two tabs summarizing data from the amino acid and nucleotide alignments, respectively.

S2 Supplementary Acknowledgement Table

The complete GISAID acknowledgement table is provided in file "**acknowledgement.xlsx**" (Zenodo, DOI: 10.5281/zenodo.3740770).

S3 Glossary

We compose a glossary, provided in file "**glossary.pdf**" (Zenodo, DOI: 10.5281/zenodo.3740770), with selected terms and concepts that are in our manuscript or that are crucial to understanding the references we cited.