Arch Virol (2003) 148: 2207–2235
DOI 10.1007/s00705-003-0162-1

Archives of Virology
Printed in Austria

A comparative sequence analysis to revise the current taxonomy of the family *Coronaviridae*

J. M. González[1], **P. Gomez-Puertas**[2], **D. Cavanagh**[3], **A. E. Gorbalenya**[4], and **Luis Enjuanes**[1]

[1]Centro Nacional de Biotecnología, CSIC, Department of Molecular and Cell Biology, Campus Universidad Autónoma, Cantoblanco, Madrid, Spain
[2]Bioinformatics Lab. Centro de Astrobiología (CSIC-INTA), Torrejón de Ardoz, Madrid, Spain
[3]Institute for Animal Health, Compton Laboratory, Compton, Newbury, U.K.
[4]Center of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands

Received April 2, 2003; accepted May 20, 2003
Published online August 18, 2003

Summary. The *Coronaviridae* family, comprising the *Coronavirus* and *Torovirus* genera, is part of the *Nidovirales* order that also includes two other families, *Arteriviridae* and *Roniviridae*. Based on genetic and serological relationships, groups 1, 2 and 3 were previously recognized in the *Coronavirus* genus. In this report we present results of comparative sequence analysis of the spike (S), envelope (E), membrane (M), and nucleoprotein (N) structural proteins, and the two most conserved replicase domains, putative RNA-dependent RNA polymerase (RdRp) and RNA helicase (HEL), aimed at a revision of the *Coronaviridae* taxonomy. The results of pairwise comparisons involving structural and replicase proteins of the *Coronavirus* genus were consistent and produced percentages of sequence identities that were distributed in discontinuous clusters. Inter-group pairwise scores formed a single cluster in the lowest percentile. No homologs of the N and E proteins have been found outside coronaviruses, and the only (very) distant homologs of S and M proteins were identified in toroviruses. Intragroup sequence conservation was higher, although for some pairs, especially those from the most diverse group 1, scores were close or even overlapped with those from the intergroup comparisons. Phylogenetic analysis of six proteins using a neighbor-joining algorithm confirmed three coronavirus groups. Comparative sequence analysis of RdRp and HEL domains were extended to include arterivirus and ronivirus homologs. The pairwise scores between sequences of the genera

Coronavirus and *Torovirus* (22–25% and 21–25%) were found to be very close to or overlapped with the value ranges (12 to 22% and 17 to 25%) obtained for interfamily pairwise comparisons, but were much smaller than values derived from pairwise comparisons within the *Coronavirus* genus (63–71% and 59–67%). Phylogenetic analysis confirmed toroviruses and coronaviruses to be separated by a large distance that is comparable to those between established nidovirus families. Based on comparison of these scores with those derived from analysis of separate ranks of several multi-genera virus families, like the *Picornaviridae*, a revision of the *Coronaviridae* taxonomy is proposed. We suggest the *Coronavirus* and *Torovirus* genera to be re-defined as two subfamilies within the *Coronavirdae* or two families within *Nidovirales*, and the current three informal coronavirus groups to be converted into three genera within the *Coronaviridae*.

Introduction

The current virus taxonomy universally uses the order, family, genus and species ranks to organize all diversity of viruses within a hierarchical system [48, 79]. To better reflect an outstanding complexity of similarities found in some virus groups, a subfamily rank is also occasionally used. Viruses are assigned to a particular taxonomic position according to results of comparative analysis of selected properties, characterizing different aspects of the genome and virion structures and the replication strategy of viruses. There is no hierarchy in the property list and most of the features used are not quantitative. Nevertheless, analysis of genomic data has *de facto* played an increasing role in the past taxonomy revisions.

The resolving power of comparative sequence analysis was clearly demonstrated in a study of the virus capsid gene sequences of the *Potyviridae* family, when diverse strains, species and genera were separated in distinct clusters according to pairwise sequence scores [83]. In another highly illustrative case, the results of comparative sequence analysis of replicative proteins [4, 41] were most vital for a decision to expel *Hepatitis E virus* from the *Caliciviridae* family, where it was originally placed using other non-sequence properties [5]. Results of comparative sequence analysis were also instrumental for the creation of the *Arteriviridae* family [17] and subsequent placement of this and *Coronaviridae* families into a newly designed *Nidovirales* order currently including also *Roniviridae*, all of which are morphologically different [13, 22, 45]. The experience proved that conserved sequence patterns common for this order are more reliable characteristics than other properties, including the spliced structural organization of subgenomic RNAs that was originally considered a hallmark of the *Nidovirales* [16, 26, 34, 46, 53, 59, 66, 71, 74, 84].

This study focuses on coronavirus taxonomy. These viruses use single-stranded positive-sense RNA genomes of between 28 and 32 kb that are packaged in enveloped virions with corona- or toro-like morphology [21]. The coronavirus genome includes multiple open reading frames (ORFs), with a large replicase being encoded in the two 5′-most and overlapping ORFs and the structural and auxiliary proteins being expressed from the downstream four or more ORFs. The replicase components are autoproteolytically derived from two polyproteins, one

of which is produced through a frameshifting during virion RNA translation [7, 90]. The backbone of the replicase polyproteins includes several uniquely arranged conserved domains, two of which have not been found outside the *Nidovirales* order [16, 26, 71]. The non-replicase ORFs are expressed from a 5′- and 3′-coterminal nested set of subgenomic viral mRNAs [22, 44]. The *Coronaviridae* family is formed by the genera *Coronavirus* and *Torovirus* [21].

Using genetic and antigenic criteria, virus species in the genus *Coronavirus* have been organized into groups 1, 2 and 3 [21]. Group 1 includes porcine *Transmissible gastroenteritis virus* (TGEV), *Feline coronavirus* (FCoV), *Canine coronavirus* (CCoV), *Human coronavirus 229E* (HCoV-229E) and *Porcine epidemic diarrhea virus* (PEDV). Group 2 members are *Murine hepatitis virus* (MHV), *Bovine coronavirus* (BCoV), *Human coronavirus OC43* (HCoV-OC43), *Porcine hemagglutinating encephalomyelitis virus* (HEV), *Rat coronavirus* (RtCoV), and Equine coronavirus (ECoV). Group 3 is formed by avian *Infectious bronchitis virus* (IBV), *Turkey coronavirus* (TCoV), and Pheasant coronavirus [10]. The current distribution of species into groups 1 to 3 agrees with previously performed phylogenetic analyses [11, 31, 67, 76], although the status of groups within a genus is rather provisional and does not correspond to a proper taxonomic category.

Toroviruses were originally proposed to form a new family separated from coronaviruses [35]. However, subsequent comparative data analyses led to its recognition as a genus within the *Coronaviridae* [9, 57]. Two torovirus species, *Bovine torovirus* (BToV), originally named Breda virus, and Equine torovirus (EToV), have been recognized although toroviruses may also infect other mammals, including human and swine [20, 42, 55, 71, 81]. The EToV is so far the only torovirus that has been propagated in tissue culture and molecularly characterized [71, 81], although partial genome sequences have also been determined for toroviruses infecting other species [20, 42].

Due to rapidly accumulating data on the genome structure, expression, and virus architecture of coronaviruses and other nidoviruses, it seems appropriate to bring up-to-date the taxonomic classification of the *Coronaviridae* family. In this study we performed a systematical quantitative analysis of sequence conservation among four structural proteins of the *Coronaviridae* and two key replicase enzymes, putative RNA-dependent RNA polymerase (RdRp) and helicase (HEL) of the *Nidovirales*. The results were correlated with non-sequence characteristics and rationalized using criteria that were derived from analysis of other virus families. Our analysis suggests that the *Coronavirus* and *Torovirus* genera should be re-defined as two subfamilies within the *Coronaviridae* or two families within *Nidovirales*, and the current informal three coronavirus groups to be converted into three genera within the *Coronaviridae*.

Materials and methods

Comparative sequence analyses

Databases searches were done using the BLAST program [1] available through the WU-BLAST2 server [87]. Amino acid sequences were obtained from the SWISS-PROT/TrEMBL

[50] and PIR [86] databases. For the structural proteins, only full-length sequences were included in the analysis. For the replicase domains, the sequences including the conserved motifs of the RdRp [40] and HEL [27, 38] that corresponded to fragments 513-820 and 1218-1512, respectively, of MHV ORF 1b (accession number P16342) were used. In total, 73 S, 44 E, 57 M, 66 N, 19 RdRp and 14 HEL sequences were analyzed and they are listed in respective figures. Note that for each protein a unique set of sequences was analyzed and the protein-specific sets overlapped to different extent.

Sequences were aligned with the CLUSTAL X program v. 1.82 [77] and the alignments were curated with T-COFFEE v. 1.32, that combines local and global multiple alignments and yields more accurate sequence alignments than other available methods [49]. Some alignments were verified using the MACAW program [62] and were manually adjusted.

The statistical significance of the similarity between the sequences included in the multiple alignments was verified applying the PSI-BLAST [2], LAMA [56], and MACAW [62] programs.

The PSI-BLAST program mediates iterative searches that start with a query and involve building a position-specific scoring matrix from sequences similar to the query to be used as input to the next round of searching. The search continues and an alignment expands with new sequences until the results convergence when no new hits above a statistically significant threshold are recorded. In this study, every sequence to be compared was used as a query in iterative PSI-BLAST searches against the non-redundant (nr) peptide sequence database with an inclusion E threshold being 0.05. This value indicates that the threshold similarity may be observed by chance once per any sequence search of a database 20 times as big as that that was actually searched. We considered similarities among all sequences in a group to be statistically significant if outputs of searches that were initiated with every group sequence formed a continuous network of matches.

The most conserved regions in sequence alignments are known to form ungapped blocks (ungapped local multiple alignments). Such blocks can be derived from multiple alignments employing the Block Maker [33] and used as a query in searches mediated by the LAMA. Both programs and other tools are run through the Blocks web server (http://www.blocks.fhcrc.org). The LAMA program searches for statistically significant similarities between blocks of an alignment and a blocks database derived from another alignment (a protein family) or from all documented families of related proteins forming the Blocks database [32]. A hit is considered relevant if its Z-score, the number of standard deviations between the blocks alignment score and a mean score previously calculated for the entire Blocks database, was above the score cut-off of 5.6. In this study, the Block Maker was used to convert multiple alignments containing groups of coronavirus or torovirus sequences into the alignment blocks databases. Then, LAMA performed inter-databases comparisons in a block-versus-block mode and also used blocks of each alignment as a query to search the complete Blocks database.

To evaluate similarity between distantly related toro- and coronavirus sequences, the MACAW program was used. The MACAW program identifies conserved ungapped blocks in a group of sequences, assesses statistical significance of intra-block similarity and combines blocks in a multiple sequence alignment containing inter-block unaligned regions. To avoid distortion of the statistical calculations, closely related sequences must be excluded from the analysis. In this study, we used MACAW to align representatives of three coronavirus groups and a torovirus sequence. If the intra-block similarity of these sequences was statistically significant (probability of finding the same or higher score by chance was not more than 0.01) and this probability became less likely after removal of any sequence from this alignment, then the intra-block relationship of *all* aligned sequences was considered to be statistically significant.

Distance and phylogenetic analyses

The obtained alignments were used as input for the distance and phylogenetic analyses. Uncorrected distances for every pairwise sequence comparison (percentage of sequence identity) were calculated with DISTANCES from the GCG package (Womble, 2000). The calculated distances were further grouped in the 2% intervals and the obtained figures were plotted on the frequency versus identity percentage histograms using Microsoft Excel 2001.

Dendrograms were computed by successively using four programs included in the PHYLIP package v.3.6a3 [25]. SEQBOOT generates resampled versions of an input data set, and it was used to create 1000 bootstrapped data sets from each alignment. Distance matrices summarizing pairwise comparisons within each one of the multiple alignment data sets were obtained with PROTDIST according to the Jones-Taylor-Thornton model of amino acid substitutions [37]. The distance matrices were fed to NEIGHBOR to compute the dendrograms by applying the Neighbor-joining method that constructs a tree by successive clustering of lineages [58]. Finally, from the multiple trees obtained for each original alignment, the majority rule consensus tree showing the bootstrap values in the nodes was calculated by CONSENSE.

Alternatively, consensus unrooted Neighbor-joining dendrograms were obtained with CLUSTAL X v1.82 starting from 1000 bootstrapped replicates of each alignment. The phylogenetic trees obtained with PHYLIP and CLUSTAL X had similar topologies. The CLUSTAL X dendrograms are shown in this article.

For dendrograms containing S and M proteins, RdRp and HEL sequences, roots were inferred with the corresponding torovirus homologous sequences as outgroups. For this purpose, the statistical significance of the relationships between coronavirus and torovirus structural protein sequences was assessed (see below).

The phylogenetic trees were plotted with the NJplot program [54] and the TreeView program v. 1.6.6 (Page, 1996) and manually edited.

Results

Generation of coronavirus-wide alignments of four structural proteins and two replicative domains

To perform a comprehensive comparative sequence study of coronaviruses, the two most conserved domains, putative RdRp [29] and HEL [28, 63, 64], that are part of the replicase polyproteins, and the four structural proteins common to all coronaviruses (N, M, E, and S) have been selected.

The Psi-Blast-mediated searches retrieved all coronavirus N, M and S proteins as separate groups that were subsequently aligned as described in the Material and Methods. Similar searches that were performed with E proteins produced four different families, two for coronavirus group 1 and one for each groups 2 and 3; these families are also listed in the protein family (PFAM) database [3]. To check whether these protein families are related, we performed LAMA-assisted across-families comparisons using 2 or 3 ungapped blocks that were derived from alignments of these protein families with the Block Maker tool. A four-families-wide network of statistically significant interblock matches was detected in pairs of different protein families excluding only the families 2 and 3 pair. These data and similar genetic positions support the common origin of the different E proteins. Accordingly, four group-specific E protein alignments were merged into one coronavirus-wide alignment using the Clustalx1.82 and T-Coffee programs.

The PSI-BLAST- and LAMA-mediated searches did not bring statistically significant matches between the structural proteins of the two genera of the *Coronaviridae*. Toroviruses have three structural proteins functionally equivalent to the S, M and N proteins of coronaviruses. The S and M proteins are of similar sizes, while the N protein is about 75% smaller in toroviruses. Based on these biological grounds, we compared S and M proteins of corona- and toroviruses using the MACAW program. Three statistically significant and colinear regions have been found in the C-terminal half of S proteins and two such regions were delineated in M proteins. The C-terminal half is the most conserved part of S protein and released as S2 moiety by a cleavage of S protein. The identified conserved regions enabled the generation of the *Coronaviridae*-wide multiple alignments of S and M proteins using the Clustal x 1.82 and T-Coffee programs.

The *Coronaviridae*-wide alignments of the most conserved regions of the RdRp and HEL were produced to include the characteristic motifs of these proteins [28, 29].

Three genetic groups are consistently evident upon analysis of pairwise distances of six proteins of the Coronavirus *genus*

The all-inclusive alignments of six coronavirus proteins were used to produce respective matrices with percentages of the pairwise sequence identity. These matrices were further processed to derive and individually plot results for three coronavirus genetic groups and four inter-group combinations for each protein. Inspection of the 42 histograms obtained showed that the calculated identity percentages are not distributed continuously, but rather group in discrete clusters. Analysis of these distributions is given below.

Overall results for the four structural proteins were similar (Fig. 1). Frequency distributions of identity percentages that were derived from intragroup 1 comparisons formed two main clusters. The rightmost one, which was discontinuous and included identity percentages from 78 to 100 (S), 75 to 100 (E), 82 to 100 (M) and 74 to 100 (N) (G1 in Figs. 1A, B, C and D), included distances between strains from the same or closely related species. The leftmost cluster was compact and showed lower identity percentages ranging from 42 to 52 (S), 23 to 31 (E), 42 to 57 (M), and 34 to 41 (N). These figures were generated from pairwise comparisons between viruses of the two group 1 subsets, one including TGEV, CCoV, and FCoV (G1-1), and the other consisting of HCoV-229E and PEDV (G1-2). Viruses that belong to the two different subsets may lack the antigenic cross-reactivity [59].

The intragroup 2 comparisons also showed identity percentages that formed two clusters (G2 in Figs. 1A, B, C and D). The rightmost one included protein identity percentages from 81 to 100 (S), 89 to 100 (E), 92 to 100 (M) and 89 to 100 (N) that were generated from comparisons between the most closely related sequences. The other cluster included percentage scores that ranged from 65 to 69 (S), 61 to 70 (E), 79 to 85 (M) and 69 to 76 (N). It corresponded to comparisons between species of two subgroups, one that includes murine coronaviruses (MHV and RtCoV), and the other including HCoV-OC43, BCoV

and HEV. It is evident that the intragroup 2 pairwise sequence differences are significantly less pronounced than those found for the group 1.

Pairwise identity percentages within group 3 that is formed by two closely related species IBV and TCoV were accordingly high and clustered compactly for three proteins, S (from 82 to 100%), E (83–100%) and M (80–100%) (G3 in Figs. 1A, B, and C). However, comparisons of protein N sequences revealed two clearly separated distance clusters, a rightmost including high percentage scores (from 88 to 100%) comparable to that of other proteins, and a leftmost with the

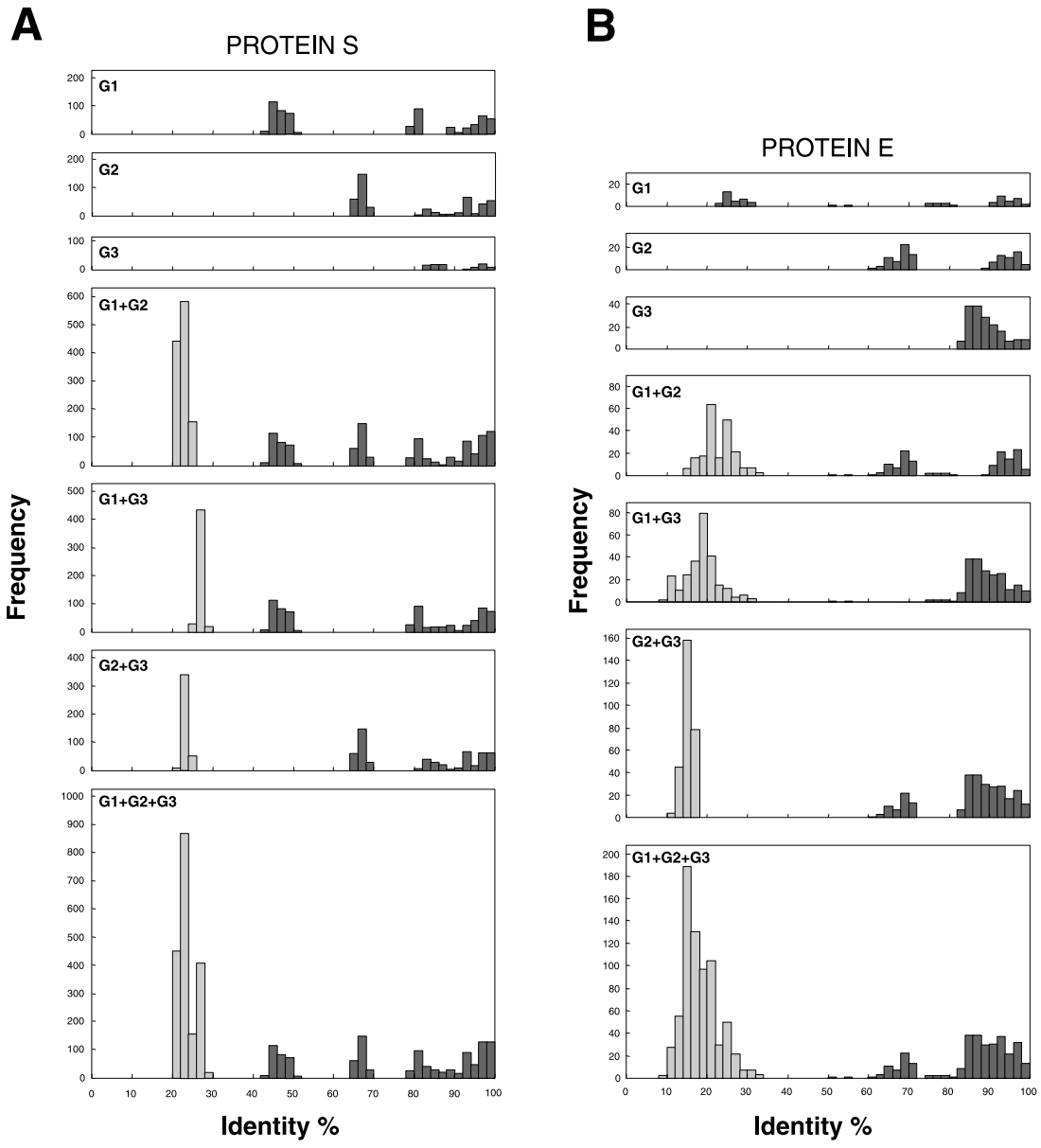


Fig. 1 (*continued*)

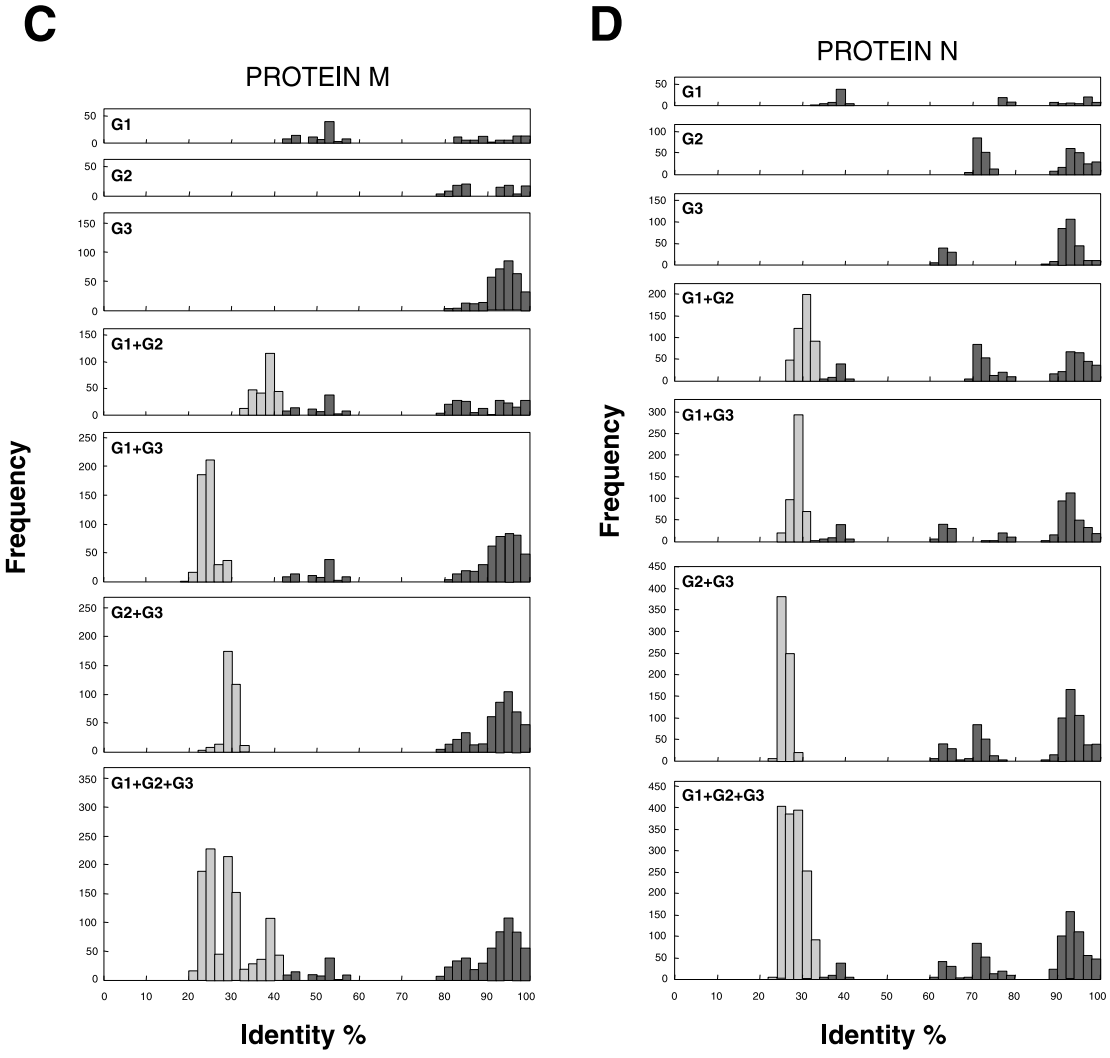


Fig. 1. Frequency distributions of pairwise identity percentages of coronavirus structural proteins. Amino acid sequences of proteins S (**A**), E (**B**), M (**C**) and N (**D**) were aligned with CLUSTAL X and T-COFFEE to generate pairwise score matrices including percentages of identical residues in each pair of sequences. Each protein matrix was produced from comparisons involving coronaviruses of three genetic groups (G1 + G2 + G3), and was used to derive submatrices involving sequences of group 1 (G1), group 2 (G2), group 3 (G3), groups 1 and 2 (G1 + G2), groups 1 and 3 (G1 + G3), and groups 2 and 3 (G2 + G3). These matrices of each protein were processed to plot frequency distributions of percentage scores that were rounded with the step of 2%. Histograms of the intra-group scores were colored in dark gray and those of the inter-group scores in light gray

pairwise identities in a range from 60 to 65% (G3 in Fig. 1D). This second unique cluster originated from comparisons involving two species subsets, one prototyped by IBV Beaudette strain and other involving three IBV strains N1/88, Q3/88 and V18/91 [60]. The observed differences in the patterns of score distribution among

four proteins may be rationalized after the genomic characterization of the above three IBV strains is extended beyond the N protein gene.

Comparisons involving sequences from two different coronavirus groups produced pairwise identity percentages distribution that combined distributions of two groups and included a new cluster containing intergroup pairwise scores (G1 + G2, G1 + G3 and G2 + G3 in Figs. 1A, B, C and D). For three different combinations of two groups, intergroup scores formed the leftmost clusters that were clearly separated from all intragroup clusters except for the leftmost G1

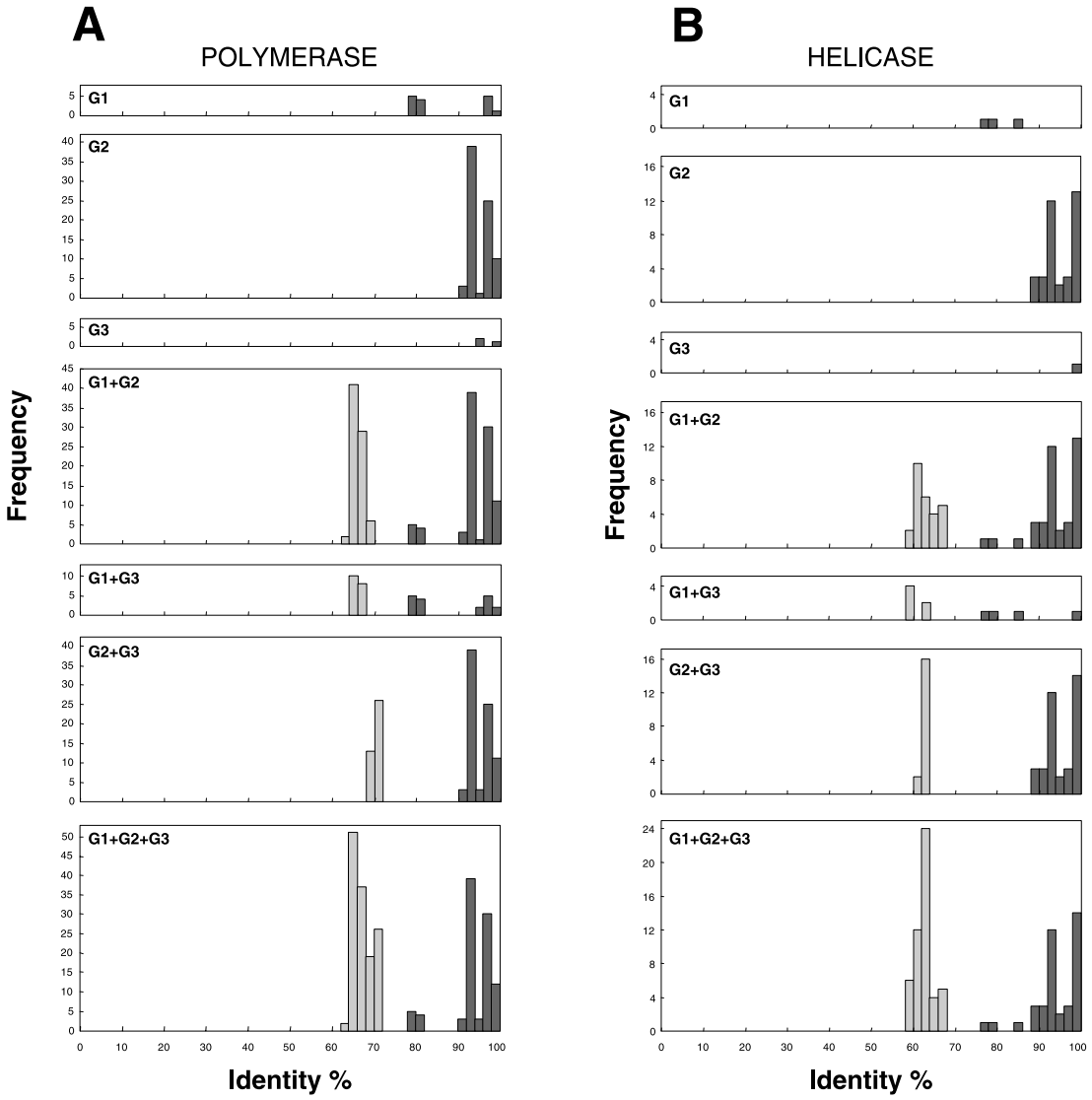
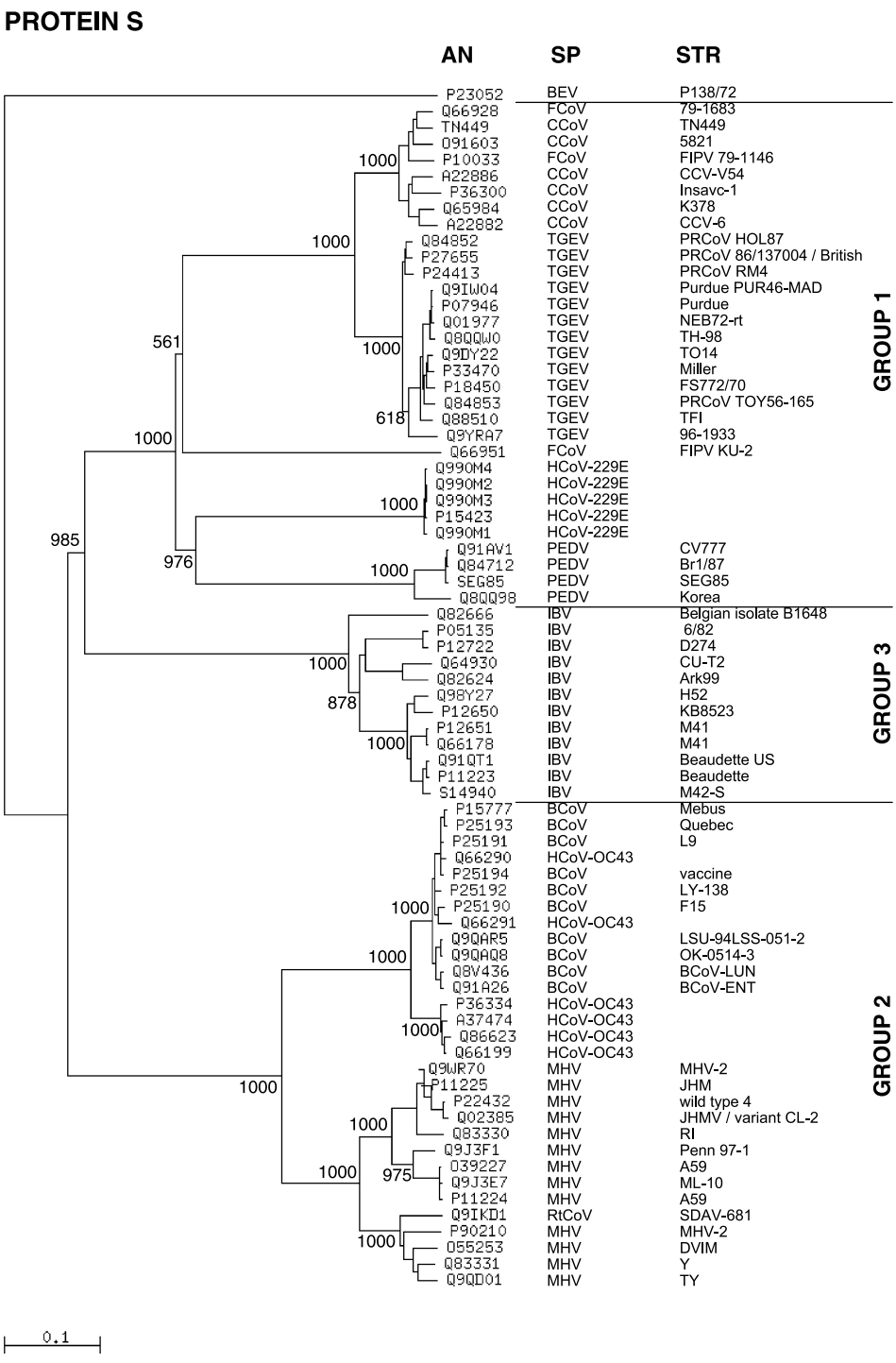


Fig. 2. Frequency distributions of pairwise identity percentages of coronavirus replicase proteins. Amino acid sequences of the RNA-dependent RNA polymerase (**A**) and the RNA helicase (**B**) domains were analyzed in a manner described in legend to Fig. 1 for structural proteins of coronaviruses


PROTEIN S
AN
SP
STR
P23052 BEV P138/72
Q66928 FCoV 79-1683
TN449 CCoV TN449
Q91603 CCoV 5821
P10033 FCoV FIPV 79-1146
A22886 CCoV CCV-V54
P36300 CCoV Insavc-1
Q65984 CCoV K378
A22882 CCoV CCV-6
Q84852 TGEV PRCoV HOL87
P27655 TGEV PRCoV 86/137004 / British
P24413 TGEV PRCoV RM4
Q9IW04 TGEV Purdue PUR46-MAD
P07946 TGEV Purdue
Q01977 TGEV NEB72-rt
Q8QQW0 TGEV TH-98
Q9DY22 TGEV TO14
P33470 TGEV Miller
P18450 TGEV FS772/70
Q84853 TGEV PRCoV TOY56-165
Q88510 TGEV TFI
Q9YRA7 TGEV 96-1933
Q66951 FCoV FIPV KU-2
Q990M4 HCoV-229E
Q990M2 HCoV-229E
Q990M3 HCoV-229E
P15423 HCoV-229E
Q990M1 HCoV-229E
Q91AV1 PEDV CV777
Q84712 PEDV Br1/87
SEG85 PEDV SEG85
Q8QQ98 PEDV Korea
Q82666 IBV Belgian isolate B1648
P05135 IBV 6/82
P12722 IBV D274
Q64930 IBV CU-T2
Q82624 IBV Ark99
Q98Y27 IBV H52
P12650 IBV KB8523
P12651 IBV M41
Q66178 IBV M41
Q91QT1 IBV Beaudette US
P11223 IBV Beaudette
S14940 IBV M42-S
P15777 BCoV Mebus
P25193 BCoV Quebec
P25191 BCoV L9
Q66290 HCoV-OC43
P25194 BCoV vaccine
P25192 BCoV LY-138
P25190 BCoV F15
Q66291 HCoV-OC43
Q9QAR5 BCoV LSU-94LSS-051-2
Q9QAQ8 BCoV OK-0514-3
Q8V436 BCoV BCoV-LUN
Q91A26 BCoV BCoV-ENT
P36334 HCoV-OC43
A37474 HCoV-OC43
Q86623 HCoV-OC43
Q66199 HCoV-OC43
Q9WR70 MHV MHV-2
P11225 MHV JHM
P22432 MHV wild type 4
Q02385 MHV JHMV / variant CL-2
Q83330 MHV RI
Q9J3F1 MHV Penn 97-1
O39227 MHV A59
Q9J3E7 MHV ML-10
P11224 MHV A59
Q9IKD1 RtCoV SDAV-681
P90210 MHV MHV-2
O55253 MHV DVIM
Q83331 MHV Y
Q9QD01 MHV TY
1000
561
1000
618
1000
985
976
1000
1000
878
1000
1000
1000
1000
1000
1000
975
1000
GROUP 1
GROUP 3
GROUP 2
0.1

clusters of E, M and N proteins, with which they were either overlapped (E and N) or formed a continuous supercluster (M).

Finally, four protein-specific histograms combining all available pairwise scores were generated (G1 + G2 + G3 in Figs. 1A, B, C and D). Inspection of these histograms revealed overlapping ranges of sequence intergroup identities for each protein: from 20 to 29% (S), 9 to 30% (E), 21 to 41% (M) and 23 to 35% (N).

Similar pair-wise comparisons involving two replicase protein domains (Fig. 2) yielded percentage identities higher than those of the structural proteins. This difference was expected given that these domains are conserved also outside coronaviruses [27, 40]. Thus, pairwise identity percentages among RdRp (Fig. 2A) and HEL (Fig. 2B) sequences from different groups ranged from 63 to 71% (G1 + G2 + G3 in Fig. 2A), and from 59 to 67% (G1 + G2 + G3 in Fig. 2B), respectively. These values border lined the respective intragroup figures that ranged from 78 to 100% (G1, G2 and G3 in Fig. 2A) and from 77 to 100% (G1, G2 and G3 in Fig. 2B), respectively. Because of the limited sampling of analyzed sequences, especially for the RNA helicase, the histograms were not as well defined as those for the structural proteins.

Thus, the results of pairwise comparisons involving structural and replicase proteins of the *Coronavirus* genus were consistent and produced percentages of sequence identities that were distributed in discontinuous clusters. Inter-group pairwise scores formed a single cluster with the lowest percentile. In contrast, intragroup pairwise scores were higher, although for some pairs, especially those from the most diverse group 1, scores were close or even overlapped with those from the intergroup comparisons.

Three genetic groups are supported by phylogenetic analysis of six proteins of the Coronavirus *genus*

The dendrograms involving six analyzed proteins revealed similar topologies (Fig. 3 to 8) that were compatible with the current distribution of coronavirus species among three groups [21, 67]. Each coronavirus group is supported by high bootstrap values for every protein analyzed. Group 3 is relatively compact, currently including data from just two species. Each of groups 1 and 2 includes two subsets of species. In the moderately diverged group 2, a subset including MHV and RtCoV species was confidently separated from the other including HCoV-OC43, BCoV and HEV species. In the most diverse and diverged group 1,

Fig. 3. Phylogenetic tree of S proteins of coronaviruses. The tree was generated using an alignment of the S2 part of S protein sequences from 74 different virus isolates by applying the Neighbor-joining method in the CLUSTAL X v1.82 program. The sequence of the Berne torovirus (BEV) S protein was used as an outgroup. Bootstrap values higher than 50% are shown in the main branch nodes. AN, SWISS-PROT/TrEMBL/PIR databases accession number; SP, abbreviation of the official species name; STR, strain name. In the AN column, TN449 and SEG85 correspond to sequences that have been obtained in the author's laboratory and have not been published

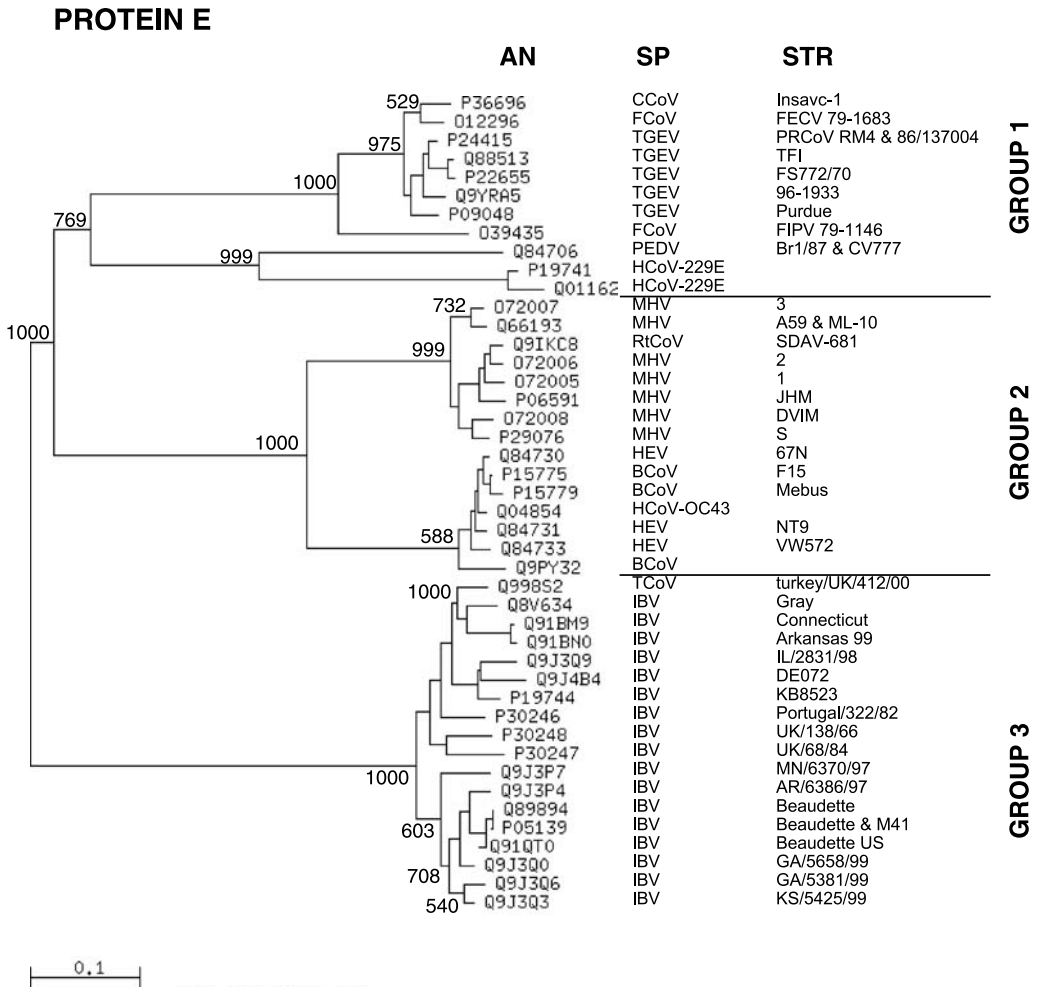


Fig. 4. Phylogenetic tree of E proteins of coronaviruses. The tree was generated using an alignment of E protein sequences from 44 virus isolates by applying the Neighbor-joining method in the CLUSTAL X v1.82 program. Bootstrap values higher than 50% are shown in the main branch nodes. AN, SP, and STR as in Fig. 1

TGEV, CCoV, FCoV, forming subgroup G1-1, were separated from substantially diverged HCoV-229E and PEDV, forming another subgroup G1-2. The topologies of group 1 sub-trees were found to be very similar for the M and N proteins but deviated for the other two proteins with respect to the positions of FCoV isolates. Pairwise similarity of the feline infectious peritonitis virus (FIPV) KU-2 strain S protein was unusually low (45% of identical residues) with homologs encoded by other group 1 coronaviruses [47] including other strains of FIPV. Furthermore, this virus failed a confidence test for the inclusion in the subgroup G1-1 (Figs. 1 and 3). The FIPV 79-1146 and the feline enteric coronavirus (FECoV) 79-1683 strains were interleaved with TGEV and other coronaviruses in the E protein tree (Fig. 4), but together with FIPV KU-2 formed the compact cluster in the trees

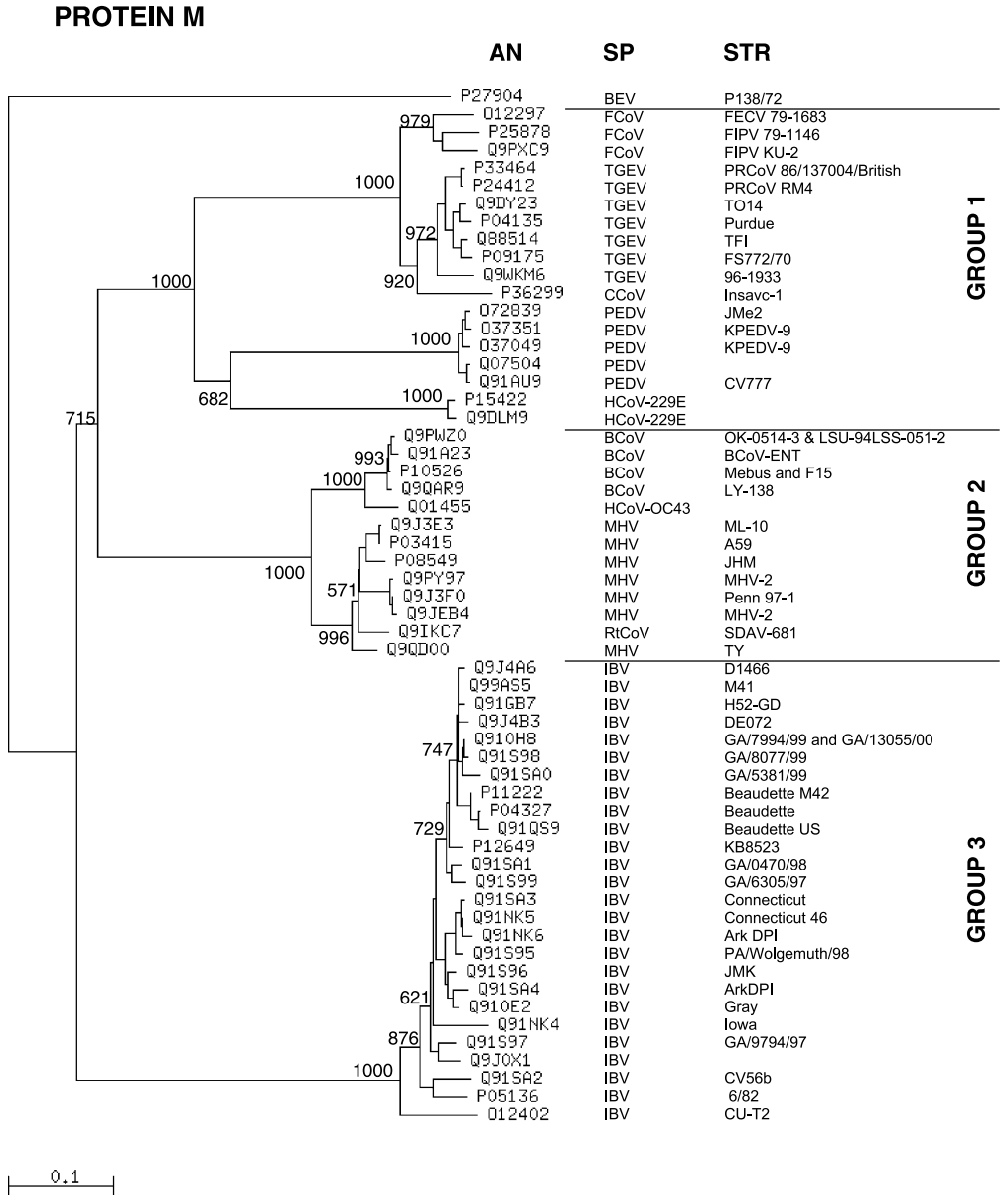


Fig. 5. Phylogenetic tree of M proteins of coronaviruses. The tree was generated using an alignment of M protein sequences from 58 virus isolates by applying the Neighbor-joining method in the CLUSTAL X v1.82 program. The sequence of the BEV M protein was used as an outgroup. Bootstrap values higher than 50% are shown in the main branch nodes. AN, SP, and STR, as in Fig. 1

of M and N proteins (Fig. 5 and 6). These topological anomalies indicate that recombination may have contributed to the evolution of group 1 viruses, and this aspect is worth further analysis that is beyond the scope of this paper. Since our findings are limited to the same group 1, they do not undermine the classification of coronaviruses in three main groups.

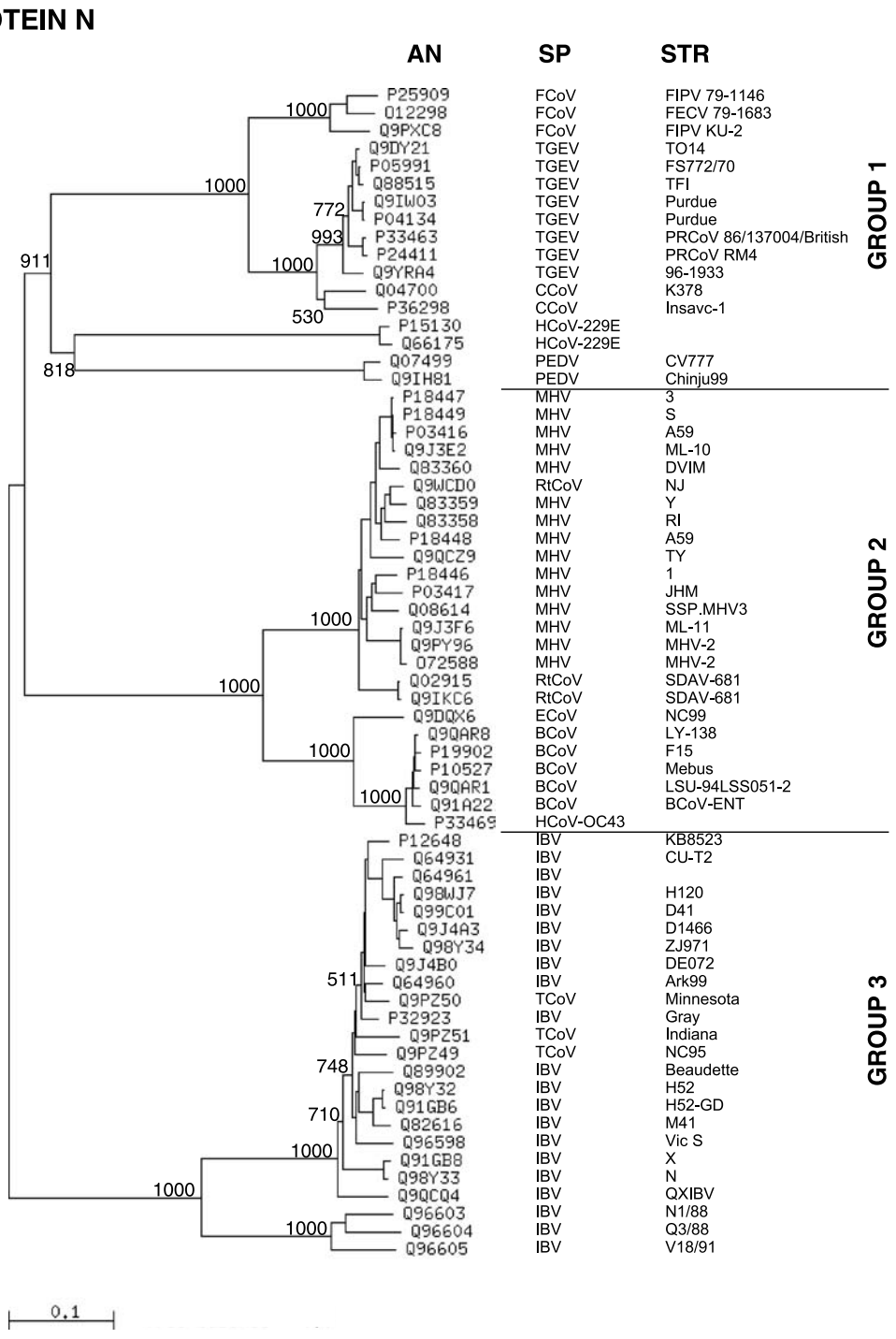


Fig. 6. Phylogenetic tree of coronavirus N protein amino acid sequences. The tree was generated using the sequences of N protein from 66 virus isolates by applying the Neighbor-joining method in the CLUSTAL X v1.82 program. Bootstrap values higher than 50% are shown in the main branch nodes. AN, SP, and STR, as in Fig. 1. EqCoV, putative equine coronavirus

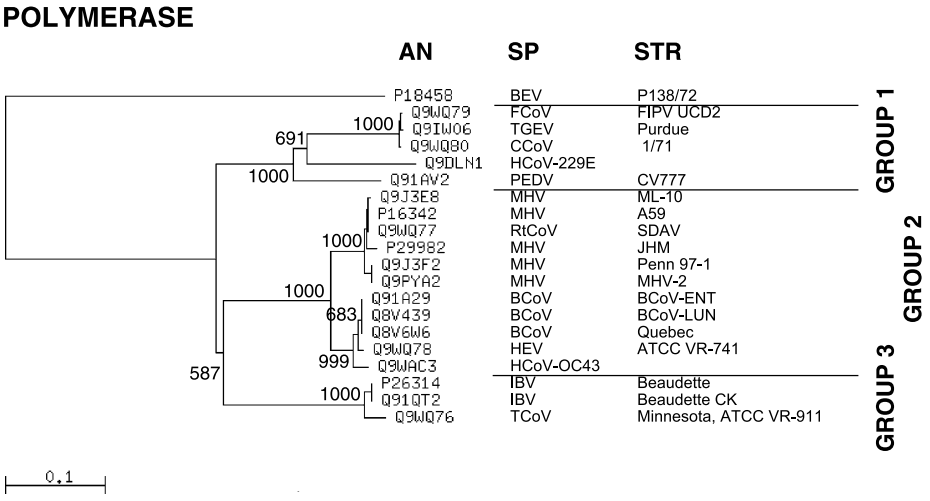


Fig. 7. Phylogenetic tree of putative RNA-dependent RNA polymerases of coronaviruses. The tree was generated using an alignment of the RdRp domain from 20 virus isolates by applying the Neighbor-joining method in the CLUSTAL X v1.82 program. The sequence of the BEV RdRp was used as an outgroup. Bootstrap values higher than 50% are shown in the main branch nodes. AN, SP, and STR, as in Fig. 1

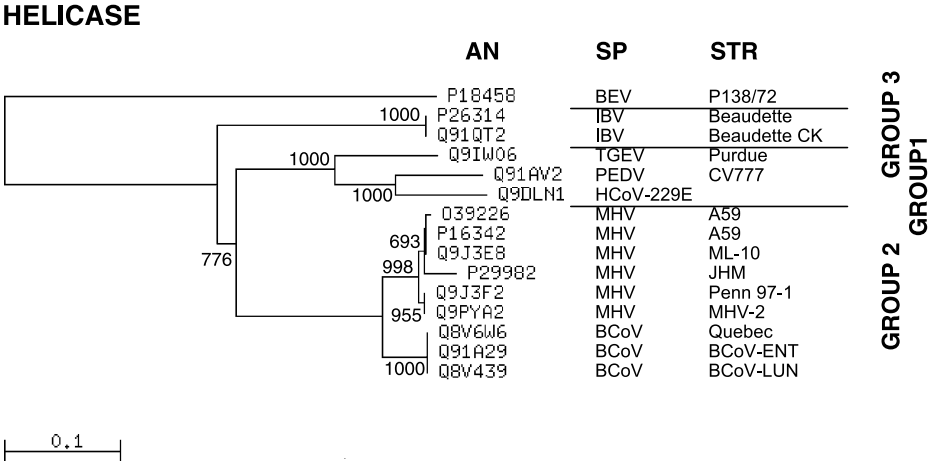


Fig. 8. Phylogenetic tree of RNA helicases of coronaviruses. The tree was generated using an alignment of the HEL domain from 15 different virus isolates by applying the Neighbor-joining method in the CLUSTAL X v1.82 program. The sequence of the BEV HEL was used as an outgroup. Bootstrap values higher than 50% are shown in the main branch nodes. AN, SP, and STR, as in Fig. 1

The trees for S (Fig. 3), M (Fig. 5), RdRp (Fig. 7) and HEL (Fig. 8) proteins were rooted using the torovirus homologs as outgroups. Several topologies are evident in these trees. The 1st and 2nd groups were clustered in the M and HEL trees, and the 1st and 3rd groups were clustered in the S tree, while no reliable intergroup clustering was observed in the RdRp tree. The possible causes of these

variations remain unknown although they may purely be due to technical reasons (e.g. small sampling size and/or large distances between the outgroup and other sequences). In this respect, the root position in the S protein tree may be the least reliable due to the most pronounce divergence of the respective torovirus sequence. However, these unresolved complexities do not compromise the group structure of coronaviruses.

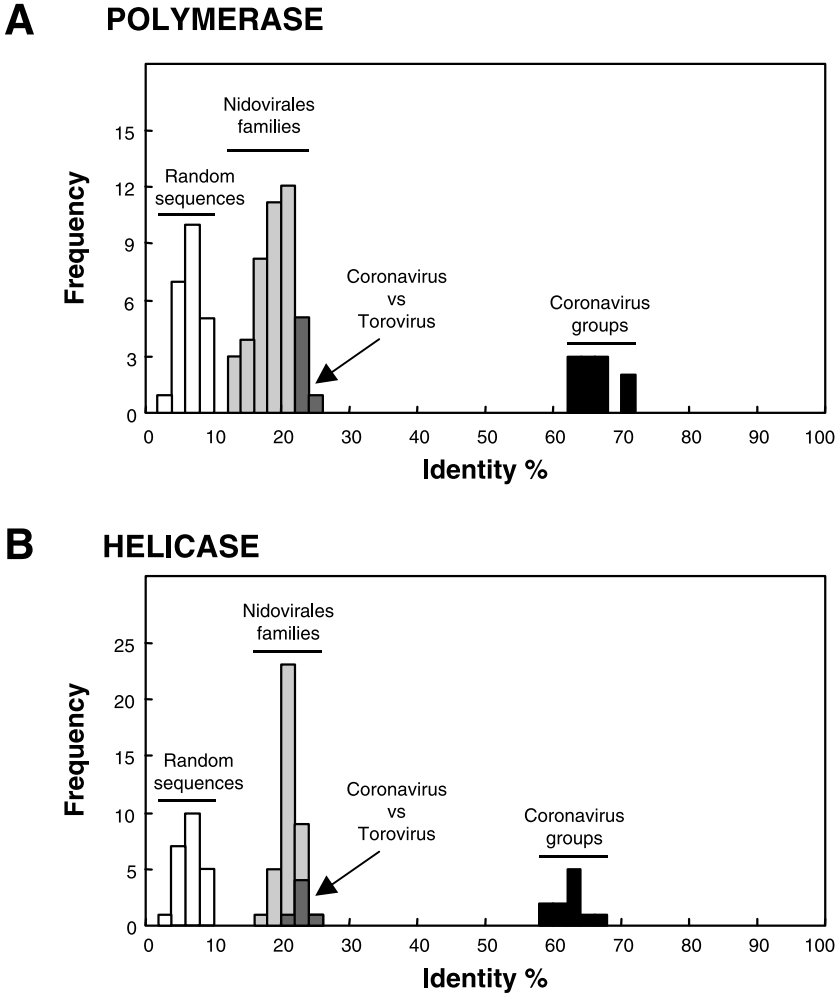


Fig. 9. Frequency distributions of pairwise identity percentages of two replicase domains in the *Nidovirales*. Amino acid sequences of putative RdRp (**A**) and the HEL (**B**) domains were retrieved from databases and aligned with CLUSTAL X and T-COFFEE to generate pairwise score matrices including percentages of identical residues in each pair of sequences. These matrices were processed to plot frequency distributions of percentage scores that were rounded with the step of 2%. Three groups of scores were colored and they correspond to comparisons: (i) between coronaviruses from different groups (black), (ii) between nidoviruses from different families (light gray), and (iii) between coronaviruses and toroviruses (dark gray). A frequency distribution of pairwise scores between random sequences (white) was also plotted

Coronaviruses and toroviruses are separated by a large interfamily-like distance

To evaluate the foundations of the genera structure of the *Coronaviridae*, comparative sequence analysis of two replicative domains of corona- and toroviruses, and two other nidovirus families – *Arteriviridae* and *Roniviridae* – was performed. The *Coronaviridae*-wide alignments of the RdRp and HEL domains (see above)

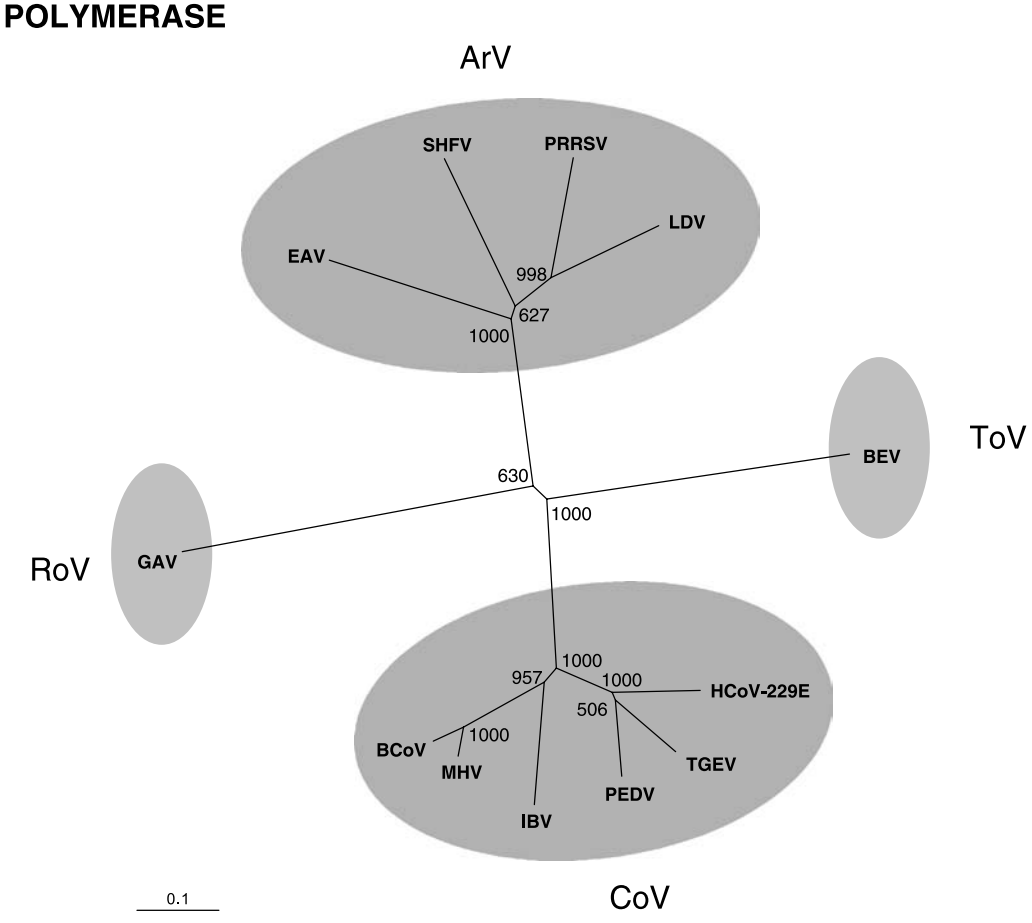


Fig. 10. Phylogenetic tree of putative RNA-dependent RNA polymerases of *Nidovirales*. The RdRp sequences of six coronaviruses (*CoV*), one torovirus (*ToV*), four arteriviruses (*ArV*), and one ronivirus (*RoV*) were aligned as described in the text. An unrooted dendrogram was generated with the Neighbor-joining method using the CLUSTAL X v1.82 program. Bootstrap values higher than 50% are shown in the main branch nodes. The sequences analyzed and their accession numbers are as follows: TGEV, strain Purdue, Q9IW06; HCoV-229E, Q9DLN1; PEDV, strain CV777, Q91AV2; MHV, strain A59, P16342; BCoV, strain Quebec, Q8V6W6; IBV, strain Beaudette, P26314; BEV, strain P138/72, P18458; PRRSV, Porcine respiratory and reproductive syndrome virus, strain Lelystad, Q04561; LDV, Lactate dehydrogenase-elevating virus, strain Plagemann, Q83018; SHFV, Simian hemorrhagic fever virus, strain LVR 42-0/M6941, P89132; EAV, Equine arteritis virus, strain Bucyrus, P19811; GAV, Gill-associated virus, Q9WPZ7

were expanded to include representative sequences of the other two families of the *Nidovirales* order.

Sequence identities of every sequence pair were extracted from the RdRp and HEL alignments. The pairwise identity percentages between sequences of the genera *Coronavirus* and *Torovirus* (from 22 to 25% and 21 to 25%) were very close to or overlapped with the value ranges (from 12 to 22% and 17 to 25%) obtained for pairwise interfamily comparisons, but were much smaller than values derived from pairwise comparisons for the *Coronavirus* genus (from 63 to 71% and 59 to 67%) (Fig. 9).

The RdRp and HEL alignments were also used to infer the neighbor-joining dendrograms that revealed different topologies. The position of the torovirus branch in the RdRp dendrogram was not confidently resolved (Fig. 10), and roniviruses interleaved between corona- and toroviruses in the HEL dendrogram (Fig. 11). Despite the observed differences, which may be due to technical reasons (see above), this phylogenetic analysis confirms toroviruses and coronaviruses to

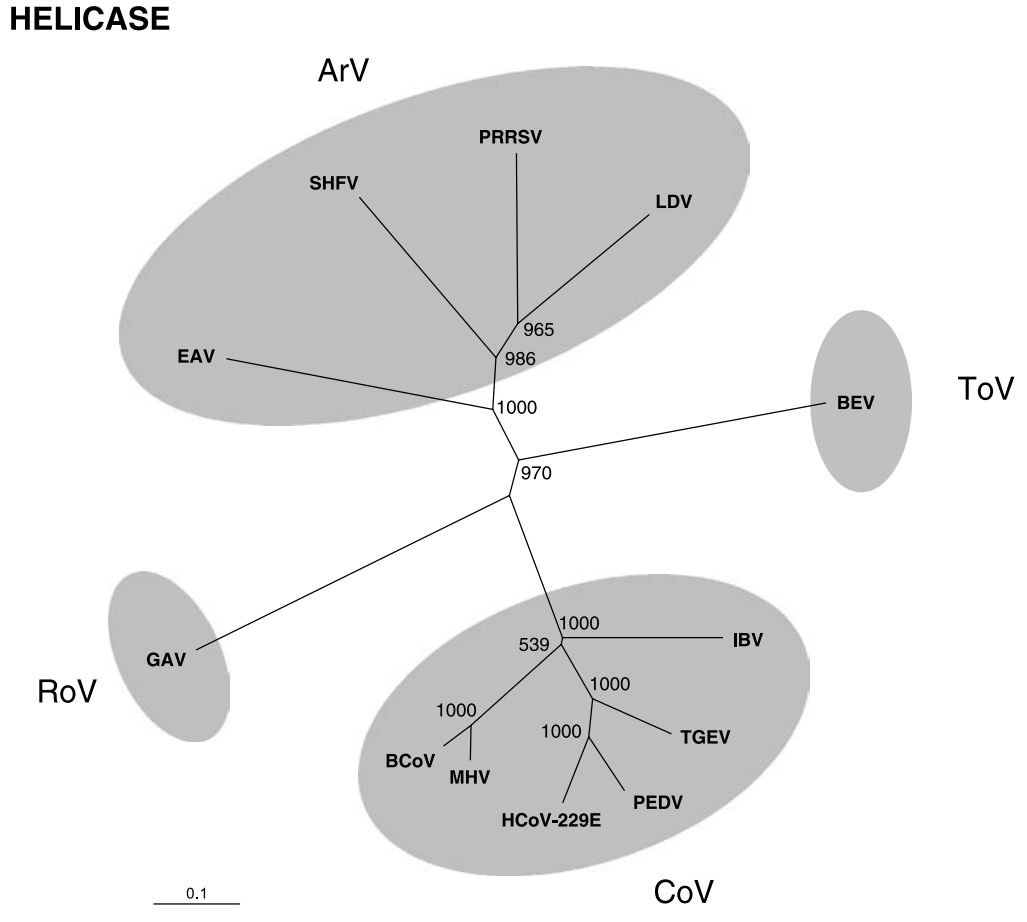


Fig. 11. Phylogenetic tree of RNA helicases of *Nidovirales*. An analysis was essentially identical to that described in Figure 10 except the HEL rather than RdRp domain sequences were processed

be separated by a large distance that is comparable to those between established nidovirus families.

Discussion

In this report we performed comparative sequence analysis of six functionally different proteins to revise the taxonomy of the *Coronaviridae*. The analyzed sets of proteins varied significantly in respect to the number and diversity of sequences, the S set being the most numerous and diverse, and the HEL and RdRp sets being the smallest. The results quantified the intervirus relationships for the *Coronaviridae* and, with some minor variations tolerable for the taxonomy classification, were consistent for all proteins analyzed.

Three coronavirus genetic groups have diverged sufficiently enough to form separate genera

Both pairwise scores and NJ-based phylogenetic analyses confirmed three genetic groups within the current *Coronavirus* genus that have been identified previously [67] and observed in analyses of other coronavirus sequences [11, 31, 76]. Do these groups belong to the same genus or are they prototypes for separate genera? The current guidelines for virus taxonomy do not provide firm quantitative sequence criteria, but rather leave it to each taxonomy study group to design and validate ranks for a family for which it is responsible [79]. As a result, intra- and inter-rank pairwise sequence conservation varies tremendously among different virus families. For instance, the *interfamily* conservation among RdRps of viruses of the genus *Sobemovirus* and a *Luteoviridae* genus is around 45% identical residues, and RdRps of the *Comoviridae* family share 37%, 26% and 24% pairwise identity with homologs of the *Sequiviridae*, *Picornaviridae* and *Caliciviridae* families, respectively. In contrast, significantly lower numbers characterize the *intergenera* RdRp conservation in the *Togaviridae* and *Flaviviridae* [88] (A. Gorbalenya, unpublished observation). These figures are correlated with other characteristics including genome organization and expression. To revise *Coronaviridae* taxonomy, we decided to follow *defacto* criteria that discriminate ranks in several complex virus families related to the *Picornaviridae* with which nidoviruses showed distant affinity [26].

Inspection of the intergroup pairwise scores obtained for the four structural proteins typical for coronaviruses show that the low end of these numbers (from 9 to 23%) is in a twilight zone occupied by highly diverged proteins [19]. Accordingly, no homologs of the N and E proteins have been found outside coronaviruses, and the only (very) distant homologs of S and M proteins are encoded by toroviruses (see below). In another test, we compared the pairwise scores obtained for structural proteins of coronaviruses and viruses of other densely populated families. It was revealed that the obtained *inter*-group pairwise scores for coronaviruses are similar to those calculated for the coat proteins of different genera of the *Potyviridae* family that have from 18 to 31% identical residues [65]. Accordingly, the coronavirus *intra*-group pairwise scores were closer to

the intragenera figures, which, for instance, show that amino acid identity in structural proteins of the *Picornaviridae* exceeds 50% [39].

Viruses belonging to different genera of the same family (e.g. *Picornaviridae*) also and typically have unique genus-specific features in their genome organisations [36, 39]. We analyzed coronaviruses from this perspective. All coronaviruses maintain a set of essential genes in an invariable order (rep-S-E-M-N), although there are other, apparently non-essential genes, whose presence and location varies and may be group-specific (Fig. 12A). Only group 2 members include the hemagglutinin-esterase (HE) gene, and only group 3 viruses have a gene located between the M and N genes [6, 44]. Also there are group-specific differences in

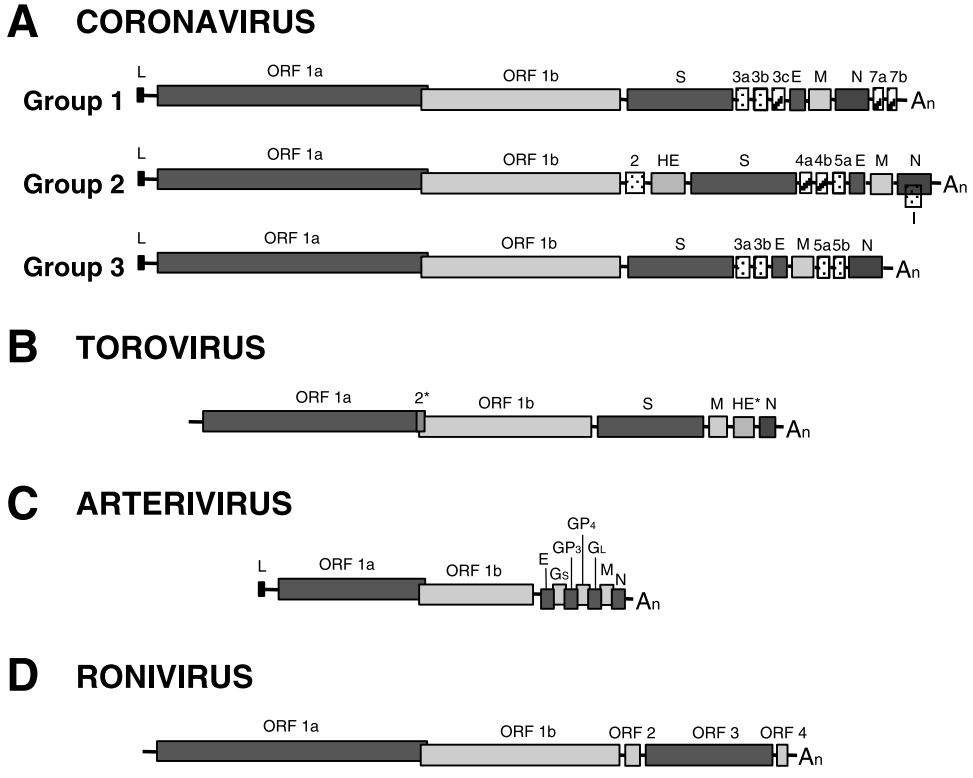


Fig. 12. Genome organizations of the prototype *Nidovirales* members. The genetic structures of prototypes members of *Nidovirales* has been deduced from previously published data [14–16, 21, 26, 44, 71, 73]. **A** Consensus genetic structure of group 1, 2 and 3 coronaviruses. Letters on the top of boxes indicate the genes for the replicase and the structural and non-structural proteins. Patterned boxes stand for nonstructural protein genes that are common to all (dotted boxes) or only some (striped boxes) species in that group. **B** Probable genetic structure of the torovirus BEV. **C** Genetic structure of the arterivirus EAV. **D** Genetic structure of the ronivirus GAV. Letters above the boxes indicate genes name. *L*, leader sequence; *HE*, hemagglutinin esterase; *S*, spike protein; *E*, envelope protein; *M*, membrane protein; *N*, nucleoprotein; *I*, internal ORF; *An*, poly-A tail; *G*S, small glycoprotein; *G*L, large glycoprotein; GP_3 and GP_4, non-essential envelope glycoproteins; BEV HE* gene has an incomplete sequence as compared with the coronavirus HE gene. 2*, sequence homologous to the non-structural protein 2 present in group 2 coronaviruses

the expression patterns of two genes. The replicase has an extra protein released from its N-terminus in the coronavirus groups 1 and 2 but not in group 3 members [91]. Likewise, the spike protein is posttranslationally cleaved into two halves in the coronavirus groups 2 and 3 but not in every group 1 members [8]. Although it is tempting to believe that the above characteristics are indeed group-specific, analysis of a more diverse set of coronavirus genomes will be essential to verify this perception.

Collectively, our analysis indicates that using criteria that were derived from analysis of the taxonomy of other virus families, the current three genetic groups of the *Coronaviridae* must each be elevated to the genus rank.

Do coronaviruses and toroviruses separate into subfamilies or families?

For the sake of consistency, the taxonomic revision proposed above must be accompanied by similar elevation of the *Coronavirus* genus, which currently unites these three groups, and the *Torovirus* genus to higher rank(s). There are two possible ranks to consider – subfamily or family. The subfamily rank is used in a few virus families and one order [48, 80]. To justify its usage in the *Coronaviridae*, the toroviruses and coronaviruses must share (significantly) more characters in common than each of them has with the other two *Nidovirales* families.

Analysis of virion architecture of nidoviruses revealed different nucleocapsid organizations in four major lineages with toroviruses having toroidal and coronaviruses having icosahedral internal virion structures [23, 24, 72]. These specifics are correlated with the lack of homologs of essential or semi-essential coronavirus N and E genes [43, 51] in toroviruses [72]. In contrast and unlike arteriviruses and roniviruses, toroviruses do have homologs of S and M proteins of coronaviruses that are encoded in the same (gene) order [70, 71]. Protein conservation is very weak and includes a previously recognized region in the M protein [18]. Accordingly, virions of coronaviruses and toroviruses slightly resemble each other (Figs. 13A and B) and differ from those of arteriviruses and roniviruses (Fig. 13C and D). The most characteristic feature common for virions of coronaviruses and toroviruses is the presence of the large peplomers that are formed by trimers of the S protein protruding from virion envelope (Figs. 13 A and B). Toroviruses also have homologs of the non-essential genes 2 and HE unique to the group 2 coronaviruses, where they are encoded between the replicase and S genes [12]. In toroviruses, these genes are located in other non-adjacent positions; an nsp2 homolog is encoded immediately upstream of the replicase frameshift as part of the replicase and an HE homolog maps between genes M and N (Fig. 12 A and B) [69]. In the equine torovirus the HE gene is partially truncated [12].

Coronaviruses, roniviruses, and arteriviruses, and most probably toroviruses [68], have the replicase gene encoded in two overlapping ORFs (ORF1a and ORF1b), the last of which is expressed through a frameshifting mechanism (Fig. 12 B, C, and D). Corona- and toroviruses, compared to roni- and arteriviruses, showed the longest collinearity in the replicase 1b region, as reported elsewhere

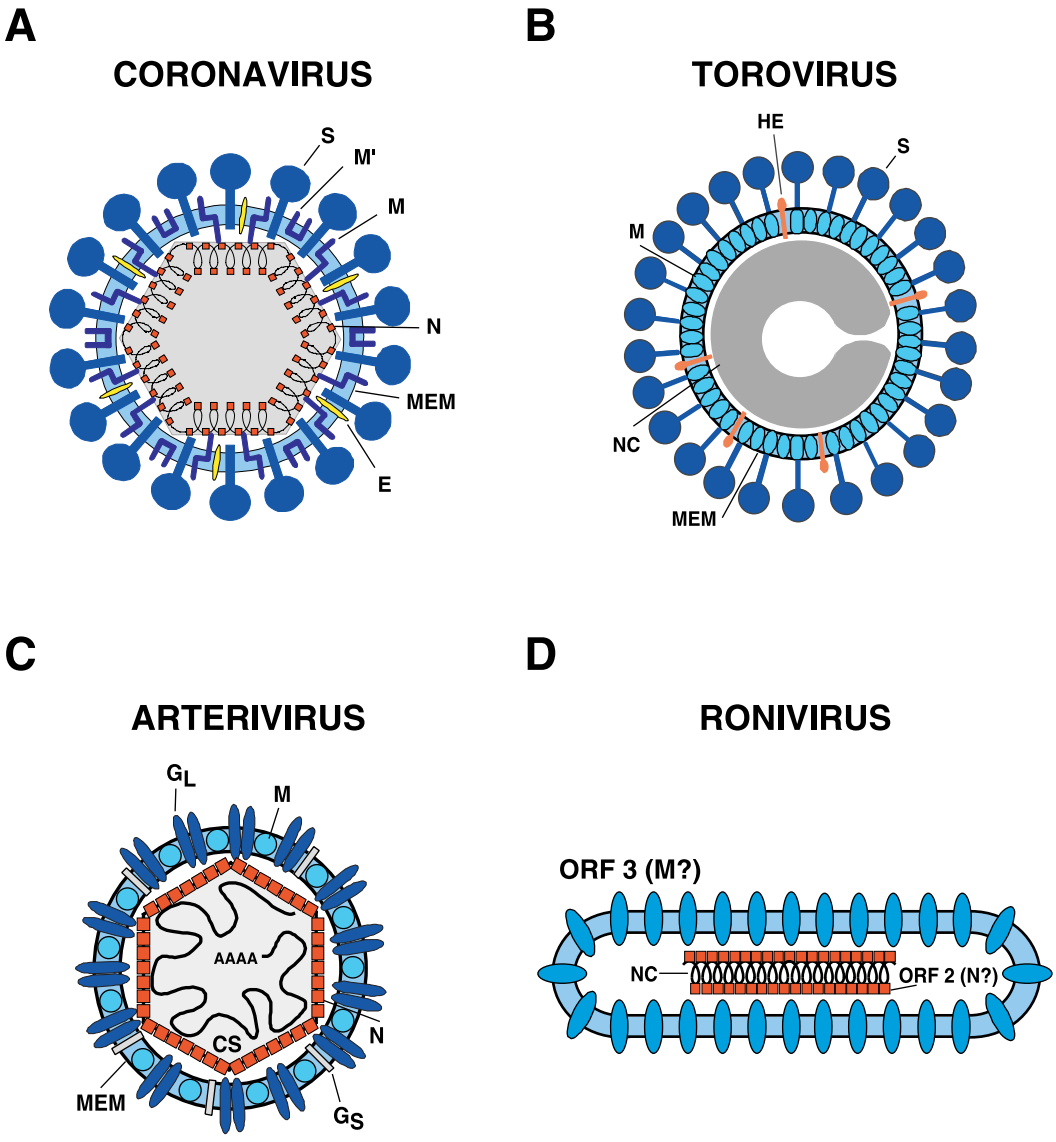


Fig. 13. Architectures of virus particles of the prototype *Nidovirales* members. Shown are cartoons depicting possible organizations of **A** Coronavirus [21]. **B** Torovirus [71]. **C** Arterivirus [73]. **D** Ronivirus (modified from an scheme by Peter Walker, Indooroopilly, Australia [75]) virions as deduced from available results. *MEM*, lipid membrane; *CS*, core shell; *NC*, nucleocapsid; *N*, nucleoprotein; *M*, membrane protein with the amino-terminal end facing the outside of the virus and the carboxyterminus located inside the virion; *M*', membrane protein with both the amino- and the carboxy-terminal end facing virus surface; *E*, envelope protein; *S*, spike protein; *HE*, hemagglutinin-esterase; *G*S, small glycoprotein; *G*L, large glycoprotein

[13, 16, 26]. For instance, arteriviruses seem to encode an unrelated protein to the putative methyltransferase domain, which is released from the C-terminus of replicase and may lack a homolog of the putative exonuclease domain, which is conserved between HEL and the nidovirus-specific domain in corona- and toroviruses

[26] (Snijder, Bredenbeek, Dobbe, Thiel, Ziebuhr, Poon, Guan, Rozanov, Spaan and Gorbalenya, submitted). This comparative analysis cannot currently be extended further to the N-terminus to include the 1a part of replicase, as the corresponding torovirus sequence is not in the public domain. This region, which includes more than 4000 amino acid residues in coronaviruses and roniviruses, is extremely poorly conserved among established families [13, 18, 26, 68, 89]. Future nidovirus-wide analysis of this region may be highly informative for taxonomic purposes.

The sequence affinity between toroviruses and coronaviruses summarized above was surprisingly not evident in our phylogenetic analysis of the two most conserved replicase domains of nidoviruses using the NJ algorithm. In contrast, in other analyses of the RdRp domain the expected clustering of coronaviruses and toroviruses was observed [13, 30]. The latter datasets included outgroup sequence(s), although the exact reasons of the observed striking variations of the tree topology may be more complex. Particularly, very large distances between four major lineages of the nidoviruses and significant under-representation of toro- and ronivirus sequences may have negatively affected the reliability of alignments and phylogenetic inference in our analyses. We believe therefore that the nidovirus phylogeny must be verified later when the number and diversity of torovirus and ronivirus sequences will match those of arteriviruses and coronaviruses.

Regardless of the actual topology of the nidovirus tree, the evident relatively large distance between RdRps of coronaviruses and toroviruses might alone be sufficient to justify separation of these viruses into different families, as has been argued by others [82]. Indeed, compared to the 22–25% amino acid residue identity between RdRps of coronaviruses and toroviruses, the RdRp conservation among a number of viruses that belong to other families is (substantially) higher (see above). Likewise, toro- and coronaviruses do not group together in respect to the transcription mechanism to express 3′-located open reading frames. For this purpose coronaviruses (and equally arteriviruses) employ discontinuous transcription of the genomic RNA [44, 61, 78], while toroviruses appear to rely upon both discontinuous and continuous transcriptions [81]. In this respect, roniviruses differ further as their sgRNAs may be produced exclusively through continuous transcription [14].

In summary, toroviruses and coronaviruses do have a number of important characteristics in common that group them together and separate them from arteriviruses and roniviruses. This list of common properties may well be extended in future when more sequences become available and characterization of these viruses advances. However, until criteria differentiating the family and subfamily ranks for viruses in general are clearly formulated, it remains a matter of personal preference to choose which characteristics – the numerous common or the few unique ones – should weigh more for revising the taxonomy of coronaviruses and toroviruses. Either of two possible decisions – to assign the family or subfamily rank to coronaviruses and toroviruses – seems to be compatible with the current guidelines of the ICTV and either of them may not satisfy everybody in the field.

Acknowledgments

This work has been supported by grants from the Comisión Interministerial de Ciencia y Tecnología (CICYT), La Consejería de Educación y Cultura de la Comunidad de Madrid, Fort Dodge Veterinaria, and the European Communities (Frame V, Key Action 2, Control of Infectious Disease Projects QLRT-1999-00002, QLRT-1999-30739, QLRT-2000-00874), and the Department of Food and Rural Affairs. JMG received a fellowship from the European Communities (Frame V, Key Action 2, Control of Infectious Diseases).

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402
3. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27: 260–262
4. Berke T, Matson DO (2000) Reclassification of the Caliciviridae into distinct genera and exclusion of hepatitis E virus from the family on the basis of comparative phylogenetic analysis. Arch Virol 145: 1421–1436
5. Bradley DW, Balayan MS (1988) Virus of enterically transmitted non-A, non-B hepatitis. Lancet 1: 819
6. Brian DA, Hogue BG, Kienzle TE (1995) The coronavirus hemagglutinin esterase glycoprotein. In: Siddell SG (ed) *The coronaviridae*, Plenum press, New York, pp 165–176
7. Brierley I (1995) Ribosomal frameshifting on viral RNAs. J Gen Virol 76: 1885–1892
8. Cavanagh D (1995) The coronavirus surface glycoprotein. In: Siddell SG (ed) *The Coronaviridae*, Plenum press, New York, pp 73–113
9. Cavanagh D, Brian DA, Briton M, Enjuanes L, Holmes KV, Lai MMC, Laude H, Plagemann PGW, Siddell SG, Spaan W, Taguchi F, Talbot P (1994) Revision of the taxonomy of the *Coronavirus, Torovirus*, and *Arterivirus* genera. Arch Virol 135: 227–237
10. Cavanagh D, Mawditt K, Welchman DdB, Britton P, Gough RE (2002) Coronaviruses from pheasants (*Phasianus colchicus*) are genetically closely related to coronaviruses of domestic fowl (infectious bronchitis virus) and turkeys. Avian Pathol 31: 81–93
11. Chouljenko VN, Lin XQ, Kousoulas KG, Gorbalenya AE (2001) Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pneumonia. J Gen Virol 82: 2927–2933
12. Cornelissen LAHM, Wierda CMH, Van Der Meer FJ, Herrewegh APM, Horzinek MC, Egberonk HF, Groot RJ (1997) Hemagglutinin-esterase, a novel structural protein of torovirus. J Virol 71: 5277–5286
13. Cowley JA, Dimmock CM, Spann KM, Walker PJ (2000) Gill-associated virus of *Penaeus monodon* prawns: an invertebrate virus with ORF1a and ORF1b genes related to arteri- and coronaviruses. J Gen Virol 81: 1473–1484
14. Cowley JA, Dimmock CM, Walker PJ (2002) Gill-associated nidovirus of *Penaeus monodon* prawns transcribes 3′-coterminal subgenomic mRNAs that do not possess 5′-leader sequences. J Gen Virol 83: 927–935

15. Cowley JA, Walker PJ (2002) The complete genome sequence of gill-associated virus of Penaeus *monodon* prawns indicates a gene organisation unique among nidoviruses. Arch Virol 147: 1977–1987
16. de Vries AAF, Horzinek MC, Rottier PJM, de Groot RJ (1997) The genome organization of the *Nidovirales*: similarities and differences between arteri-, toro-, and coronaviruses. Semin Virol 8: 33–47
17. den Boon JA, Snijder EJ, Chirnside ED, De Vries AAF, Horzinek MC, Spaan WJM (1991a) Equine arteritis virus is not a togavirus but belongs to the coronaviruslike superfamily. J Virol 65: 2910–2920
18. den Boon JA, Snijder EJ, Locker JK, Horzinek MC, Rottier JM (1991b) Another triple-spanning envelope protein among intracellularly budding RNA viruses: the torovirus E protein. Virology 182: 655–663
19. Doolittle RF (1986) Of URFs and ORFs: A primer on how to analyze derived amino acid sequences. University Science Books, Mill Valley, CA
20. Duckmanton L, Luan B, Devinish J, Tellier R, Petric M (1997) Characterization of torovirus from human fecal specimens. Virology 239: 158–168
21. Enjuanes L, Brian D, Cavanagh D, Holmes K, Lai MMC, Laude H, Masters P, Rottier P, Siddell SG, Spaan WJM, Taguchi F, Talbot P (2000a) *Coronaviridae*. In: van Regenmortel MHV, Fauquet CM, Bishop DHL, Carsten EB, Estes MK, Lemon SM, McGeoch DJ, Maniloff J, Mayo MA, Pringle CR, Wickner RB (eds) Virus taxonomy. Classification and nomenclature of viruses. Academic Press, San Diego, pp 835–849
22. Enjuanes L, Spaan W, Snijder E, Cavanagh D (2000b) *Nidovirales*. In: van Regenmortel MHV, Fauquet CM, Bishop DHL, Carsten EB, Estes MK, Lemon SM, McGeoch DJ, Maniloff J, Mayo MA, Pringle CR, Wickner RB (eds) Virus taxonomy. Classification and nomenclature of viruses. Academic Press, San Diego, pp 827–834
23. Escors D, Camafeita E, Ortego J, Laude H, Enjuanes L (2001b) Organization of two transmissible gastroenteritis coronavirus membrane protein topologies within the virion and core. J Virol 75: 12228–12240
24. Escors D, Ortego J, Laude H, Enjuanes L (2001a) The membrane M protein carboxy terminus binds to transmissible gastroenteritis coronavirus core and contributes to core stability. J Virol 75: 1312–1324
25. Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c: Distributed by the author. Department of Genetics, University of Washington, Seattle
26. Gorbalenya AE (2001) Big nidovirus genome. When count and order of domains matter. In: Lavi E, Weiss S, Hingley ST (eds) *The Nidoviruses* (*Coronaviruses and Arteriviruses*). Kluwer Academic/Plenum Publishers, New York, pp 1–17
27. Gorbalenya AE, Koonin EV (1993) Helicases: amino acid sequence comparisons and structure-function relationships. Curr Opin Struc Biol 3: 419–429
28. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1988) A novel superfamily of nucleoside triphosphate-bining motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. FEBS Lett 235: 16–24
29. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1989) Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. Nucleic Acids Res 17: 4847–4861
30. Gorbalenya AE, Pringle FM, Zeddam J-L, Luke BT, Cameron CE, Kalmakoff J, Hanzlik TN, Gordon KHJ, Ward VK (2002) The Palm Subdomain-based Active Site is Internally Permuted in Viral RNA-dependent RNA Polymerases of an Ancient Lineage. J Mol Biol 324: 47–62

31. Hegyi A, Friebe A, Gorbalenya AE (2002) Mutational analysis of the active centre of coronavirus 3C-like proteases. J Gen Virol 83: 581–593
32. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the Blocks Database servers. Nucleic Acids Res 28: 228–230
33. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 163: GC17–GC26
34. Hogue BG, King B, Brian DA (1984) Antigenic relationships among proteins of bovine coronavirus, human respiratory coronavirus OC43, and mouse hepatitis coronavirus A59. J Virol 51: 384–388
35. Horzinek MC, Flewett TH, Saif LJ, Spaan WJ, Weiss M, Woode GN (1987) A new family of vertebrate viruses: Toroviridae. Intervirology 27: 17–24
36. Johansson S, Niklasson B, Maizel J, Gorbalenya AE, Lindberg AM (2002) Molecular analysis of three Ljungan virus isolates reveals a new, close-to-root lineage of the Picornaviridae with a cluster of two unrelated 2A. J Virol 76: 8920–8930
37. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282
38. Kadare G, Haenni A-L (1997) Virus-encoded RNA helicases. J Virol 71: 2583–2590
39. King AMQ, Brown F, Christian P, Hovi T, Hyypia T, Knowles NJ, Lemon SM, Minor PD, Palmenberg AC, Skern T, Stanway G (2000) Picornaviridae. In: van Regenmortel MHV, Fauquet CM, Bishop DHL, Carsten EB, Estes MK, Lemon SM, McGeoch DJ, Maniloff J, Mayo MA, Pringle CR, Wickner RB (eds) Virus taxonomy. Classification and nomenclature of viruses. Academic Press, San Diego, pp 657–678
40. Koonin EV (1991) The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. J Gen Virol 72: 2197–2206
41. Koonin EV, Gorbalenya AE, Purdy MA, Rozanov MN, Reyes GR, Bradley DW (1992) Computer-assisted assignment of functional domains in the nonstructural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses. Proc Natl Acad Sci USA 89: 8259–8263
42. Kroneman A, Cornelissen LAHM, Horzinek MC, de Groot RJ, Egberink HF (1998) Identification and characterization of a porcine torovirus. J Virol 72: 3507–3511
43. Kuo L, Hurst R, Masters PS (2002) MHV E protein plays a critical, but not essential, role in MHV replication. In: 21st Annual Meeting, ASV (ed). American Society for Virology, Lexington, Kentucky
44. Lai MMC, Cavanagh D (1997) The molecular biology of coronaviruses. Adv Virus Res 48: 1–100
45. Mayo MA (2002) A summary of taxonomic changes recently approved by ICTV. Arch Virol 147: 1655–1656
46. McIntosh K (1974) Coronaviruses: a comparative review. Cur Top Microbiol Immunol 63: 86–129
47. Motokawa K, Hohdatsu T, Aizawa C, Koyama H, Hashimoto H (1995) Molecular cloning and sequence determination of the peplomer protein gene of feline infectious peritonitis virus type I. Arch Virol 140: 469–480
48. Murphy FA, Fauquet CM, Bishop DHL, Ghabrial SA, Jarvis AW, Martelli GP, Mayo MA, Summers MD (1995) Part I: Introduction to the universal system of virus taxonomy. In: Murphy FA, Fauquet CM, Bishop DHL, Ghabrial SA, Jarvis AW, Martelli GP, Mayo MA, Summers MD (eds) Virus Taxonomy. Sixth Report of the International Committee on Taxonomy of Viruses. Springer, Wien New York, pp 1–13
49. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217

50. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform 3: 275–284
51. Ortego J, Sola I, Almazan F, Ceriani JE, Riquelme C, Balasch M, Plana-Durán J, Enjuanes L (2003) Transmissible gastroenteritis coronavirus gene 7 is not essential but influences *in vivo* virus replication and virulence. Virology 308: 13–22
52. Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 12: 357–358
53. Pedersen NC, Ward J, Mengeling WL (1978) Antigenic relationship of the feline infectious peritonitis virus to coronaviruses of other species. Arch Virol 58: 45–53
54. Perriere G, Gouy M (1996) WWW-Query: An on-line retrieval system for biological sequence banks. Biochimie 78: 364–369
55. Petric M, Tellier R (2000) Torovirus: emerging pathogens of humans and animals. In: Scheld WM, Craig WA, Hughes JM (eds) Emerging infections. ASM Press, Washington, pp 23–32
56. Pietrokovski S (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 24: 3836–3845
57. Pringle CR (1992) Committee pursues medley of virus taxonomic issues. ASM News 58: 475–476
58. Saituo NM, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425
59. Sánchez CM, Jiménez G, Laviada MD, Correa I, Suñé C, Bullido MJ, Gebauer F, Smerdou C, Callebaut P, Escribano JM, Enjuanes L (1990) Antigenic homology among coronaviruses related to transmissible gastroenteritis virus. Virology 174: 410–417
60. Sapats SI, Ashton F, Wright PJ, Ignjatovic J (1996) Novel variation in the N protein of avian infectious bronchitis virus. Virology 226: 412–417
61. Sawicki SG, Sawicki DL (1998) A new model for coronavirus transcription. Adv Exp Med Biol 440: 215–220
62. Schuler GD, Altschul SF, Lipman DJ (1991) A workbench for multiple alignment construction and analysis. Proteins: Structure, Function, and Genetics 9: 180–190
63. Seybert A, Hegyi A, Siddell SG, Ziebuhr J (2000a) The human coronavirus 229E superfamily 1 helicase has RNA and DNA duplex-unwinding activities with 5′-to-3′ polarity. RNA 6: 1056–1068
64. Seybert A, van Dinten LC, Snijder EJ, Ziebuhr J (2000b) Biochemical characterization of the equine arteritis virus helicase suggests a close functional relationship between arterivirus and coronavirus helicases. J Virol 74: 9586–9593
65. Shukla DD, Ward CW, Brunt AA (1994) The *Potyviridae*. Cab International, Oxon, pp. 516
66. Siddell SG (1995a) The *Coronaviridae*. In: Fraenkel-Conrat H, Wagner RR (eds) *The Viruses*. Plenum Press, New York, pp. 418
67. Siddell SG (1995b) The Coronaviridae: an introduction. In: Siddell SG (ed) *The Coronaviridae*. Plenum Press, New York, pp 1–10
68. Snijder EJ, den Boon JA, Bredenbeek PJ, Horzinek MC, Rijnbrand R, Spaan WJM (1990a) The carboxy-terminal part of the putative Berne virus polymerase is expressed by ribosomal frameshifting and contains sequence motifs which indicate that toro- and coronaviruses are evolutionarily related. Nucleic Acids Res 18: 4535–4542
69. Snijder EJ, den Boon JA, Horzinek MC, Spaan WJM (1991) Comparison of the genome organization of toroviruses and coronaviruses – evidence for two nonhomologous RNA recombination events during Berne virus evolution. Virology 180: 448–452

70. Snijder EJ, den Boon JA, Spaan WJM, Weiss M, Horzinek MC (1990b) Primary structure and post-translational processing of the Berne virus peplomer protein. Virology 178: 355–363
71. Snijder EJ, Horzinek MC (1993) Toroviruses: replication, evolution and comparison with other members of the coronavirus-like superfamily. J Gen Virol 74: 2305–2316
72. Snijder EJ, Horzinek MC (1995) The molecular biology of toroviruses. In: Siddell SG (ed) *The Coronaviridae*. Plenum Press, New York, pp 219–238
73. Snijder EJ, Meulenberg JJM (1998) The molecular biology of arteriviruses. J Gen Virol 79: 961–979
74. Spaan W, Cavanagh D, Horzinek MC (1990) Coronaviruses. In: van Regenmortel MHV, Neurath AR (eds) Immunochemistry of viruses, II. The basis for serodiagnosis and vaccines. Elsevier, Amsterdam, pp 359–379
75. Spann KM, Vickers JE, Lester RJG (1995) Lymphoid organ virus of *Penaeus monodon* from Australia. Dis Aquat Org 23: 127–134
76. Stephensen CB, Casebolt DB, Gangopadhyay NN (1999) Phylogenetic analysis of a highly conserved region of the polymerase gene from 11 coronaviruses and development of a consensus polymerase chain reaction assay. Virus Res 60: 181–189
77. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignement aided by quality analysis tools. Nucleic Acids Res 24: 4876–4882
78. van Marle G, Dobbe JC, Gultyaev AP, Luytjes W, Spaan WJM, Snijder EJ (1999) Arterivirus discontinuous mRNA transcription is guided by base pairing between sense and antisense transcription-regulating sequences. Proc Nat Acad Sc USA 96: 12056–12061
79. van Regenmortel MHV (2000) Introduction to the species concept in virus taxonomy. In: van Regenmortel MHV, Fauquet CM, Bishop DHL, Carsten EB, Estes MK, Lemon SM, Maniloff J, Mayo MA, McGeoch DJ, Pringle CR, Wickner RB (eds) Virus taxonomy. Classification and nomenclature of viruses. Academic Press, San Diego, pp 3–16
80. van Regenmortel MHV, Fauquet CM, Bishop DHL, Carsten EB, Estes MK, Lemon SM, McGeoch DJ, Maniloff J, Mayo MA, Pringle CR, Wickner RB (2000) Virus taxonomy. Classification and nomenclature of viruses, Academic Press, pp. 1–1162
81. van Vliet ALW, Smits SL, Rottier PJM, de Groot RJ (2002) Discontinuous and non-discontinuous subgenomic RNA transcription in a nidovirus. EMBO J 21: 6571–6580
82. Ward CW (1993) Progress towards a higher taxonomy of viruses. Res Virol 144: 419–453
83. Ward CW, McKern NM, Frenkel MJ, Shukla DD (1992) Sequence data is the major criterion for potyvirus classification. Arch Virol Suppl 5: 283–297
84. Wege H, Siddell S, Ter Meulen V (1982) The biology and pathogenesis of coronaviruses. Curr Top Microbiol Immunol 99: 165–200
85. Womble DD (2000) GCG: The Wisconsin Package of sequence analysis programs. Method Mol Biol 132: 3–22
86. Wu CH, Yeh L-SL, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J, Barker WC (2003) The Protein Information Resource. Nucleic Acids Res 31: 345–347
87. Yuan YP, Eulenstein O, Vingron M, Bork P (1998) Towards detection of orthologues in sequence databases. Bioinformatics 14: 285–289
88. Zanotto PMA, Gibbs MJ, Gould EA, Holmes EC (1996) A reevaluation of the higher taxonomy of viruses based on RNA polymerases. J Virol 70: 6083–6096
89. Ziebuhr J, Bayer S, Cowley JA, Gorbalenya AE (2003) The 3C-like proteinase of an invertebrate nidovirus links coronavirus and potyvirus homologs. J Virol 77: 1415–1426

90. Ziebuhr J, Snijder EJ, Gorbalenya AE (2000) Virus-encoded proteinases and proteolytic processing in the *Nidovirales*. J Gen Virol 81: 853–879
91. Ziebuhr J, Thiel V, Gorbalenya AE (2001) The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. J Biol Chem 276: 33220–33232.

Authors' addresses: Luis Enjuanes, Department of Molecular and Cell Biology, Centro Nacional de Biotecnología, CSIC, Campus Universidad Autónoma, Cantoblanco, 28049 Madrid, Spain; e-mail: L.Enjuanes@cnb.uam.es and Alexander E. Gorbalenya, Center of Infectious Diseases, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands; e-mail: a.e.gorbalenya@lumc.nl