On Classification and Taxonomy of Coronaviruses (*Riboviria*, *Nidovirales*, *Coronaviridae*) with Special Focus on Severe Acute Respiratory Syndrome-Related Coronavirus 2 (SARS-CoV-2)

Evgeny Mavrodiev (evgeny@ufl.edu)
University of Florida https://orcid.org/0000-0002-9670-7662
Melinda L Tursky
St Vincent Clinical School, Faculty of Medicine, UNSW, Sydney, 210, Australia
Nicholas E Mavrodiev
Cornerstone Academy, Gainesville, FL, USA
Malte C Ebach
Paleontology, Geobiology and Earth Archives Research Center (PANGEA) of School of Biological Sciences, UNSW, Australia
David M Williams
London Museum of Natural History, Department of Life Sciences
Laura Schroder
University of Alaska Anchorage

Research Article

Keywords: pathogenic, cladistic analysis, infrageneric taxa, Coronaviridae

Posted Date: June 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-282371/v1

Abstract

Coronaviruses are highly pathogenic and therefore important human and veterinary pathogens viruses worldwide. This study is the first that produced phylogenies based on all 39 species of *Coronaviridae* recognized by the Coronaviridae Study Group and International Committee on Taxonomy of Viruses and has, uniquely, used several methods of cladistic analysis in combination with the Maximum Likelihood method. Resultant trees were utilized to test for monophyly of all available non-monotypic genera and infrageneric taxa of Coronaviridae. Monophyly was confirmed, thereby validating they are representative of a nature hierarchy. This study therefore presents the first natural hierarchical classification of *Coronaviridae* and the most accurate taxonomic representation of *Coronaviridae*'s relationships to date. The authors additionally seek at add to the current discussion regarding the nomenclature of viruses, demonstrating and supporting a "one-step" solution to incorporate the principles of binary nomenclature into *Coronaviridae*, which will aid future recognition of numerous virus species, particularly in currently monotypic subgenus *Sarbecovirus*. Commenting on the nature of SARS-CoV-2, the authors emphasize that no member of the *Sarbecovirus* clade is an ancestor of this virus, and humans are the only *natural* known host.

Introduction

Coronaviridae is a virus family of the order *Nidovirales* (realm *Riboviria*) [1, 2]. According to the current summaries of International Committee on Taxonomy of Viruses (ICTV) [1–7] this family is divided into two subfamilies - *Letovirinae* and *Orthocoronavirinae*. These two subfamilies circumscribe five genera: monotypic genus *Alphaletovirus* of the subfamily *Letovirinae* (with single species *Microhyla letovirus* 1), and four non-monotypic genera of subfamily *Orthocoronavirinae*, namely:

1. genus *Alphacoronavirus* with 12 subgenera: *Colacovirus* (monotypic), *Decacovirus* (two species), *Duvinacovirus* (monotypic), *Luchacovirus* (monotypic), *Minacovirus* (two species), *Minunacovirus* (two species), *Myotacovirus* (monotypic), *Nyctacovirus* (monotypic), *Pedacovirus* (two species), *Rhinacovirus* (monotypic), *Setracovirus* (two species) and *Tegacovirus* (monotypic) (Fig. 1, Supplemental Table S1);
2. genus *Betacoronavirus* with five subgenera: *Embecovirus* (four species), *Hibecovirus* (monotypic), *Merbecovirus* (four species), *Nobecovirus* (two species) and *Sarbecovirus* (currently monotypic) (Fig. 1, Table S1);
3. genus *Gammacoronavirus* with two monotypic subgenera: *Cegacovirus* and *Igacovirus* (Fig. 1, Table S1);
4. genus *Deltacoronavirus* with three monotypic subgenera *Andecovirus*, *Herdecovirus* and *Moordecovirus* and one non-monotypic subgenus *Buldecovirus* (four species) (Fig. 1, Table S1).

The viruses of *Coronaviridae*, such as the severe acute respiratory syndrome (SARS-CoV) and related Middle East respiratory syndrome (MERS-CoV), are highly pathogenic [8]. The recently discovered virus SARS-CoV-2, which is also a member of this family, causes the Coronavirus disease 2019 (COVID-19) that

was declared a pandemic by WHO in March 2020. Fifteen months later the total cases of COVID-19 are approaching 180,000,000 and cumulative deaths have exceeded 3,800,000 (www.who.int, June 2021). Thus, *Coronaviridae* is of high medical importance worldwide. Such importance has resulted in an urgency to further understand the relationships within the coronavirus family, and the viruses most closely related to SARS-CoV-2.

On the motivations of the study

The primary focus of this study is the utilization of all 39 ICTV-recognized species, with four different methods of analysis to produce basal-rooted phylogenetic trees, and validation of these trees as the natural hierarchical representations of relationships within *Coronaviridae*.

Second to this, the authors seek to add to the current and future discussion of the nomenclature used for virology, from the perspectives of the authors who are well-published on the nomenclature and classification of various living organisms, and the diversity of scientific fields represented by the authors from medicine to virology to biological classification.

1. Natural hierarchical representation of relationships within *Coronaviridae*.

a. The sampling of phylogenetic studies of *Coronaviridae* and rooting of the obtained trees.

Coronaviridae is a virus family with potentially high cryptic diversity at the species level [6, 7, 9], currently summarized by CSG in “the current classification of coronaviruses recognizes 39 species” [5, 7]. The recognition of these 39 species is the result of the multiple years of dedicated work by the members of CSG that involved numerous analyses of more than 2000 virus genomes, including the genomes of prospective species and the other virus taxa [7].

Many of the phylogenetic analyses in the literature of *Coronaviridae* to date have included use of unrooted or mid-point rooted trees, many including viruses from other families in the matrices, and some cases also using limited or arbitrary taxonomic sampling (recent examples include, [7, 10–13]).

Within conventional phylogenetic frameworks the root of cladogram can be defined as a most basal taxon of cladogram on which all characters were polarized [15]. Thus, rooting is a procedure of the *assigning* a root to a cladogram; this can be done either *a priori* or *a posteriory* to the results of the analyses [15]. Basal rooting is a means of restructuring data that results in more stable and rigorous hierarchical classification. By comparison, procedures such as mid-point rooting rely on artificial assumptions of the molecular clock [15] and are unnecessary if the actual outgroup taxon of some particular relationship is actually known or inferred with a high degree of likelihood. The simple presence of the other viruses within the working matrices of *Coronaviridae* is not a warranty of the basal rooting of the resulted trees, and there is no warranty at all that the trees will appear to be rooted “automatically” if no distant relatives of coronaviruses were included to the analyses [12, 13]. For example, in case of the application of the phylogenetic methods to the molecular matrices of *Coronaviridae* and prospective relatives [e.g., 10], the resulted tree(s) still must be rooted relatively to one of the *a posteriori* selected

outgroups. By operating this way, the shape of the final tree may be dramatically changed if compared with the initial outputs of phylogenetic analysis. For the *Coronaviridae* suitable outgroup taxa are known and can be used for basal rooting, and the resultant trees will represent a more stable hierarchical classification.

Both the lack of basal rooting and insufficient or biased sampling in phylogenetic studies can result in gaps in vital hierarchical information needed to understand the relationships between the viruses of this family. To address this urgent need, this study has undertaken a comprehensive taxonomic analysis of the family *Coronaviridae*, using the best taxonomic sampling available that is based on the genomes of all 39 ICTV-recognized coronavirus species. Additional viruses were also included to enable testing of the veracity of recently published proposals, included recently identified viruses such as SARS-CoV-2.

b. **On the cladistic analyses of *Coronaviridae*.**

Almost all available phylogenetic studies of coronaviruses have been based on the parametric Maximum Likelihood (ML) method. In order to increase the veracity of representations of the relationships within *Coronaviridae*, in addition to the ML method we have used the following three non-parametric cladistic approaches. The last two methods have not previously been used to resolve relationships within *Coronaviridae*.

1. Standard Maximum Parsimony (MP),
2. Three-taxon statement analysis (3TA), and the
3. Average Consensus (AC) analysis as applied to the array of the maximal relationships.

Methods of cladistic analysis may be of benefit to virology classification. The following provides some formal definitions to explain the relevance of the methods chosen in this study. “Phylogenetic tree” is a hypothesis of genealogical relationships among a group of taxa with specific connotations of common ancestry and a time axis [15]. Most trees that have been obtained in the phylogenetic studies of *Coronaviridae* aree defined as “phylogenetic trees” [7, 10–14]. A more general term is a “cladogram,” which is a branching diagram specifying hierarchical relationship among taxa based upon homologies [15]. For simplicity, cladogram can also be named as a “tree”. *But contrary to a phylogenetic tree per se, a cladogram in general includes no references to common ancestry and implies no time axis* [15].

Within conventional cladistics, Maximum Parsimony analysis, the cladogram may also be treated just as a phylogenetic tree [16]. However, this is not true within more general interpretations of cladistics [15, 17–22]. Under the latter interpretations, Maximum Parsimony must be considered as a method of classification that groups taxa hierarchically into nested sets representing such sets as a cladograms [15, 17–22]. The major goal of cladistic analysis is therefore a search for clades or monophyletic groups (relationships that are based solely on synapomorphies or true homologies [e. g., 15, 21, 22]). In its essence it is the extension of the comparative thinking that forms the heart of traditional biology (such as zoology or botany)[20, 21]. By comparison, as it has been clearly stated by the most well-known modern proponent of the statistic phylogenetics Joe Felsenstein [23, 24], that the Maximum likelihood (ML)

estimation is the evolutionary model-based method of statistical inference that involves finding the phylogenetic tree which yields the highest probability of evolving the observed data. All data (characters), not solely synapomorphies (or homologies), are used in forming a major focus of ML analysis (and of the other phenetic procedures [25]).

As much of the discussion within the field of virology taxonomy focuses on classification, methods of cladistics analysis may be of benefit to virologists as model-free methods of hierarchical classification *per se*, that focuses on synapomorphies (homologies), while simultaneously being very flexible for use in process-based explanations of the observed taxonomic patterns and its hierarchies.

c. **Monophyly and hypothesis on hierarchical relationships within *Coronaviridae*.**

The trees produced from the four methods used (Figs. 1–2, Supplemental Fig. S1–S4) enabled testing for monophyly of all available non-monotypic genera and infrageneric taxa of *Coronaviridae* to determine whether they are representative of a natural hierarchy. Confirmation of monophyly would validate the representation of relationships between all of the available genera/infrageneric taxa or particular virus species of *Coronaviridae* (incl. SARS-CoV-2).

The first assessment of monophyly was that the newly described monotypic genus *Alphaletovirus* (Table S1), a member of subfamily *Letovirinae*, is a sister group of family *Coronaviridae* [10].

The second monophyly assessment is that the general relationship within the *Coronaviridae* subfamily *Orthocoronavirinae* as a simple hierarchy:

((({*Alphacoronavirus*},{*Betacoronavirus*}){*Gammacoronavirus*}){*Deltacoronavirus*})[26]. This proposition is named the (((A, B) G) D) hypothesis of the general relationship within the subfamily. Unfortunately, this hypothesis [26] was originally based on limited taxonomic sampling.

Thirdly, the subgenus *Hibecovirus* has been placed as sister taxon of the presumably monotypic subgenus *Sarbecovirus* (for instance [7, 27]). This proposal is based on unrooted phylogenies, so it has been tested herein with special attention to the placement of SARS-CoV-2.

In the fourth proposal, the close relationship of SARS-CoV-2 with bat coronaviruses RaTG13 [28, 29] and RmYN02 [29, 30] as well as with pangolin coronavirus (reviewed in [12]) is also tested.

2. Discussion points regarding virology nomenclature.

The current taxa nomenclature for the family *Coronaviridae* was developed by the CSG and approved by the ICTV (see above). Around 2010 the CSG paused introduction of a binomical nomenclatures for viral species to allow discussion. Revision of the nomenclatures for viral species is likely to resume soon. It is therefore now timely to consider how the principles of binary nomenclature could be applied to *Coronaviridae*, and how this would impact future recognition of virus species and in particular those in the currently monotypic subgenus *Sarbecovirus*.

As authors from a diversity of scientific fields, from medicine, to virology and biological classification, we strongly support the adoption of a binary nomenclature for viruses [3, 4, 31]. Such nomenclature would be relatively easy to implement [31], and would be clear and practical to understand by people from a diversity of fields. We believe this would reduce difficulties in interpretation and synthesis of results arising from the usage of cumbrous abbreviations and trivial names of different coronaviruses used by virologists in various scientific publications, past and recent.

Our suggestions herein do not seek to create conflict but offer options for discussion that might minimize conflict and confusion in the future. For example, below we offered for discussion introduction of the rank of the tribe. Our paper does not formally describe new tribes, genera, or species, but aims to demonstrate a potential efficient and practical binary virus nomenclature for consideration in the making of future nomenclature decisions.

Materials And Methods

1. Taxonomic sampling of the study

Following others [32], all 39 ICTV-approved genomes of species of *Coronaviridae* have been used in this study, with essential information including virus names, correspondent abbreviations, proposed hosts (if any), the GenBank numbers etc. summarized in Table S1.

In addition, the following were included with the final alignments (see Table S1 for details):

a. the published genome MN908947 of SARS-CoV-2 [14];
b. an unpublished genome of the same virus species (MN988713);
c. the genome of bat SARS-like coronavirus (bat-SL-CoV-ZC45; MG772933) previously used in other published comparisons to SARS-CoV-2 (for instance, [13, 14]);
d. - f. bat coronavirus RaTG13 (MN996532) [28], pangolin coronavirus (MT121216) [12] and RmYN02 (GISAID: EPI_ISL_412977) [30], previously proposed to be the closest known relatives of SARS-CoV-2 (reviewed in [12, 29, 30, 33];
g. an additional SARS related genome ZS-B (AY394996);
h. the genome of recently discovered *Microhyla letovirus* 1 or MLeV virus (subgenus *Milecovirus*) [10], obtained upon courtesy request from Prof. B. W. Neuman (Texas A&M University-Texarkana, TX, US).
i. genus *Torovirus* (family Tobaniviridae, order Nidovirales) was assumed as the best outgroup taxon of *Coronaviridae* [8], and ICTV-approved genome of *Torovirus* (AY427798) has been selected for this study.

The names of major relationships on all trees obtained from the trees produced in this study (Figs. 1–2, Supplemental Fig. S1–S4) are written in italics due to the strong congruence with different taxonomic entities. Depending on the context, the names of the clades are provided in square and/or curved

brackets. As the utility of phylogenetic trees depends on their clarity, the use of abbreviations and trivial names of viruses has been avoided whenever possible.

2. Matrices

All genomic alignments have been performed using MAFFT [34–36] following FFT-NS-I strategy with the command: mafft --inputorder --adjustdirection --anysymbol --kimura 1 --maxiterate 1000 --6merpair input.

To removed poorly aligned positions from the obtained genomic alignments we used the simple, easy and efficient program G-block [37] as implemented in SeaView [38], thereby solving the saturation problem with the alignment of virus genomes [39] *directly* through the moving of the saturated sites out of the molecular matrix. Moving these sites out of consideration makes future analyses relatively noise-less. Based on the high accuracy and efficiency of this method, we focused on the G-Block-based whole genome alignment of *Coronaviridae* plus *Torovirus*, rather than be limited to the codon-based alignment of the same group of taxa.

Under the conditions of "less stringent" strategy of the G-block algorithm [37, 38], the saturated positions have been successfully removed from

1. the genomic alignment of *Coronaviridae* plus *Torovirus,* and
2. the genomic alignment of subfamily *Orthocoronavirinae* with no outgroups included.

The G-block version of the genomic alignment of *Coronaviridae* plus *Torovirus* was also established as a binary matrix using simple "presence – absence" coding (reviewed in [15]) with the future inclusion of the "all-plesiomorphic" ("all-zero") artificial taxon. In short, G-block based genomic alignment of the family *Coronaviridae* plus *Torovirus* was rewritten as a binary (01) matrix, where "zero" means "the absence of a nucleotide in this particular position of the alignment", and "one" means "the presence of the nucleotide in this particular position of the same alignment". For example, if the character-state of the character number 253 is equal to A (Adenine), this can be written as "1000", where "1" means "the A is present in position 253", and subsequent "0" indicates that U(T), G and C are simultaneously absent on the same position.

Assuming the "absence of the nucleotide" (the character-state "zero") is a plesiomorphic character-state, we can add to the binary matrix "all-plesiomorphic" or "all-zeros" outgroup. The binary matrix with an "all-zeros" outgroup added was later used as an input into the script Forester v. 1.0 [22] following the command ruby trees.rb path_to_matrix_file with future selection of the "Additional" forest of the maximal trees (relationships) [22] for Average Consensus (AC) analysis [40, 41, 22].

Manipulations with either the molecular or binary matrices and the tree-files have been performed with Mesquite v. 3.51 [42], PAUP* v. 4.0a [43] and FigTree v. 1.4.4 [44].

3. Trees and analyses

The G-block version of the molecular alignment of *Coronaviridae* plus *Torovirus* was analyzed by the standard Maximum Parsimony (MP) approach (Fitch Parsimony, reviewed in [15]), and by the three-taxon statement analysis (3TA) with fractional weighting (reviewed in [15, 20, 21, 45], as implemented in [45]).

Following the logic of Williams-Siebert (WS) representation of the unordered multistate data (reviewed in [45]) the three-taxon statement (3TS) permutations of the G-block-based alignment of *Coronaviridae* plus *Torovirus* were conducted with TAXODIUM version 1. [45] using the command: taxodium.exe input_file_name.csv –idna –ob –og –fw –nex taking values of the operational outgroup as equal to the values of *Torovirus*.

All MP analyses have been performed in PAUP* [43] as in [22, 45]. The resulted most parsimonious tree resulted standard MP analysis was *a posteriori* rooted relative to *Torovirus*.

The AC of the array of maximal trees was calculated using the program Clann version 4.1.5 [46, 47] as described in [22]. The distance optimality criterion for the AC analysis was specified as a “distance with non-weighted least squares” [40, 41].

Following others [48], Maximum Likelihood (ML) analysis of G-block alignment of *Coronaviridae* plus *Torovirus* was conducted with W-IQ-TREE [49] with implemented automatic model selection procedure [49]. The resulted most probable tree was *a posteriori* rooted relative to *Torovirus*.

The MP Bootstrap support (BS) values have been calculated as described in [45]. In the case of the ML analysis, the Approximate Likelihood ratio Tests (aLRT) support values (reviewed in [49]) have been calculated instead of the ML BS supports, as implemented W-IQ-TREE [49].

The simplest “total” character differences between the G-block modified aligned genome sequences of subfamily *Orthocoronavirinae* (Supplemental Table S2), as well as between three aligned genomes of bat coronaviruses RmYN02, RaTG13 and severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) (MN908947) with no aligned positions excluded (Table S3), was calculated in PAUP* [43] under the default options.

Results

1. Matrices.

The genomic alignment of *Coronaviridae* + *Torovirus* (outgroup) consists of 52,990 molecular characters (base pairs (bp)), the G-block version of this alignment is of 22,489 characters with 19,550 of those are parsimony-informative. This alignment was utilized in MP, 3TA, AC and ML analyses.

The genomic alignment of subfamily *Orthocoronavirinae* with no outgrop included is of 49,881 bp. The G-block version of this alignment consists of 23,431 bp. This alignment was used to calculate the simplest pairwise distances (“total character differences” (TCD) between all of the members of the subfamily included in the analyses (Table S2). The genomic alignment of bat coronaviruses RmYN02, RaTG13 and

severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) (MN908947) is of 29,907 molecular characters. This alignment with no characters excluded was used to calculate the total character differences (Table S3) between newly discovered SARS-CoV-2 and two of its closest relatives.

The TCD is the simplest expression of the pairwise distance between aligned molecular sequences and indicates solely the total number of different single nucleotide positions (or SNPs) between them. For example, the number 1138 in the Table S3 means that the aligned genomes of human coronavirus SARS Cov 2 (accession “a”) and the bat coronavirus RaTG13 are different from each other by 1138 SNPs. For instance, in position # 37 of the same molecular alignment the value of SARS Cov 2 is equal to “C” and the value of RaTG13 is equal to “G”. The total number of such SNPs equals 1138.

The standard MP analysis of the 22,489 bp G-block alignment resulted in the single most parsimonious tree of 208,176 steps (CI = 0.2505, RI = 0.4954) (Fig. S1).

The 3TS representation of the same 22,489 bp G-block alignment resulted 39,621,820 3TSs (binary characters), all parsimony-informative and fractionally weighted, with the most parsimonious fit of 20,786,459.7424 steps (RI = 0.5706) (Fig. S2).

The presence-absence re-coding of the genomic alignment of *Coronaviridae* + *Torovirus* resulted in a matrix of 73,258 binary characters from which 65,435 characters can be established as rooted trees (maximal relationships). The AC analyses of the forest of these 65,435 relationships resulted in a single AC tree of the score 0.00911 (Fig. S3).

For ML analysis of the 22,489 bp G-block matrix of *Coronaviridae* and outgroup (*Torovirus*), GTR+F+R10 model has been automatically selected by W-IQ-TREE as a best-fit model based on either corrected or non-corrected Akaike Information Criteria, as well as on Bayesian Information Criterion. The resulted single most probable (ML) tree has the best score (log likelihood) equal to -766940.2344 (Fig. S4).

All obtained alignments contains the huge number of variable characters and every *Coronaviridae* virus seems to separate from the others sometimes by hundreds or (more commonly) thousands of SNPs (Tables S2 and S3). For example, the minimal relationship {RmYN02 {RaTG13 plus SARS-CoV-2}} (Figs. S2–S3), based on the 29,907 bp complete genomic alignment of three taxa (RmYN02, RaTG13 and SARS-CoV-2), implies 1,329 informative SNPs (or, respectively, 1,329 3TSs) from the total 2,467 variable characters.

2. Trees and obtained relationships.

Hierarchical classification of the coronaviruses (*Riboviria*, *Nidovirales*, *Coronaviridae*) was established as a simplified Strict Consensus of four trees, produced by three different methods of cladistic analysis as well as by Maximum Likelihood (Fig. 3). All four trees have several major consistencies, as shown in the Strict Consensus (Figs. 1–3, Supplemental Fig. S1–S4).

a. **General pattern of the relationships within *Coronaviridae* and the placement of *Microhyla letovirus* 1** (Figs. 1–3, Supplemental Fig. S1–S4, Table S1).

The results of all analyzes have demonstrated that the hierarchy ((({*Alphacoronavirus*}, {*Betacoronavirus*}){*Gammacoronavirus*}){*Deltacoronavirus*}), with *Microhyla letovirus* 1 (subgenus *Milecovirus*, genus *Alphaletovirus*, subfamily *Letovirinae*) which has been defined as its sister taxon, form a general pattern of the relationship within *Coronaviridae*.

b. **{*Alphacoronavirus*} clade** (Figs. 1–3, Supplemental Fig. S1–S4)

{*Alphacoronavirus*} clade includes all known species of the subgenus *Alphacoronavirus.* Within this clade, all four trees showed that the following taxa are sisters:

a. Ferret coronavirus and Mink coronavirus 1 (subgenus *Minacovirus*);
b. Porcine epidemic diarrhea virus and *Scotophilus* bat coronavirus 512 (subgenus *Pedacovirus*);
c. Human coronavirus NL63 and NL63-related bat coronavirus BtKYNL63-9b (subgenus *Setracovirus*);
d. Bat coronavirus HKU10 and *Rhinolophus ferrumequinum* alphacoronavirus HuB-2013 (subgenus *Decacovirus*);
e. *Miniopterus* bat coronavirus 1 and *Miniopterus* bat coronavirus HKU8 (subgenus *Minunacovirus*).

Thus, within {*Alphacoronavirus*} clade, we were able to find five smaller clades (subclades): {*Decacovirus*}, {*Minacovirus*}, {*Minunacovirus*}, {*Pedacovirus*} and {*Setracovirus*}with the exact correspondence of each of these clades to the previously established subgenera of the genus *Alphacoronavirus.*

From our results (Figs. 1–3, Supplemental Fig. S1–S4) the following is also clear:

a. monotypic subgenus *Colacovirus* (Bat coronavirus CDPHE15) is a sister of *Pedacovirus* clade in all of the trees;
b. monotypic subgenus *Duvinacovirus* (Human coronavirus 229E) is a sister of the *Setracovirus* clade in all of the trees, and
c. monotypic subgenus *Tegacovirus* (Alphacoronavirus 1) is a sister of the *Minacovirus* clade in all of the trees.

The phylogenetic placement of monotypic subgenera *Luchacovirus* (Lucheng Rn rat coronavirus), *Myotacovirus* (*Myotis ricketti* alphacoronavirus Sax-2011) and *Rhinacovirus* (*Rhinolophus* bat coronavirus HKU2) depends on the method of the analyses (Figs. 1–3, Supplemental Fig. S1–S4). It is worth stressing, however, that the standard MP (Fig. S1), as well as Cladistic methods (Figs. S2–S3), but not the ML method (Fig. S4) have placed *Luchacovirus* as a sister of the {*Alphacoronavirus*}.

c. **{*Betacoronavirus*}** (Figs. 1–3, Supplemental Fig. S1–S4).

{*Betacoronavirus*} clade includes all known species of the subgenus *Betacoronavirus.* All of the analyses argue in favor of the general simple hierarchical relationship ({*Embecovirus*} ({*Merbecovirus*} ({*Nobecovirus*} (*Hibecovirus*, {*Sarbecovirus*})))) within *Betacoronavirus* clade (Figs. 1–3, Supplemental Fig. S1–S4).

The relationships of four species of the subgenus *Embecovirus* (Betacoronavirus 1, China Rattus coronavirus HKU24, Human coronavirus HKU1 and Murine coronavirus), that formed a clade with the same name {*Embecovirus*}, the sister of {*Betacoronavirus*} clade, depend on the method of the analysis (Figs. 1–3, Supplemental Fig. S1–S4)

All four species of subgenus *Merbecovirus* form a clade {*Merbecovirus*}. In all trees, the Hedgehog coronavirus 1 is a sister to the clade that contains three other members of the subgenus *Merbecovirus*: namely, Middle East respiratory syndrome-related coronavirus, *Pipistrellus* bat coronavirus HKU5, *Tylonycteris* bat coronavirus HKU4 (Figs. 1–3, Supplemental Fig. S1–S4).

The {*Sarbecovirus*} clade that corresponds to the subgenus *Sarbecovirus* includes the viruses of severe acute respiratory syndrome-related coronavirus (SARS), the newly discovered monophyletic SARS-CoV-2 (two accessions have been included to the analyses, SARS-CoV-2a and SARS-CoV-2b), as well as CoV-ZC45 and SARS Cov ZS B.

Clade {*Sarbecovirus*-SARS plus SARS Cov ZS B} is a sister of the remaining species of *Sarbecovirus*.

Depending on the analysis, either bat coronaviruses RmYN02 or RaTG13 have been placed as a sister of SARS-CoV-2 (Figs. 1–3, Supplemental Fig. S1–S4). The pangolin coronavirus (isolate MP789) has been defined as a sister of the clade {RmYN02 plus RaTG13 plus SARS-CoV-2} in all of the analyses (Figs. 1–3, Supplemental Fig. S1–S4).

All trees define the monotypic subgenus *Hibecovirus* (Bat Hp-betacoronavirus) as a sister of {*Sarbecovirus*} clade.

Two members of subgenus *Nobecovirus,* namely *Rousettus* bat coronavirus GCCDC1 and *Rousettus* bat coronavirus HKU9 are sister taxa (Figs. 1–3, Supplemental Fig. S1–S4).

d. **{*Gammacoronavirus*} clade** (Figs. 1–3, Supplemental Fig. S1–S4)

Two subgenera of the genus *Gammacoronavirus*, namely subgenus *Cegacovirus* (with the single species Beluga whale coronavirus SW1) and subgenus *Igacovirus* (with a single species Avian coronavirus) formed a clade {*Gammacoronavirus*} in all of the analyses (Figs. 1–3, Supplemental Fig. S1–S4).

e. **{*Deltacoronavis*} clade** (Figs. 1–3, Supplemental Fig. S1–S4).

Five subgenera of genus *Deltacoronavis,* namely subgenus *Andecovirus* (with single species Wigeon coronavirus HKU20), subgenus *Buldecovirus* (1-4) (with four species: Bulbul coronavirus HKU11, Coronavirus HKU15, Munia coronavirus HKU13 and White-eye coronavirus HKU16), monotypic

subgenus *Herdecovirus* (Night heron coronavirus HKU19) and monotypic subgenus *Moordecovirus* (with single species Common moorhen coronavirus HKU21), have formed the {*Deltacoronavis*} clade in all of the analyzes. Also, all of the analyses argue in favor of the simplest hierarchy of the relationships within this clade: (*Andecovirus* (*Herdecovirus* (*Moordecovirus* ({*Buldecovirus*})))). Within {*Buldecovirus*} clade Coronavirus HKU15 and Munia coronavirus HKU13 appeared to be sisters in all of the analyses, the relationships between the other members of the {*Buldecovirus*} clade depends on the method of the analysis. Monotypic subgenus *Andecovirus* (Wigeon coronavirus HKU20) is a sister of {*Deltacoronavis*} clade.

f. **On monophyly of non-monotypic taxa of *Coronaviridae*** (Figs. 1–3, Supplemental Fig. S1–S4).

As demonstrated above, all four trees show that all four current subgenera of subfamily *Orthocoronavirinae* (family *Coronaviridae*), namely subgenus *Alphacoronavirus*, subgenus *Betacoronavirus*, subgenus *Deltacoronavirus*, and subgenus *Gammacoronavirus*, are monophyletic (Figs. 1–3, Supplemental Fig. S1–S4).

All non-monotypic subgenera of all four genera of subfamily *Orthocoronavirinae* (family *Coronaviridae*), namely subgenus *Decacovirus* (genus *Alphacoronavirus*), subgenus *Minacovirus* (genus *Alphacoronavirus*), subgenus *Minunacovirus* (genus *Alphacoronavirus*), subgenus *Pedacovirus* (genus *Alphacoronavirus*), subgenus *Setracovirus* (genus *Alphacoronavirus*), subgenus *Embecovirus* (genus *Betacoronavirus*), subgenus *Merbecovirus* (genus *Betacoronavirus*), subgenus *Nobecovirus* (genus *Betacoronavirus*), and subgenus *Buldecovirus* (genus *Deltacoronavirus*) are monophyletic in all of the analyses (Figs. 1–3, Supplemental Fig. S1–S4).

Discussion

Several phylogenies of *Coronaviridae* (or parts thereof) have been previously published. However, none have been based on all 39 ICTV-recognized species of the family. In addition, almost all were either unrooted or mid-point rooted, or taxonomically incomplete (for example, [10–14]). On occasions when the phylogenetic tree appears as rooted (for instance, [26, 50]) the taxonomic sampling of some such studies was incomplete or arbitrary. Finally, the majority of published data is based on the ML method of analysis. The study presented herein has sought to address these important gaps in the current presentation of relationships within *Coronaviridae*.

1. Rigorous rooted trees validate tested proposals of the relationships within *Coronaviridae*

The basal rooted trees produced by the four methods herein allow for the effective testing of various proposals regarding the relationships of viruses within the *Coronaviridae* family.

1. We confirm (Figs. 1–3, Supplemental Fig. S1–S4) that two subfamilies of *Coronaviridae*, namely subfamily *Letovirinae* (family *Abyssoviridae*) and subfamily *Orthocoronavirinae* are sisters [10]. This solution, which was previously based on the unrooted or mid-point rooted trees and limited

taxonomic sampling, is consistent with the familial rank of both taxa [10]. We would recommend to accept the monotypic subfamily *Letovirinae* at the familiar rank [10].

2. Our results (Figs. 1–3, Supplemental Fig. S1–S4) clearly argue in favor of (((A, B) G) D) hypothesis of the general pattern of relationship within subfamily *Orthocoronavirinae* [26 = Tokarz R et al 2015]. This previously published proposition was based on limited taxonomic sampling and an unrooted or mid-point rooted network [26]. With basal placement of subgenus *Milecovirus* (MLeV) (Figs. 1–3, Supplemental Fig. S1–S4), we can extend this pattern to the relationship: ((((A, B) G) D) MLeV). This natural hierarchy within family *Coronaviridae* is in principle congruent to the general pattern of their hosts: (((Mammals) Birds plus Mammals) Amphibia) (Table S1).
3. Subgenus *Hibecovirus* has been proposed as a sister taxon of the subgenus *Sarbecovirus* (reviewed in [27]). This proposal, made on the basis of unrooted or mid-point rooted trees, was reexamination using the comprehensive basal rooted trees produced in this study. All four methods validate the sister relationship of the subgenera *Hibecovirus* and *Sarbecovirus* clade (Figs. 1–3, Supplemental Fig. S1–S4).
4. Due to particular interest in SARS-CoV-2 across diverse scientific fields, comment on the relationships with relation to SARS-CoV-2 are of interest. SARS-CoV-2Importantly, all four methods have placed pangolin coronavirus as a sister of the relationship {RaTG13 plus RmYN02 plus SARS-CoV-2} (Figs. 1–3, Supplemental Fig. S1–S4), confirming that bat coronaviruses RaTG13 or RmYN02, but not the pangolin coronavirus, are indeed the closest relatives of SARS-CoV-2 (Figs. 1–3, Supplemental Fig. S1–S4). In other words, contrary to some recent suggestions (reviewed in [33, 51]) the basal rooted trees produced herein confirm that either bat coronavirus RaTG13 [28, 29] or RmYN02 [29, 30], but not the pangolin coronavirus [12, 33], is an immediate sister of SARS-CoV-2 (Figs. 1–3, Supplemental Fig. S1–S4).

Differences in details within clade relationships do exist between the ML method and each of the cladistic methods newly applied to molecular sequence data of *Coronaviridae*. For example the MP, AC and 3TA trees (Figs. S1–S3), but not ML trees (Fig. S4), have placed Lucheng Rn rat coronavirus (genus *Alphacoronavirus*, subgenus *Luchacovirus*) as a sister taxon to the clade that contains all of the remaining members of genus *Alphacoronavirus*.

Similarly, both conventional phylogenetic methods (MP and ML) (Figs. S1, S4) have defined bat coronavirus RmYN02 as a weakly supported sister of the SARS-CoV-2 (Figs. S1, S4). However, both Hennigian methods (3TA and AC, as applied to the array of the basal rooted trees, each corresponding to the binary representation of the standard molecular characters) (Figs. S2–S3), in contrast, placed bat coronavirus RaTG13, but not RmYN02, as a sister of this newly discovered coronavirus (Figs. S2–S3).

All four analyses are initially based on a common molecular matrix (22,489 bp G-Block version of the 47-genomes alignment). Differences between analyses are likely a result of how each method deals with conflict to form the optimal trees. Nevertheless, the similarity between the tree topologies suggests that,

regardless of method, many of the nodes are 'true' summaries of the data and that the data themselves are relatively noise-free (Figs. 1–3, Supplemental Fig. S1–S4).

2. On relationships, taxonomy and hosting of newly discovered *Sarbecovirus* SARS-CoV-2

From the elementary comparative point of view, it is clear that every *Coronaviridae* virus seems to be well defined, with each separated from the others by hundreds or more frequently thousands of SNPs (Tables S2 and S3). The closely related viruses from the *Sarbecovirus* clade (including the newly discovered SARS-CoV-2 and bat coronaviruses RaTG13 and RmYN02), are also all remarkably different from one another from the comparative molecular standpoint (see the Results and Supplementary Tables S2 and S3 for more detail). The same is true for every relationship within *Coronaviridae* we have recovered in our analyses (Tables S2 and S3).

Such simple observations automatically exclude the possibility of simple or recent recombination origins of SARS-CoV-2 [12, 51–53] as well as other semantically similar propositions [51]. Based on the data available to date, including the comprehensive trees produced herein (Figs. 1–5, Fig. S1), the focus should shift to the comparative aspect of the problem.

The monophyly of all the genera of *Coronaviridae*, as well as all of its non-monotypic subgenera, and the general relationship ((((A, B) G) D) within the subfamily *Orthocoronavirinaeis*, can also be demonstrated within the pure comparative analytical framework (Fig. S2–S3). Within the later, no member of *Sarbecovirus* clade is an ancestor of SARS-CoV-2. This comparative view may be critical in discussing the general simple pattern in hosting of SARS-CoV-2.

Recent studies of SARS-related coronaviruses have suggested that bats harbor close relatives to SARS-CoV-2 (for example, [13, 52]), and that pangolins may be natural hosts of this member of genus *Betacoronavirus* ([33, 54], see also [12] for the review), leading to the hypothesis of animal to human transmission of SARS-CoV-2 ("The presence in pangolins of an RBD very similar to that of SARS-CoV-2 means that we can infer this was also probably in the virus that jumped to humans" [33].

However, the search of other hosts, as well as the related exotic ways of transmission of SARS-CoV-2 from these hypothetical hosts to humans, is based on a set of the complicated assumptions (for example, "The sequence similarity in the spike receptor binding domain between SARS-CoV-2 and a sequence from pangolin is probably due to an ancient intergenomic introgression..." [54]) and ignores the possibility of the original human-based natural hosting of SARS-CoV-2. The last proposal is semantically similar to propositions of the original bat-based natural hosting of the severe acute respiratory syndrome-related coronavirus (SARS-CoV), or its closest relatives (Table S1).

In fact, the hosting of the viruses that are genetically related to SARS-CoV-2 by bats or pangolins is not, strictly speaking, an argument in favor of the animal hosting of SARS-CoV-2, especially as the latter virus has not yet been detected in animals such as bats or pangolins (see also [55]).

The clear possibility of the original human hosting of SARS-CoV-2, in other words of the *endemism* of recently discovered SARS-CoV-2, unfortunately, is yet to be discussed in the scientific literature. Furthermore, the extremely high contagiousness of SARS-CoV-2 is arguable unlikely to have arisen within only 1.5 years of transmission from animals to humans. The alternative is that a form of SARS-CoV-2 may have existed in some parts of the human population before the recognized origin of the current pandemic (either within a geographically isolated group or a particular age group, such as infants) but was less pathological or perhaps more effectively suppressed by the human immune system for some time prior to acquisition of increased pathogenicity.

Seven human coronaviruses have been identified to date (reviewed in [56]). Examples of four common, human endemic, coronavirus are Human coronavirus 229E from subgenus *Duvinacovirus* (genus *Alphacoronavirus*), Human coronaviruses OC43 (HCoV-OC43) and HKU1 (HCoV-HKU1) from subgenus *Embecovirus* (genus *Betacoronavirus*) and Human coronavirus NL63 (HCoV-NL63) from subgenus *Setracovirus* (genus *Betacoronavirus*) (Table S1)[56–61]. These are globally distributed viruses where no animal hosts have even been proposed [56–58], and may be associated with severe pathogenesis in the lower respiratory tract, such as bronchiolitis or pneumonia [58, 59].

Anthroponotic transmissions of SARS-CoV-2 do occur (reviewed in [62–64], but can be easily be considered secondary transmission. Recent conclusions state so far little is known concerning the role of pets and other animals (such as hamsters, minks, ferrets, lions, monkeys, tigers and some others) in the transmission of COVID-19 [63]. Thus, “as of now, there is no strong evidence for natural animal-to-human transmission or sustained animal-to-animal transmission of SARS-CoV-2” [62].

With the information available to date, there is no direct evidence to suggest which hypothesis is true: 1) animal to human transmission (either recently or less recently [29, 54], with or without the involvement of intermediate hosts [33]), or 2) *the original human hosting*, as proposed in this paper. The evidence only tells us how viruses are related within a hierarchical classification (Fig. 1–2).

3. On taxonomy, classification, naming, and nomenclature of the coronaviruses

Because the taxonomy and nomenclature of the viruses are still “under construction” [1–7], the names of the virus species frequently remain non-binary, even within the taxonomic statements of ICTV or GSG [1–7] (Fig. 1, Table S1). Simultaneously, the current circumscriptions of the two largest genera of the *Coronaviridae* (*Alphacoronaviruses* and *Betacoronaviruses*) are very complicated. For example, current genus *Alphacoronavirus* currently circumscribes 12 subgenera, and current genus *Betacoronavirus* circumscribes five subgenera. These two observations cause issues with a clear naming of viruses on the phylogenetic trees of *Coronaviridae*, as well as with the reading of these trees, especially by non-specialists.

Herein we resolved these issues by using the ICTV summaries [1–5] of the **subgeneric** names of *Coronaviridae* as the basic units of our notation, where possible, in the analyses and trees (Fig. 2, Supplemental Fig. S1–S4). We partly avoided the nomenclature methods traditional for virologists,

instead demonstrating the possibilities of a simple and clear binary nomenclature of the coronavirus family (Fig. 2, Supplemental Fig. S1–S4). Accepting traditional genera of the family at the rank of a tribe and, simultaneously, the current subgenera (all of which are monophyletic (Figs. 1–3, Supplemental Fig. S1–S4) at the generic rank seems to be an easy heuristic way to incorporate the Linnaean principles of the binary nomenclature to the classification of the family literally “in one step” (Fig. 2).

This solution also implying the future recognition of numerous cryptic virus species, particularly in currently monotypic subgenus *Sarbecovirus* (Fig. 2, Supplemental Fig. S1–S4). For instance, the intra-species variation within the SARS-Cov, the current single species of subgenus *Sarbecovirus* (the same-name clade)(subgenus *Betacoronavirus*) is comparable with inter-species variation within the other clades/subgenera of coronaviruses (e. g., within current subgenus/clade *Merbecovirus*) (Table S2). Such simple comparison may also be easily connected with the potential endemism of SARS-Cov2 (as explained above, no other *natural* hosts besides humans have been proposed for this virus) as well as with clear morphological differences between SARS-Cov and SARS-Cov2 [65].

As a point for discussion, based on this suggestion it may be valid to treat newly discovered "severe acute respiratory syndrome-related coronavirus 2" as a prospective new species of *Sarbecovirus*. In doing so the coronavirus SARS-CoV-2 might be named a ***Sarbecovirus* species, abbreviation: sp. (e.g.: *Sarbecovirus* sp.)**, where “species” (sp.) is any available epithet, even a trivial one. Numerous currently discovered variants or strains of it (summarized in [11]) can be *in principle* established as forms or varieties of the same species. The trivial name “severe acute respiratory syndrome-related coronavirus 2” and correspondent abbreviation “SARS-CoV-2” could be listed in the description of the species. The same operations could be done with any other viruses of *Coronaviridae* (Fig. 2). Using such suggestions the binary renaming of all of the 39 species of the family, as well as other related viruses (either known or cryptc), might not be a difficult task.

Additionally, for the purpose of discussion, we suggest consistent involvement of the type method into the taxonomy of the viruses. As a simple example, we would like to suggest that the ICTV approved GenBank Accession number (ideally the reference to the whole genome of the virus) be incorporates as a nomenclature type for any virus species. For example the GenBank Accession number MN908947 could be treated as a nomenclature type of newly described *Sarbecovirus* SARS-CoV-2. Such a number implies the name/abbreviation of the biological isolate as well as other useful information. If adopted, the higher nomenclature categories (tribes, genera, families etc.) could be typified by the names of the species (genera etc.), exactly in a manner of plant or animal names. The nomenclature classification of plants and animals has been developed over hundreds of years, and as such is robust and well tested. Adopting the Linnaean binary nomenclature for viruses will increase the universality of the system, and thereby lead to more consistent information content and information exchange across and within disciplines.

We believe that such simple recommendations, as well as our Hierarchical Classification of the family *Coronaviridae* (Fig. 3), are fully consistent with the currently established principles of the future binary nomenclature and classification of viruses [1–7] and may be useful for different virus families.

Declarations

Funding: Not applicable

Conflict of interest / Completing interests: Not applicable.

Availability of data and materials: Data is available upon request from the first author

Code availability: No specific code has been developed during this study

Authors contributions as per CRediT taxonomy: E.V.M.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. M.L.T.: Conceptualization, Investigation, Data Curation, Writing – original draft, Writing – review & editing, Supervision. N.E.M.: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Resources, Writing – review & editing, Visualization. LS: Investigation, Writing – review & editing. M.C.E.: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision. D.M.W.: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision.

References

1. Adams MJ et al (2017) 50 years of the International Committee on Taxonomy of Viruses: progress and prospects Arch. Virol 162:1441–1446. https://doi.org/10.1007/s00705-016-3215-y
2. Walker PJ et al (2019) Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019) Arch. Virol 164:2417–2429. https://doi.org/10.1007/s00705-019-04306-w
3. Siddell SG et al (2020a) Binomial nomenclature for virus species: a consultation Arch. Virol 165:519–525. https://doi.org/10.1007/s00705-019-04477-6
4. Siddell SG et al (2020b) Correction to: Binomial nomenclature for virus species: a consultation. Arch Virol 165:1263. https://doi.org/10.1007/s00705-020-04555-0
5. Siddell SG et al (2019) Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). Arch. Virol. 164: 943–946 (2019)
6. Gorbalenya AE et al (2020b) [International Committee on Taxonomy of Viruses Executive Committee], The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol 5:668–674. https://doi.org/10.1038/s41564-020-0709-x
7. Gorbalenya AE et al (2020a) [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. https://doi.org/10.1038/s41564-020-0695-z

8. MacLachlan NJ, Dubovi EJ, Barthold SW, Swayne DF, Winton RJ (2016) Fenner's Veterinary Virology, 5th edn. Elsevier Inc, Amsterdam
9. Edgar RC et al (2020) Petabase-scale sequence alignment catalyses viral discovery BioRxiv https://doi.org/10.1101/2020.08.07.241729
10. Bukhari K, Mulley G, Gulyaeva AA, Zhao L, Shu G, Jiang J, Neuman BW (2018) Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the *Coronavirinae*. the proposed genus Alphaletovirus Virology 524:160–171. https://doi.org/10.1016/j.virol.2018.08.010
11. Candido DS et al (2020) Evolution and epidemic spread of SARS-CoV-2. in Brazil Science 369:1255–1260. https://doi.org/10.1126/science.abd2161
12. Liu P et al (2020) Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog 16:e1008421. https://doi.org/10.1371/journal.ppat.1008421
13. Lu R et al (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574. https://doi.org/10.1016/S0140-6736(20)30251-8
14. Wu F et al (2020) A new coronavirus associated with human respiratory disease in. China Nature 579:265–269. https://doi.org/10.1038/s41586-020-2008-3
15. Kitching IJ, Forey P, Forey PL, Humphries C, Williams D (1998) Cladistics: the theory and practice of parsimony analysis, 2nd edn. Oxford University Press, Oxford
16. Farris JS (1983) The logical basis of phylogenetic analysis. In: Platnick NI, Funk V (eds.) Adv. cladistics, 2. Proceedings of the 2nd meeting of the Willi Hennig society; Ann, Arbor, Mich., USA, Oct. 1–4; 1981. Columbia University Press, New York, pp. 7–36
17. Nelson G, Platnick NI (1981) Systematics and biogeography: cladistics and vicariance. Columbia University Press, New York
18. Nelson G, Platnick NI (1991) Three-taxon statements: a more precise use of parsimony? Cladistics 7:351–366. https://doi.org/10.1111/j.1096-0031.1991.tb00044.x
19. Brower AVZ (2000) Evolution is not a necessary assumption of cladistics. Cladistics 16:143–154. https://doi.org/10.1111/j.1096-0031.2000.tb00351.x
20. Williams DM, Ebach MC (2008) Foundations of systematics and biogeography. Springer, New York
21. Williams DM, Ebach MC (2020) Cladistics: A guide to biological classification, Systematics Association Special Volume Series, Series Number 88, 3rd edn. Cambridge University Press, Cambridge
22. Mavrodiev EV, Dell C, Schroder L (2017) A laid-back trip through the. Hennigian Forests PeerJ 5:e3578. https://doi.org/10.7717/peerj.3578
23. Felsenstein J (1981) Evolutionary trees from DNA sequences: A Maximum Likelihood approach. J Mol Evol 17:368–376. https://doi.org/10.1007/BF01734359

24. Felsenstein J (2004) Inferring phylogenies, Oxford University Press. Sinauer Associates Inc, Sunderland. https://www.sinauer.com/media/wysiwyg/tocs/InferringPhylogenies.pdf
25. Williams DM, Ebach MC, Wheeler QD (2010) Beyond belief: The steady resurrection of phenetics. In: Williams DM, Knapp S (eds) Beyond cladistics: The branching of a paradigm. University of California Press, Berkeley, pp 169–196
26. Tokarz R et al (2015) Discovery of a novel nidovirus in cattle with respiratory disease. J Gen Virol 96:2188–2193. https://doi.org/10.1099/vir.0.000166
27. Wong ACP, Li X, Lau SKP, Woo PCY (2019) Global epidemiology of bat coronaviruses Viruses 11:174. https://doi.org/10.3390/v11020174
28. Zhou P et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7
29. Pipes L, Wang H, Huelsenbeck J, Nielsen R (2020) Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny BioRxiv https://doi.org/10.1101/2020.06.19.160630
30. Zhou H et al (2020) 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. https://doi.org/10.1016/j.cub.2020.05.023
31. Postler T et al (2017) Possibility and challenges of conversion of current virus species names to Linnaean binomials. Syst Biol 66:463–473. https://doi.org/10.1093/sysbio/syw096
32. Helmy YA, Fawzy M, Elaswad A, Sobieh A, Kenney SP, Shehata AA (2020) The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control J. Clin Med 9:1225. https://doi.org/10.3390/jcm9041225
33. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2 Nat. Med 26:450–452. https://doi.org/10.1038/s41591-020-0820-9
34. Katoh K, Misawa K, Kuma Ki, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform Nucleic Acids Res. 30:3059–3066. https://doi.org/10.1093/nar/gkf436
35. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability Mol. Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
36. Miller MA, Pfeiffer W, Schwartz T (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Gateway Computing Environments Workshop (GCE) 14 Nov. 2010, Curran Associates, Inc. New Orleans, pp 1–8. https://doi.org/10.1109/GCE.2010.5676129
37. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis Mol. Biol Evol 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334
38. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building Mol. Biol Evol 27:221–224. https://doi.org/10.1093/molbev/msp259

39. Krishnamurthy SR, Wang D (2017) Origins and challenges of viral dark matter. Virus Res 239:136–142. https://doi.org/10.1016/j.virusres.2017.02.002
40. Lapointe FJ, Cucumel G (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa Syst. Biol 46:306–312. https://doi.org/10.1093/sysbio/46.2.306
41. Lapointe FJ, Levasseur C (2004) Everything you always wanted to know about the average consensus, and more. In: Bininda-Emonds ORP (ed) Phylogenetic supertrees: Combining information to reveal the tree of life. Kluwer Academic Publisher, Dordrecht, pp 87–105
42. Maddison WP, Maddison DR (2019) Mesquite: a modular system for evolutionary analysis Version 3.61 http://www.mesquiteproject.org
43. Swofford DL (2002) PAUP*. Phylogenetic analysis using Parsimony (*and other methods). Sinauer Associates, Sunderland. https://paup.phylosolutions.com
44. Rambaut A, Drummond AJ (2018) FigTree version 1.4.4 University of Edinburgh, Institute of Evolutionary Biology http://tree.bio.ed.ac.uk/software/figtree/
45. Mavrodiev EV, Madorsky A (2012) TAXODIUM Version 1.0: A simple way to generate uniform and fractionally weighted three-item matrices from various kinds of biological data PLoS ONE 7: e48813. https://doi.org/10.1371/journal.pone.0048813
46. Creevey C (2004) Clann: Construction of supertrees and exploration of phylogenomic information from partially overlapping datasets (version 3.0.0), User Manual vol 1, 3.0 edn. The lab. of James McInerney, Manchester, Great Britan http://chriscreevey.github.io/clann/
47. Creevey CJ, McInerney JO (2009) Trees from trees: Construction of phylogenetic supertrees using Clann. In: Posada D (ed) Bioinformatics for DNA Sequence Analysis. Springer Humana Press, New York, pp 139–161
48. Zhou X, Shen XX, Hittinger CT, Rokas A (2018) Evaluating fast Maximum Likelihood-based phylogenetic programs using empirical phylogenomic data sets Mol. Biol Evol 35:486–503. https://doi.org/10.1093/molbev/msx302
49. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for Maximum Likelihood analysis Nucleic Acids Res. 44:W232-W235. https://doi.org/10.1093/nar/gkw256
50. Ferron F, Debat HJ, Shannon A, Decroly E, Canard B (2020) A N7-guanine RNA cap methyltransferase signature-sequence as a genetic marker of large genome, non-mammalian *Tobaniviridae* Nucleic Acids Res. Genomics and Bioinformatics 2:1–10. https://doi.org/10.1093/nargab/lqz022
51. Liu SL, Saif LJ, Weiss SR, Su L (2020) No credible evidence supporting claims of the laboratory engineering of SARS-CoV-2. Emerg Microbes Infect 9:505–507. https://doi.org/10.1080/22221751.2020.1733440
52. Boni MF et al (2020) Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. 5:1408–1417 (2020). https://doi.org/10.1038/s41564-020-0771-4

53. Li X et al (2020) Emergence of SARS-CoV-2 through recombination and strong purifying selection Sci. Adv 6:eabb9153. https://doi.org/10.1126/sciadv.abb9153
54. Chaw SM et al (2020) The origin and underlying driving forces of the SARS-CoV-2 outbreak J. Biomed Sci 27:1–12. https://doi.org/10.1186/s12929-020-00665-8
55. Conceicao C et al (2020) The SARS-CoV-2 Spike protein has a broad tropism for mammalian ACE2 proteins. PLoS Biol 18(12):e3001016. https://doi.org/10.1371/journal.pbio.3001016
56. Liu DX, Liang JQ, Fung TS (2020) Human Coronavirus-229E,-OC43,-NL63, and-HKU1 Ref. Mod Life Sci. https://doi.org/10.1016/B978-0-12-809633-8.21501-X
57. Enjuanes L, Zuñiga S, Castaño-Rodriguez C, Gutierrez-Alvarez J, Canton J, Sola I (2016) Molecular basis of coronavirus virulence and vaccine development. In: Ziebuhr J (ed) Coronaviruses, Advances in Virus Research, 96, 8, Academic Press. ScienceDirect, Elsevier Inc., New York, pp 245–286
58. Fehr AR, Perlman S (2015) Coronaviruses: An overview of their replication and pathogenesis. In: Maier HJ, Bickerton E, Britton P (eds) Coronaviruses: Methods and protocols. Methods in Molecular Biology, Springer Humana Press, New York, pp 1–26
59. Gaunt ER 1, Hardie A, Claas ECJ, Simmonds P, Templeton KE (2010) Epidemiology and clinical presentations of the four human coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time. PCR method 48:2940–2947. https://doi.org/10.1128/JCM.00636-10
60. van der Hoek L (2007) Human coronaviruses: what do they cause? Antivir Ther 12:651–658 PMID: 17944272
61. Woo PC, Lau SK, Chu CM, Chan KH, Tsoi HW, Huang Y et al (2005) Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol 79:884–895. https://doi.org/10.1128/JVI.79.2.884-895.2005
62. Abdel-Moneim AS, Abdelwhab EM (2020) Evidence for SARS-CoV-2 infection of animal hosts Pathogens 9:529. https://doi.org/10.3390/pathogens9070529
63. Kiros M et al (2020) COVID-19 pandemic: current knowledge about the role of pets and other animals in disease transmission. Virol J 17:1–8. https://doi.org/10.1186/s12985-020-01416-9
64. Oreshkova N et al (2020) SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020 Euro Surveille 2020 25:pii = 2001005. https://doi.org/10.2807/1560-7917.ES.2020.25.23.2001005
65. Rossi GA, Sacco O, Mancino E, Cristiani L, Midulla F (2020) Differences and similarities between SARS-CoV and SARS-CoV-2: spike receptor-binding domain recognition and host cell infection with support of cellular serine proteases. Infection 48:665–669. https://doi.org/10.1007/s15010-020-01486-5

Figures

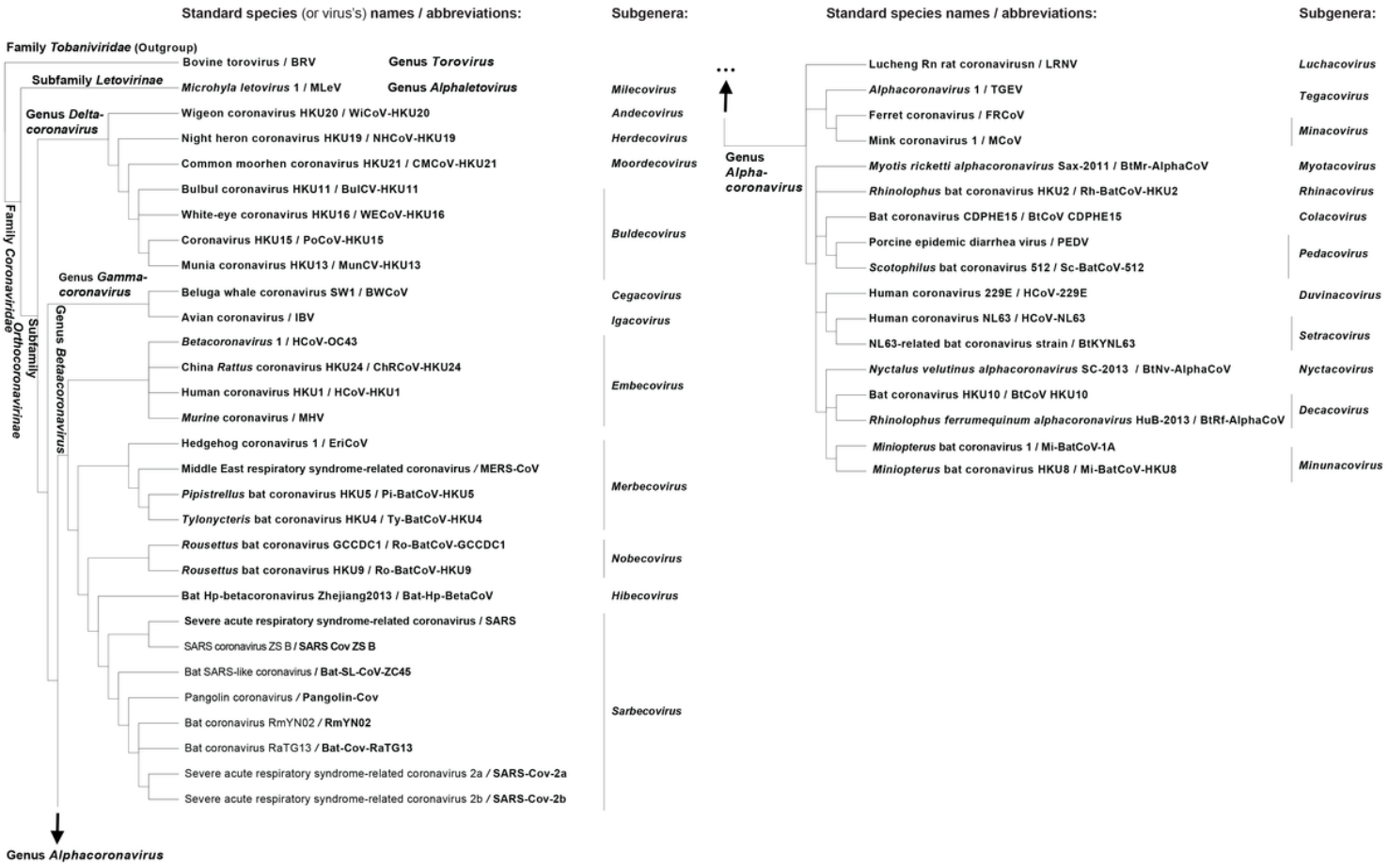

Figure 1

Strict Consensus of four trees produced by three different methods of cladistic analysis as well as by Maximum Likelihood method using modified genomic alignment of Coronaviridae + Torovirus drawn following standard species (or virus's) names and abbreviations. The standard virus names (not in bold) have been used only in case of the viruses from subgenus Sarbecovirus. See Figs S1–S4 for more detail including the tree node support values.

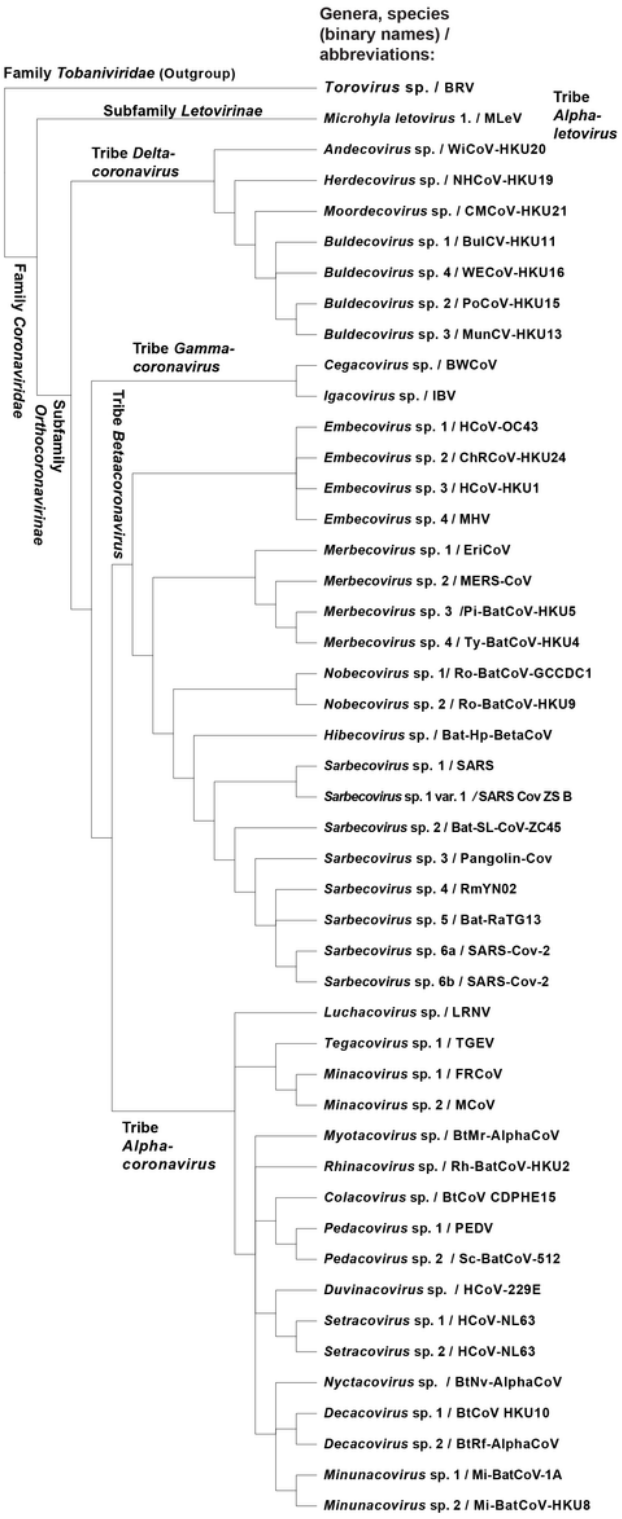

Figure 2

The same as Fig. 1, but drawn following suggested changes in the nomenclature of family Coronaviridae

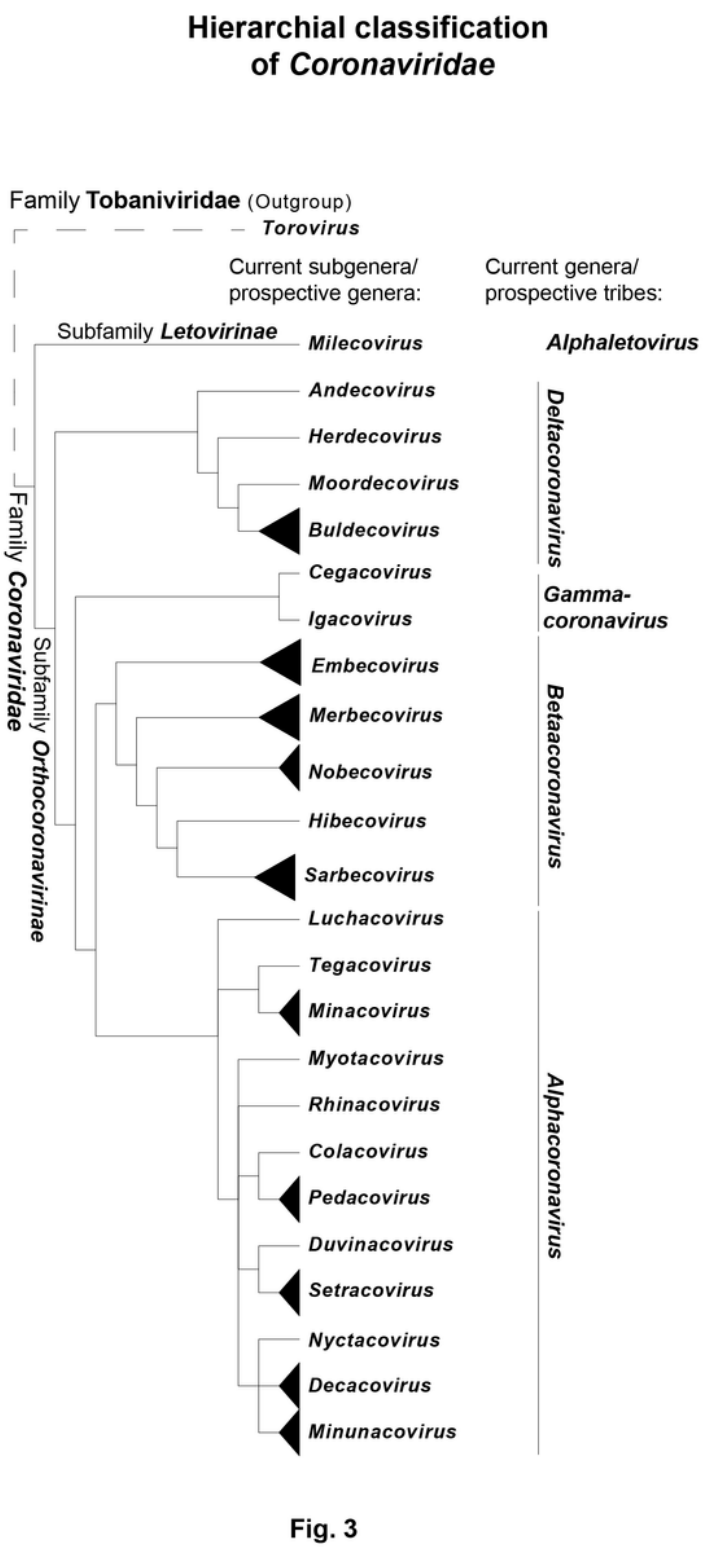

Fig. 3

Figure 3

Hierarchical classification of the coronaviruses (Riboviria, Nidovirales, Coronaviridae), established as a simplified Strict Consensus of four trees, produced by three different methods of cladistic analysis as well as by Maximum Likelihood method using G-block version of the genomic alignment of Coronaviridae + Torovirus. See Figs 1–2 and S1–S4 for more detail including the tree node support values. SARS-CoV-2

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fig.S1.pdf
- Fig.S2.pdf
- Fig.S3.pdf
- Fig.S4.pdf
- SupplementalTableS1R1.xlsx
- SupplementalTableS2R1.xlsx
- SupplementalTableS3R1.xlsx