Arch Virol (2004) 149: 621–624
DOI 10.1007/s00705-003-0244-0

Archives of
Virology
Printed in Austria

The phylogeny of SARS coronavirus

Brief Report

A. J. Gibbs, M. J. Gibbs, and **J. S. Armstrong**

School of Botany and Zoology, Australian National University, Canberra, Australia

Received July 10, 2003; accepted September 29, 2003
Published online January 5, 2004

Summary. Different tree-building methods consistently place the SARS coronavirus (SARS-CoV) as a basal Group 2 coronavirus rather than as an ungrouped species as concluded by others. Detailed comparisons of the SARS-CoV genomic sequence with those of six other coronaviruses failed to find evidence of recombination or genomic rearrangement using computational methods designed for that purpose.

*

In their report of the SARS-CoV genomic sequence, Marra et al. [4] constructed unrooted phylogenetic trees by boot strapped distance methods using the encoded major protein sequences of SARS-CoV and representatives of the three coronavirus groups and reported that those encoded by SARS-CoV “do not cluster more closely with any one group”, and concluded that SARS-CoV should “be considered the first representative of “Group 4” coronaviruses”. Likewise Rota et al. [5] using the same methods reported that SARS-CoV “is approximately equidistant from all previously characterized coronaviruses” and “forms a distinct group within the genus *Coronavirus*”. However neither of those studies had outgroup sequences in the analyses, and therefore it was not possible to be sure whether SARS-CoV showed consistent relationships with other coronaviruses.

The largest gene of the SARS-CoV genome is that which encodes the multifunctional polymerase and comprises 70.7% of the genome, and the next largest is the spike (peplomer) gene, a further 12.7%. We found the polymerase genes of the arteriviruses to be clearly homologous to those of coronaviruses as GenBank searches using BLASTX gave E values around $1 \times 10^{-4-5}$ between arterivirus and coronavirus sequences, and they therefore provide an appropriate outgroup to coronavirus polymerases, however similar searches using BLASTN −X and −P failed to find outgroup sequences for the spike protein gene.

The complete polymerase genes of SARS-CoV, six other coronaviruses and two arteriviruses were aligned [3, 8] either as amino acids or as nucleotides, or as nucleotides grouped as codons. The resulting gapped nucleotide sequences had one or more sequences with gaps at over 50% of the homologous positions, so separate comparisons were made of the fully gapped sequences, the parts of the alignment present in all sequences, and also the gap-containing parts that had been removed when degapping. Trees obtained by the neighbor-joining method and, especially, using a maximum-likelihood algorithm [7] (Fig. 1), placed

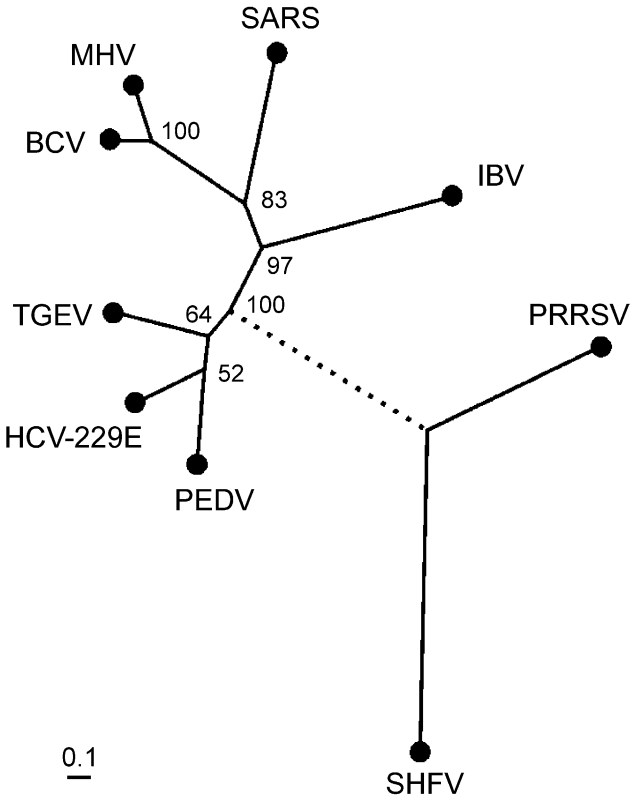

Fig. 1. The relationships of seven coronavirus and two arterivirus polymerase genes. Group 1; TGEV, *Transmissible gastroenteritis virus* (Accession Code NC_002306); HCV-229E, *Human coronavirus-229E* (NC_002645); PEDV, *Porcine epidemic diarrhea virus* (NC_003436). Group 2: MHV, *Murine hepatitis virus* (NC_001846); BCV, *Bovine coronavirus* (NC_003045). Group 3: IBV, *Infectious bronchitis virus* (NC_001451). Arterivirus outgroup: PRRSV, *Porcine respiratory and reproductive syndrome virus* (NC_001961); SHFV, *Simian hemorraghic fever virus* (NC_003092). The sequences were aligned and gapped using the CLUSTALX [3, 8] alignment of their encoded proteins, and the trees computed by TREE-PUZZLE 5.0 [7] using the HKY model of substitution, the mixed model of rate heterogeneity and 1000 puzzlings steps which gave 83% support for the SARS-CoV, MHV and BCV cluster; the scale shows maximum likelihood branch length, and the broken line connecting the outgroup had a length of 6.5

SARS-CoV as a well-supported sister lineage to the two Group 2 coronaviruses, *Murine hepatitis virus* (MHV) and *Bovine coronavirus* (BCV), with the Group 3 coronavirus, *Infectious bronchitis virus* (IBV), consistently placed as a sister lineage to the Group 2-SARS-CoV clade.

This topology is totally consistent with those of the trees reported in the first SARS-CoV sequence analyses [4, 5] if we assume that they have the same root, and it is also consistent with rootless trees calculated from the spike protein genes.

The taxonomy of the remaining 16.6% of the SARS-CoV genome is less certain. For example, whereas taxonomies place the nucleocapsid proteins in the same topology as the polymerase and spike genes, the nucleocapsid genes (4.3% of the SARS-CoV genome) place SARS-CoV as sister to IBV.

The SARS-CoV genomic sequence was also compared with those of the three most closely related coronaviruses, MHV, BCV and IBV, in pairwise dotplots [2, 6] using a window of 25 nucleotides. These comparisons showed that the SARS-CoV genome is co-linear with those of the other three coronaviruses, and is not significantly rearranged, and although, when the sequence was compared with itself, many regions of short repetitions were found, none were significantly different from those in the other three coronavirus genomes.

Group 2 coronaviruses differ from other coronaviruses in possessing a haemagglutinin-esterase (HE) gene between the polymerase and spike protein genes. This has presumably been acquired recently but is of unknown origin [10]. Differences in the enzymic specificity of the HE gene, and congruent gene sequence differences, place Group 2 coronaviruses into two sub-groups; the MHV-like (Group 2a) and BCV-like (Group 2b) viruses [10]. We therefore suggest that it will be most useful to consider SARS-CoV as a Group 2c coronavirus, and to represent the original Group 2 lineage from which the Group 2a and 2b lineages were derived after acquiring a progenitor HE gene rather than as the first species of a novel group, as an emphasis on taxonomic similarities, rather than differences, may generate more useful predictions.

We also examined the aligned sequences using the SiScan [1] and PhylPro [9] programs that are specifically designed to detect the phylogenetic anomalies resulting from recombination, and found no significant anomalies with subsequences as short as 100 nts, irrespective of whether all differences or just synonymous or non-synonymous differences were examined.

Note added after submission

A study that found the same phylogenetic placement of SARS-CoV and other coronaviruses using a different outgroup sequence and different algorithms has been reported by Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus Group 2 lineage. J Mol Biol 29: 991–1004.

References

1. Gibbs MJ, Armstrong JS, Gibbs AJ (2000) Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16: 573–582; http://www.anu.edu.au/BoZo/software/
2. Gibbs AJ, McIntyre GA (1970) The diagram, a method for comparing sequences, its use with amino acid and nucleotide sequences. Eur J Biochem 16: 1–11
3. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23: 403–405
4. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY, Cloutier A, Coughlin SM, Freeman D, Girn N, Griffith OL, Leach SR, Mayo M, McDonald H, Montgomery SB, Pandoh PK, Petrescu AS, Robertson AG, Schein JE, Siddiqui A, Smailus DE, Stott JM, Yang GS, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth TF, Bowness D, Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A, Kabani A, Li Y, Normand S, Stroher U, Tipples GA, Tyler S, Vogrig R, Ward D, Watson B, Brunham RC, Krajden M, Petric M, Skowronski DM, Upton C, Roper RL (2003) The genome sequence of the SARS-associated coronavirus. Science 300: 1377–1404
5. Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH, Tong S, Tamin A, Lowe L, Frace M, DeRisi JL, Chen Q, Wang D, Erdman DD, Peret TC, Burns C, Ksiazek TG, Rollin PE, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus AD, Drosten C, Pallansch MA, Anderson LJ, Bellini WJ (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300: 1394–1399
6. Sonnhammer ELL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167: 1–10
7. Strimmer K, von Haeseler A (1996) Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 13: 964–969
8. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W. Improving the sensitively of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680
9. Weiller GF (1998) Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol 15: 326–335
10. Wurzer WJ, Obojes K, Vlasak R (2002) The sialate-4-O-acetylesterases of coronaviruses related to mouse hepatitis virus: a proposal to reorganize group 2 Coronaviridae. J Gen Virol 83: 395–402

Author's address: Prof. Adrian Gibbs, School of Botany and Zoology, Australian National University, Canberra, ACT 0200, Australia; e-mail: adrian.gibbs@anu.edu.au