Sciencexpress Research Article

The Genome Sequence of the SARS-Associated Coronavirus

Marco A. Marra,[1]* Steven J. M. Jones,[1] Caroline R. Astell,[1] Robert A. Holt,[1] Angela Brooks-Wilson,[1] Yaron S. N. Butterfield,[1] Jaswinder Khattra,[1] Jennifer K. Asano,[1] Sarah A. Barber,[1] Susanna Y. Chan,[1] Alison Cloutier,[1] Shaun M. Coughlin,[1] Doug Freeman,[1] Noreen Girn,[1] Obi L. Griffith,[1] Stephen R. Leach,[1] Michael Mayo,[1] Helen McDonald,[1] Stephen B. Montgomery,[1] Pawan K. Pandoh,[1] Anca S. Petrescu,[1] A. Gordon Robertson,[1] Jacqueline E. Schein,[1] Asim Siddiqui,[1] Duane E. Smailus,[1] Jeff M. Stott,[1] George S. Yang[1]

Francis Plummer,[2] Anton Andonov,[2] Harvey Artsob,[2] Nathalie Bastien,[2] Kathy Bernard,[2] Timothy F. Booth,[2] Donnie Bowness,[2] Michael Drebot,[2] Lisa Fernando,[2] Ramon Flick,[2] Michael Garbutt,[2] Michael Gray,[2] Allen Grolla,[2] Steven Jones,[2] Heinz Feldmann,[2] Adrienne Meyers,[2] Amin Kabani,[2] Yan Li,[2] Susan Normand,[2] Ute Stroher,[2] Graham A. Tipples,[2] Shaun Tyler,[2] Robert Vogrig,[2] Diane Ward,[2] Brynn Watson[2]

Robert C. Brunham,[3] Mel Krajden,[3] Martin Petric,[3] Danuta M. Skowronski[3]

Chris Upton,[4] Rachel L. Roper[4]

[1]British Columbia Cancer Agency Genome Sciences Centre, 600 West 10th Avenue, Vancouver, British Columbia V5Z 4E6, Canada. [2]National Microbiology Laboratory, 1015 Arlington Street, Winnipeg, Manitoba R3E 3R2, Canada. [3]British Columbia Centre for Disease Control and University of British Columbia Centre for Disease Control, 655 West 12th Avenue, Vancouver, British Columbia V5Z 4R4, Canada. [4]Department of Biochemistry and Microbiology, University of Victoria, Post Office Box 3055 STN CSC, Victoria, British Columbia V8W 3P6, Canada.

*To whom correspondence should be addressed. E-mail: mmarra@bccgsc.ca

We sequenced the 29,751-base genome of the severe acute respiratory syndrome (SARS)–associated coronavirus known as the Tor2 isolate. The genome sequence reveals that this coronavirus is only moderately related to other known coronaviruses, including two human coronaviruses, HCoV-OC43 and HCoV-229E. Phylogenetic analysis of the predicted viral proteins indicates that the virus does not closely resemble any of the three previously known groups of coronaviruses. The genome sequence will aid in the diagnosis of SARS virus infection in humans and potential animal hosts (using PCR and immunological tests), in the development of antivirals (including neutralizing antibodies), and in the identification of putative epitopes for vaccine development.

An outbreak of atypical pneumonia, referred to as severe acute respiratory syndrome (SARS) and first identified in Guangdong Province, China, has spread to several countries. The severity of this disease is such that the mortality rate appears to be ~3 to 6%. A number of laboratories worldwide have undertaken the identification of the causative agent (*1, 2*). The National Microbiology Laboratory in Canada obtained the Tor2 isolate from a patient in Toronto, and succeeded in growing a coronavirus-like agent in African Green Monkey Kidney (Vero E6) cells. This coronavirus has been named publicly by the World Health Organization and member laboratories as "SARS virus" (press release issued by WHO April 16, 2003) following tests of causation according to Koch's postulates, including monkey inoculation (*3*). This virus was purified and its RNA genome extracted and sent to the British Columbia Centre for Disease Control in Vancouver for genome sequencing by the Genome Sciences Centre at the BC Cancer Agency.

The coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cells (*4*). They are distinguished by the presence of a single-stranded plus sense RNA genome approximately 30 kb in length that has a 5′ cap structure and 3′ polyA tract. Upon infection of an appropriate host cell, the 5′ most open reading frame (ORF) of the viral genome is translated into a large polyprotein that is cleaved by viral-encoded proteases to release several nonstructural proteins including an RNA-dependent RNA polymerase (Rep) and an ATPase helicase (Hel). These proteins in turn are responsible for replicating the viral genome as well as generating nested transcripts that are used in the synthesis of the viral proteins. The mechanism by which these subgenomic mRNAs are made is not fully understood, however recent evidence indicates that transcription regulating sequences (TSRs) at the 5′end of each gene represent signals that regulate the discontinuous transcription of subgenomic mRNAs (sgmRNAs). The TRSs include a partially conserved core sequence (CS) that in some coronaviruses is 5′-CUAAAC-3′. Two major models have been proposed to explain the discontinuous transcription in coronaviruses and arterioviruses (*5, 6*). The discovery of transcriptionally active, subgenomic-size minus strands containing the antileader sequence and of transcription intermediates active in the synthesis of mRNAs (*7–10*) favors the model of discontinuous transcription during the minus strand synthesis (*6*).

The viral membrane proteins, including the major proteins S (Spike) and M (Membrane), are inserted into the endoplasmic reticulum Golgi intermediate compartment (ERGIC) while full length replicated RNA (+ strands)

assemble with the N (nucleocapsid) protein. This RNA protein complex then associates with the M protein embedded in the membranes of the ER and virus particles form as the nucleocapsid complex buds into the ER. The virus then migrates through the Golgi complex and eventually exits the cell, likely by exocytosis (*4*). The site of viral attachment to the host cell resides within the S protein.

The coronaviruses include a large number of viruses that infect different animal species. The predominant diseases associated with these viruses are respiratory and enteric infections, although hepatic and neurological diseases also occur. Coronaviruses are divided into three serotypes, Groups 1, 2 and 3 (*11*). Phylogenetic analysis of coronavirus sequences also identifies three main classes of these viruses, corresponding to each of the three serotypes. Group 2 coronaviruses contain a hemagglutinin esterase (HE) gene homologous to that of Influenza C virus. It is presumed that the precursor of the Group 2 coronaviruses acquired HE as a result of a recombination event within a doubly infected host cell. We note that the Tor2 genome sequence appears to lack an HE gene.

Purification of viral particles and RNA, and DNA sequencing. Virus isolation was performed on a bronchoalveolar lavage specimen of a fatal SARS case belonging to the original case cluster from Toronto, Canada. Viral particles from this Tor2 isolate were purified and the genetic material (RNA) was extracted (*12*) from the Tor2 isolate (*13*). The RNA was converted to cDNA using a combined random-priming and oligo-dT priming strategy (*12*). Size selected cDNA products were cloned and single sequence reads were generated from each end of the insert from randomly picked clones. Sequences were assembled and the assembly edited to produce a draft sequence of the viral genome on April 12, 2003 (*12*). RACE (*12*) was performed to capture the 5´ end of the viral genome. The SARS genomic sequence has been deposited into Genbank (Accession AY274119.3). The final sequence we produced (also available as Release 3; www.bcgsc.bc.ca) is essentially identical to that released independently by the CDC (*14*). We report additional bases in the Tor2 sequence that correspond to the 3´ (encoded) polyA tail. Eight base differences between the two sequences could represent sequencing errors, PCR artifacts or mutable sites in the genome. The differences we detect between our sequence and that of the CDC are summarized in Table 1.

Non-protein coding features of the Tor2 SARS-CoV genome sequence. At the 5´ end of the genome we detected a putative 5´ leader sequence with similarity to the conserved coronavirus core leader sequence, 5´-CUAAAC-3 (*5, 6*). Putative TRS sequences were determined through manual alignment of sequences upstream of potential initiating methionine codons (see below) to the region of the coronavirus genome sequence containing the leader sequence (Table 2). Candidate TRS sequences were scored as strong, weak or absent based on inspection of the alignments.

The 3´ UTR sequence contains a 32 base-pair region corresponding to the conserved s2m motif (*15*). The s2m motif is believed to be a universal feature of astroviruses that has also been identified in avian infectious bronchitis virus (avian IBV) and the ERV-2 equine rhinovirus. The high degree of conservation between the s2m motifs in these different viruses and their evolutionary distance suggests that the avian IBV and ERV-2 have acquired the s2m motif through separate horizontal RNA transfer events (*15*). The inferred distance of the SARS coronavirus to IBV from our phylogenetic analysis (Fig. 1) would also suggest that the SARS coronavirus has obtained its s2m motif through a horizontal transfer event.

Predicted protein coding features of the Tor2 SARS-CoV genome sequence. Open reading frames were determined initially through sequence similarity to known coronavirus proteins. This approach identified replicases 1a and 1b, the S protein, the E protein, the M protein and the N protein. Orfs that did not match database sequences were identified if they were larger than 40 amino acids, unless a strong match to the TRS consensus was found close to and upstream of the potential initiating methionine residue. We note that Rota *et al.* (*14*) did not identify potential proteins of less than 50 amino acids. We attempted to identify putative TRSs upstream of all Orfs, both known and predicted (Tables 2 and 3). However, TRSs are not required for transcription of all coronavirus genes, as internal initiation from larger RNA transcripts is also able to facilitate translation (*16, 17*). Certain Orfs overlap (Orf 10 and 11, by 12 amino acids; Fig. 2), and some are contained entirely within another Orf or Orfs (Orf 4 and Orfs 13 and 14; Fig. 2). The biological relevance of these Orf predictions remains to be established, but in the cases of Orfs 10 and 11, we detect strong matches to the TRS consensus in close proximity to their respective initiating methionine codons (Table 2). Construction of unrooted phylogenetic trees using the set of known proteins and representatives of the three known coronaviral groups reveals that the proteins encoded by the SARS virus do not readily cluster more closely with any one group (Fig. 1). Hence, we propose that this isolate be considered the first representative of "Group 4" coronaviruses.

The coding potential of the 29,751-base genome is depicted in Fig. 2. Recognizable open reading frames include the replicase 1a and 1b translation products, the S glycoprotein, the E protein, the M protein and the N protein. We have, in addition, conducted a preliminary analysis of the nine novel Orfs, in an attempt to ascribe to them a possible functional role. These analyses are summarized below.

The replicase 1a (265-13,398 bp) and 1b Orfs (13,398 – 21,485 bp) occupy 21.2 kb of the SARS virus genome (Fig. 2). Conserved in both length and amino acid sequence to other coronavirus replicase proteins, the genes encode a number of proteins that are produced by proteolytic cleavage of a large polyprotein (*18*). As seen in other coronaviruses and as anticipated, a frame shift interrupts the protein-coding region, separating the 1a and 1b reading frames.

The S (Spike glycoprotein) (Fig. 2; 21,492 to 25,259 bp) encodes a surface projection glycoprotein precursor predicted to be 1,255 amino acids in length. Mutations in this gene have previously been correlated with altered pathogenesis and virulence in other coronaviruses (*4*). In some coronaviruses, the mature spike protein is inserted in the viral envelope with the majority of the protein exposed on the surface of the viral particles. It is believed that three molecules of the Spike protein form the characteristic peplomers or corona-like structures of this virus family. Our analysis of the spike glycoprotein with SignalP (*19*) reveals a high probability of a signal peptide (probability 0.996) with cleavage between residues 13 and 14. TMHMM (*20*) reveals a strong transmembrane domain near the C-terminal end. Together these data predict a type I membrane protein with the N-terminus and the majority of the protein (residues 14-1195) on the outside of the cell-surface or virus particle, in agreement with other coronavirus spike protein data. Supporting this conclusion, it has recently been shown that

for HCoV-229E virions, residues 417-546 are required for binding to the cellular receptor, aminopeptidase N (*21*). However it is known that various coronaviruses use different receptors, hence it is likely that different receptor binding sites are also used.

Orf 3 (Fig. 2; 25,268 - 26,092) encodes a predicted protein of 274 amino acids that lacks significant BLAST (*22*), FASTA (*23*) or PFAM (*24*) similarities to any known protein. Analysis of the N-terminal 70 amino acids with SignalP provides weak evidence for the existence of a signal peptide and a cleavage site (probability 0.540). Both TMpred (*25*) and TMHMM predict the existence of three trans-membrane regions spanning approximately residues 34-56, 77-99, and 103-125. The most likely model from these analyses is that the C-terminus and a large 149 amino - terminal domain would be located inside the viral or cellular membrane. The C-terminal (interior) region of the protein may encode a protein domain with ATP-binding properties (PD037277).

Orf 4 (Fig. 2; 25,689 - 26,153) encodes a predicted protein of 154 amino acids. This Orf overlaps entirely with Orf 3 and the E protein. Our analysis failed to locate a potential TRS sequence at the 5´ end of this putative Orf. However, it is possible this protein is expressed from the Orf3 mRNA using an internal ribosomal entry site. BLAST analyses failed to identify matching sequences. Analysis with TMPred predicts a single transmembrane helix.

The small envelope protein E (Fig. 2; 26,117 - 26,347) encodes a predicted protein of 76 amino acids. BLAST and FASTA comparisons indicate that the predicted protein exhibits significant matches to multiple envelope (alternatively known as small membrane) proteins from several coronaviruses. PFAM analysis of the protein reveals the predicted protein is a member of the well-characterized NS3_EnvE protein family (*24*). InterProScan (*26, 27*) analysis reveals that the protein is a component of the viral envelope, and conserved sequences are also found in other viruses, including gastroenteritis virus and murine hepatitis virus. SignalP analysis predicts the presence of a transmembrane anchor (probability 0.939). TMpred analysis of the predicted protein revealed a similar trans-membrane domain at positions 17-34, consistent with the known association of this protein with the viral envelope. TMHMM predicts a type II membrane protein with the majority of the hydrophilic domain (46 residues) and C terminus to be located on the surface of the viral particle. We note that in some coronaviruses such as transmissible gastroenteritis coronavirus (TGEV) the E protein is essential for virus replication (*28*) while in mouse hepatitis virus (MHV) it has been shown that although deletion of gene E reduces virus replication by more than ten thousand fold, the virus still can replicate (*29*).

The Membrane glycoprotein M (Fig. 2; 26,398 - 27,063) encodes a predicted protein of 221 amino acids. BLAST and FASTA analysis of the protein revealed significant matches to a large number of coronaviral matrix glycoproteins. The association of the spike glycoprotein (S) with the matrix glycoprotein (M) is an essential step in the formation of the viral envelope and in the accumulation of both proteins at the site of virus assembly (*4*). Analysis of the amino acid sequence with SignalP predicts a signal sequence (probability 0.932) that is not likely cleaved. TMHMM and TMpred analysis both indicate the presence of three trans-membrane helices, located at approximately residues 15-37, 50-72 and 77-99, with the 121 amino acid hydrophilic domain on the inside of the virus particle, where it is believed to interact with the nucleocapsid. PFAM analysis reveals a match to PFAM domain PF01635, and alignments to 85 other sequences in the PFAM database bearing this domain, which is indicative of the coronavirus matrix glycoprotein.

Orf7 (Fig. 2; 27,074-27,265) encodes a predicted protein of 63 amino acids. BLAST and FASTA searches yield no significant matches indicative of function. TMHMM and SignalP predict no transmembrane region, however, TMpred analysis predicts a likely trans-membrane helix located between residues 3 and 22, with the N-terminus located outside the viral particle. Similarly, the gene encoding Orf8 (Fig. 2; 27,273-27,641) encoding a predicted protein of 122 amino acids, has no significant BLAST or FASTA matches to known proteins. Analysis of this sequence with SignalP indicates a cleaved signal sequence (probability 0.995), with the predicted cleavage site located between residues 15 and 16. TMpred and TMHMM analysis also predicts a trans-membrane helix located approximately at residues 99-117. Together these data indicate that Orf7 is likely to be a type I membrane protein with the major hydrophilic domain of the protein (residues 16-98) and the amino-terminus are oriented inside the lumen of the ER/Golgi, or on the surface of the cell membrane or virus particle, depending on the membrane localization of the protein.

Orf9 (Fig. 2; 27,638-27,772) encodes a predicted protein of 44 amino acids. FASTA analysis of this sequence revealed some weak similarities (37% identity over a 35 amino acid overlap) to Swiss-Prot accession Q9M883, annotated as a putative sterol-C5 desaturase. A similarly weak match to a hypothetical *Clostridium perfringens* protein (Swiss-Prot accession CPE2366) was also detected. The functional implications, if any, of these matches are unknown. TMpred predicted the existence of a single strong trans-membrane helix, with little preference for alternate models in which the N-terminus was located inside or outside the particle. Similarly Orf10 (Fig. 2; 27,779-27,898) encoding a predicted protein of 39 amino acids, exhibited no significant matches in BLAST and FASTA searches but was predicted to encode a trans-membrane helix by TMPred, with the N-terminus located within the viral particle. The region immediately upstream of Orf10 exhibits a strong match to the TRS consensus (Table 2), providing support for the notion that a transcript initiates from this site. Orf11 (Fig. 2; 27,864-28118) encodes a predicted protein of 84 amino acids exhibited only very short (9-10 residues) matches to a region of the human coronavirus E2 glycoprotein precursor (starting at residue 801). Analysis by SignalP and TMHMM predict a soluble protein. As was the case for Orf 10, a detectable alignment to the TRS consensus sequence was found (Table 2).

The protein (422 amino acids) encoded by the Nucleocapsid gene (Fig. 2; 28,120-29,388) aligns well with nucleocapsid proteins from other representative coronaviruses, although a short lysine rich region (KTFPPTEPKKDKKKKTDEAQ) (*30*) appears to be unique to SARS. This region is suggestive of a nuclear localization signal, and while it contains a hit to InterProDomain IPR001472 (bipartite nuclear localization signal), the function of this insertion remains unknown. It is possible that the SARS virus nucleocapsid protein has a novel nuclear function, which could play a role in pathogenesis. In addition, the basic nature of this peptide suggests it may assist in RNA binding.

Orf 13 (Fig. 2; 28,130 – 28,426) encodes a predicted protein of 98 amino acids. BLAST analysis failed to identify

similar sequences, and no transmembrane helices are predicted. Orf 14 (Fig. 2; 28,583 – 28,795) encodes a predicted protein of 70 amino acids. BLAST analysis failed to identify similar sequences. TMPred predicts a single transmembrane helix.

Conclusions. We have determined through genome sequencing that the virus named by the WHO as causally associated with SARS is a novel coronavirus. This has been confirmed by the sequence of two independent isolates, the Tor2 isolate reported here and the Urbani isolate, reported by the CDC (*14*). Although morphologically a coronavirus (*2*), this SARS virus is not more closely related to any of the three known classes of coronavirus, and we propose that it defines a fourth class of coronavirus (Group 4) and that it be referred to as SARS-CoV. Our sequence data do not support a recent inter-viral recombination event between the known coronavirus groups in the etiology of this virus, but this may be due to the limited number of known coronavirus genome sequences. Apart from the s2m motif located in the 3´UTR, there is also no evidence of any exchange of genetic material between the SARS virus and non-Coronaviridae. These data are consistent with the hypothesis that an animal virus for which the normal host is currently unknown recently mutated and developed the ability to productively infect humans. There also remains the possibility that the SARS virus evolved from a previously harmless human coronavirus. However, preliminary evidence suggests antibodies to this virus are absent in those not infected with SARS-CoV (*2*), implying a benign virus closely related to the Tor2 isolate is not resident in humans. Identification of the normal host of this coronavirus and comparison of the sequences of the ancestral and SARS forms will further elucidate the process by which this virus arose.

Availability of the SARS virus genome sequence is important from a public health perspective. It will allow the rapid development of PCR-based assays for this virus that capitalize on novel sequence features allowing the discrimination between this and other circulating coronaviruses. Such assays will allow the diagnosis of SARS virus infection in humans and, critically, will consolidate the association of this virus with SARS. If the association is further borne out, SARS virus genome-based PCR assays may form an important part of a public health strategy to control the spread of this syndrome. In the longer term, this information will assist in the development of antiviral treatments including neutralizing antibodies and development of a vaccine to treat this emerging and deadly new disease.

References and Notes

1. J. S. M. Peiris *et al*., *Lancet*, published online 8 April 2003 (http://image.thelancet.com/extras/03art3477web.pdf).
2. T. G. Ksiazek *et al*., *N. Engl. J. Med.*, published online 10 April 2003 (10.1056/NEJMoa030781).
3. R. Munch, *Microbes Infect.* **5**, 69 (2003).
4. B. N. Fields, D. M. Knipe, P. M. Howley, D. E. Griffin, *Fields Virology* (Lippincott Williams & Wilkins, Philadelphia, ed. 4, 2001).
5. M. M. C. Lai, D. Cavanagh, *Adv. Virus Res.* **48**, 1 (1997).
6. S. G. Sawicki, D. L. Sawicki, *Adv. Exp. Med. Biol.* **440**, 215 (1998).
7. D. L. Sawicki *et al*., *J. Gen. Virol.* **82**, 386 (2001).
8. S. G. Sawicki, D. L. Sawicki, *J. Virol.* **64**, 1050 (1990).
9. M. Schaad, R. S. J. Baric, *J. Virol.* **68**, 8169 (1994).
10. P. B. Sethna *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5626 (1989).
11. L. Enjuanes *et al*., in *Virus Taxonomy. Classification and Nomenclature of Viruses,* M. H. V. van Regenmortel *et al*., Eds. (Academic Press, New York, 2000), pp. 835–849.
12. Information on materials and methods is available on *Science* Online.
13. S. M. Poutanen *et al*., *N. Engl. J. Med.*, published online 31 March 2003 (10.1056/NEJMoa030634).
14. P. A. Rota *et al*., *Science*, in press; published online 1 May 2003 (10.1126/science.1085952).
15. C. M. Jonassen, T. O. Jonassen, B. Grinde, *J. Gen. Virol.* **79**, 715 (1998).
16. W. Lapps, B. G. Hogue, D. A. Brian, *Virology* **157**, 47 (1987).
17. R. Krishnan, R. Y. Chang, D. A. Brian, *Virology* **218**, 400 (1996).
18. J. Ziebuhr, E. J. Snijder, A. E. Gorbalenya, *J. Gen. Virol.* **81**, 853 (2000).
19. H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, *Protein Eng.* **10**, 1 (1997).
20. E. L. Sonnhammer, G. von Heijne, A. Krogh, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175 (1998).
21. J. C. Tsai, B. D. Zelus, K. V. Holmes, S. R. Weiss, *J. Virol.* **77**, 841 (2003).
22. S. F. Altschul *et al*., *Nucleic Acids Res.* **25**, 3389 (1997).
23. W. R. Pearson, D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (1988).
24. A. Bateman *et al*., *Nucleic Acids Res.* **30**, 276 (2002).
25. K. Hofman, W. Stoffel, *Biol. Chem. Hoppe-Seyler* **374**, 166 (1993).
26. R. Apweiler *et al*., *Nucleic Acids Res.* **29**, 37 (2001).
27. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
28. J. Ortego *et al*., *J. Virol.* **76**, 11518 (2002).
29. L. Kuo *et al*., paper presented at the annual meeting of the American Society for Virology, Lexington, KY, 20 to 24 July 2002.
30. Abbreviations for amino acids: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
31. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
32. J. Felsenstein, *PHYLIP (Phylogeny Inference Package)* version 3.5c (1993). Distributed by the author, Department of Genetics, University of Washington, Seattle.

33. We would like to thank all the staff at the BCCA Genome Sciences Centre for helping to facilitate the rapid sequencing of the SARS-CoV genome. We would like to acknowledge Raymond Tellier, Hospital for Sick Children, for providing us with information on primer sequences that amplify a 216 base-pair region of the Pol gene. We also thank Ivan Sadowski, Dept. Biochemistry and Molecular Biology and John Hobbs and his staff at the Nucleic Acid and Protein Services Unit at the University of British Columbia for rapid synthesis of PCR primers. We also acknowledge the advice and assistance of Francis Ouellette, University of British Columbia Bioinformatics Centre and the staff at the National Center for Biotechnology Information for rapidly processing and making available our sequence data. We are indebted to anonymous reviewers who provided useful criticisms. The British Columbia Cancer Agency Genome Sciences Centre is supported by the BC Cancer Foundation, Genome Canada/Genome British Columbia, Western Economic Diversification, Canada Foundation for Innovation, BC

Knowledge Development Fund, Canadian Institutes of Health Research, the Michael Smith Foundation for Health Research, and the Natural Sciences and Engineering Research Council of Canada. Clones derived from the SARS virus are available from the Genome Sciences Centre (www.bcgsc.bc.ca).

Supporting Online Material
www.sciencemag.org/cgi/content/full/1085953/DC1
Materials and Methods
References

19 April 2003; accepted 30 April 2003
Published online 1 May 2003; 10.1126/science.1085953
Include this information when citing this paper.

Fig. 1. Phylogenetic analysis of SARS proteins. Unrooted phylogenetic trees were generated by clustalw 1.74 (*31*) using the BLOSUM comparison matrix and a bootstrap analysis of 1000 iterations. Numbers indicate bootstrap replicates supporting each node. Phylogenetic trees were drawn with the Phylip Drawtree program 3.6a3 (*32*). Branch lengths indicate the number of substitutions per residue. Genbank accessions for protein sequences: (**A**) Replicase 1A: BoCoV (Bovine Coronavirus): AAL40396, HCoV-229E (Human Coronavirus): NP_07355, MHV (Mouse Hepatitis Virus): NP_045298, IBV (Avian Infectious bronchitis virus): CAC39113, TGEV (Transmissible Gastroenteritis Virus): NP_058423. (**B**) Membrane Glycoprotein: PHEV (Porcine hemagglutinating encephalomyelitis virus): AAL80035, BoCoV (Bovine Coronavirus): NP_150082, IBV & IBV2 (Avian infectious bronchitis virus): AAF35863 & AAK83027, MHV (Mouse hepatitis virus): AAF36439, TGEV (Transmissible gastroenteritis virus): NP_058427, HCoV-229E & HCoV-OC43 (Human Coronavirus): NP_073555 & AAA45462, FCoV (Feline coronavirus): BAC01160. (**C**) Nucleocapsid: MHV (Mouse hepatitis virus): P18446, BoCoV (Bovine coronavirus): NP_150083, IBV 1 & 2 (Avian infectious bronchitis virus): AAK27162 & NP_040838, FCoV (Feline coronavirus): CAA74230, PTGV (Porcine transmissible gastroenteritis virus): AAM97563, HCoV-229E & HCoV-OC43 (Human coronavirus): NP_073556 & P33469, PHEV (porcine hemagglutinating encephalomyelitis virus): AAL80036, TCV (Turkey coronavirus): AAF23873. (**D**) S (Spike) Protein: BoCoV (Bovine coronavirus): AAL40400, MHV (Mouse hepatitis virus): P11225, HCoV-OC43 & HCoV-229E (Human coronavirus): S44241 & AAK32191, PHEV (Porcine hemagglutinating encephalomyelitis virus): AAL80031, PRCoV (Porcine respiratory coronavirus): AAA46905, PEDV (Porcine epidemic diarrhea virus): CAA80971, CCoV (Canine coronavirus): S41453, FIPV (Feline infectious peritonitis virus): BAA06805, IBV (Avian infectious bronchitis virus): AAO34396.

Fig. 2. Map of the predicted Orfs and s2m motif in the Tor2 SARS virus genome sequence.

Table 1. Nucleotide base differences between the Tor2 sequence and the Urbani sequence [(*14*), www.cdc.gov/ncidod/sars/sequence.htm]. Yellow indicates a base difference resulting in an amino acid change (*30*); X indicates a non-conservative amino acid substitution.

Table 2. The nucleotide position, associated open-reading frame and putative transcription regulatory sequences (see text for details). Numbers in parentheses within the alignment indicate distance to the putative initiating codon. The conserved core sequence is indicated in bold in the putative leader sequence. Contiguous sequences identical to region of the leader sequence containing the core sequence are colored. No putative TRSs were detected for Orfs 4, 13 and 14, although Orf 13 could share the TRS associated with the N protein.

Table 3. Features of the Tor2 genome sequence.

Tor2			Urbani				
Position*	Base	Amino acid	Base	Amino acid		Frame	Protein
7,919	C	A	T	V		1	Replicase 1A
16,622	C	A	T	A		3	Replicase 1B
19,064	A	E	G	E		3	Replicase 1B
19,183	T	V	C	A		3	Replicase 1B
23,220	G	A	T	S	X	3	S (spike) glycoprotein
24,872	T	L	C	L		3	S (spike) glycoprotein
25,298	A	R	G	G	X	2	ORF3
26,857	T	S	C	P	X	1	M protein

*GenBank AY274119.3.

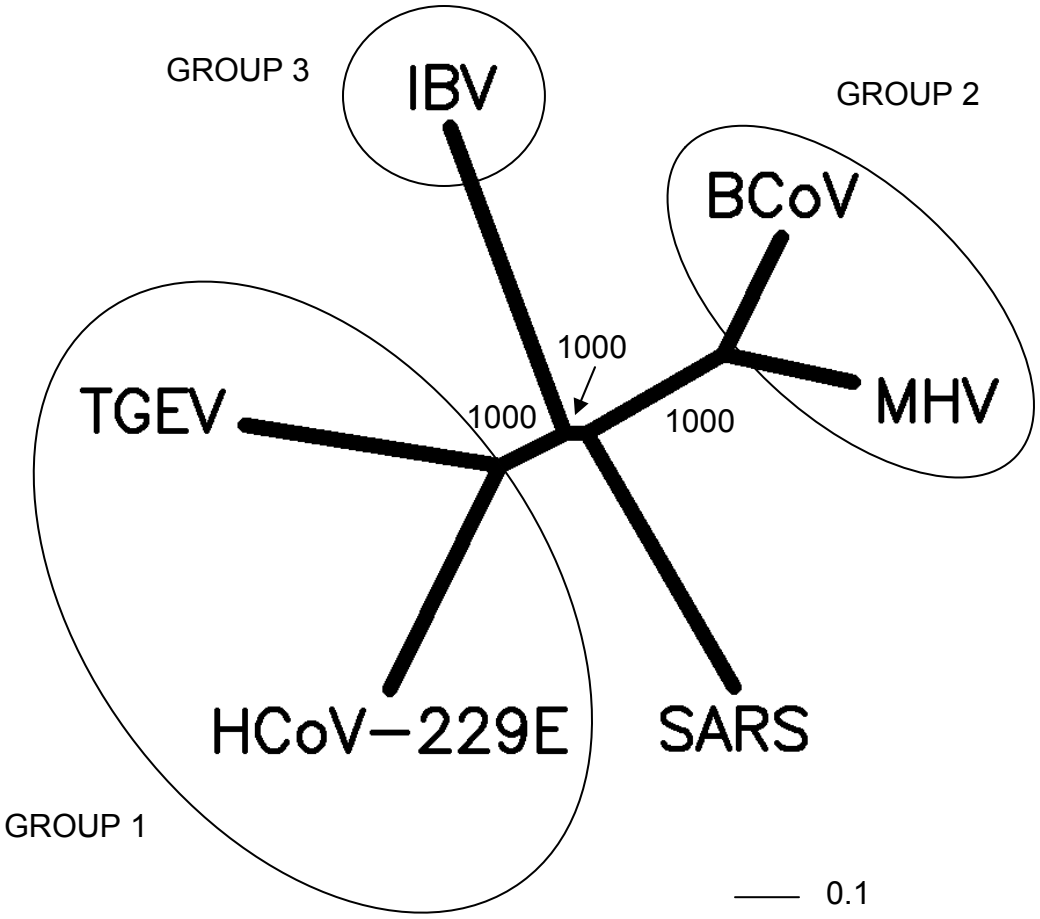
.

Base	ORF	TRS sequence
60	Leader	UCU**CUAAAC**GAACUUUAAAAUCUGUG
21,479	S(Spike)	CAACUAAACGAAC**AUG**
25,252	ORF3	CACAUAAACGAACUU**AUG**
26,104	Envelope	UGAGUACGAACUU**AUG**
26,341	M	GGUCUAAACGAACUAACU (40)**AUG**
27,001	ORF7	AACUAUAAAUU (62)**AUG**
27,259	ORF8	UCCAUAAAACGAAC**AUG**
27,590	ORF9	UGCUCUA---GUAUUUUUAAUACUUUG (24)**AUG**
27,766	ORF10	AGUCUAAACGAAC**AUG**
27,852	ORF11	CUAAUAAAACCUC**AUG**
28,099	Nucleocapsid	UAAAUAAACGAACAAAUUAAA**AUG**

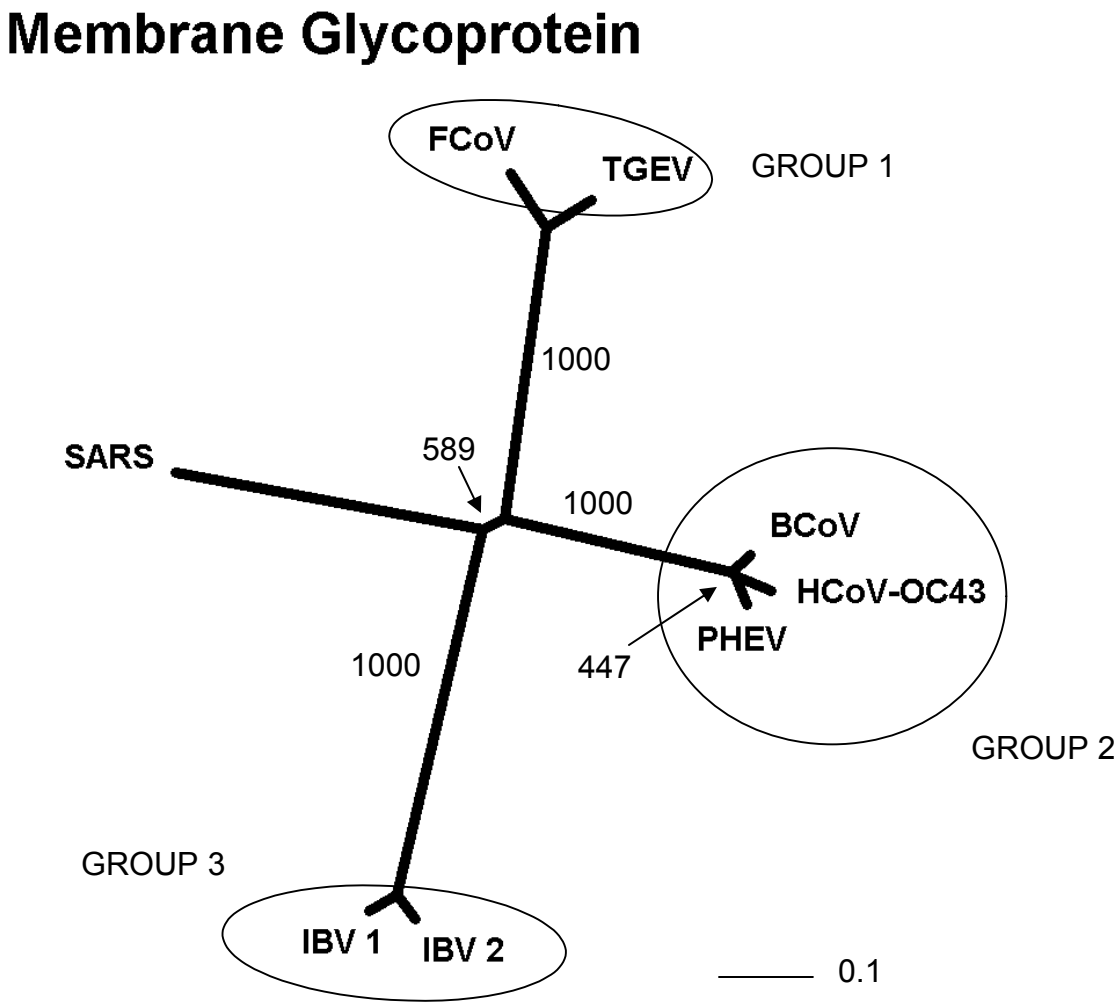
Feature	Start–End*	No. of amino acids	No. of bases	Frame	Candidate TRS†	Rota *et al.* ORF‡
Orf 1a	265–13,398	4,382	13,149	+1	N/A‖	1a
Orf 1b	13,398–21,485	2,628	7,887	+3	N/A	1b
S protein	21,492–25,259	1,255	3,768	+3	Strong	S
Orf 3	25,268–26,092	274	825	+2	Strong	X1
Orf 4	25,689–26,153	154	465	+3	Absent§	X2
E protein	26,117–26,347	76	231	+2	Weak	E
M protein	26,398–27,063	221	666	+1	Strong	M
Orf 7	27,074–27,265	63	192	+2	Weak	X3
Orf 8	27,273–27,641	122	369	+3	Strong	X4
Orf 9	27,638–27,772	44	135	+2	Weak	N/R
Orf 10	27,779–27,898	39	120	+2	Strong	N/R
Orf 11	27,864–28,118	84	255	+3	Weak	X5
N protein	28,120–29,388	422	1,269	+1	Strong	N
Orf 13§	28,130–28,426	98	297	+2	Absent§	N/R
Orf 14§	28,583–28,795	70	213	+2	Absent	N/R
s2m motif	29,590–29,621	N/A	30	N/A	N/A	N/R

*End coordinates include the stop codon, except for ORF 1a and s2m. †See text for details. ‡Corresponding ORFs from Rota *et al.* (*14*). N/R indicates the feature was not reported. §These ORFs overlap substantially or completely with others and may share TRSs. ‖N/A indicates not applicable.

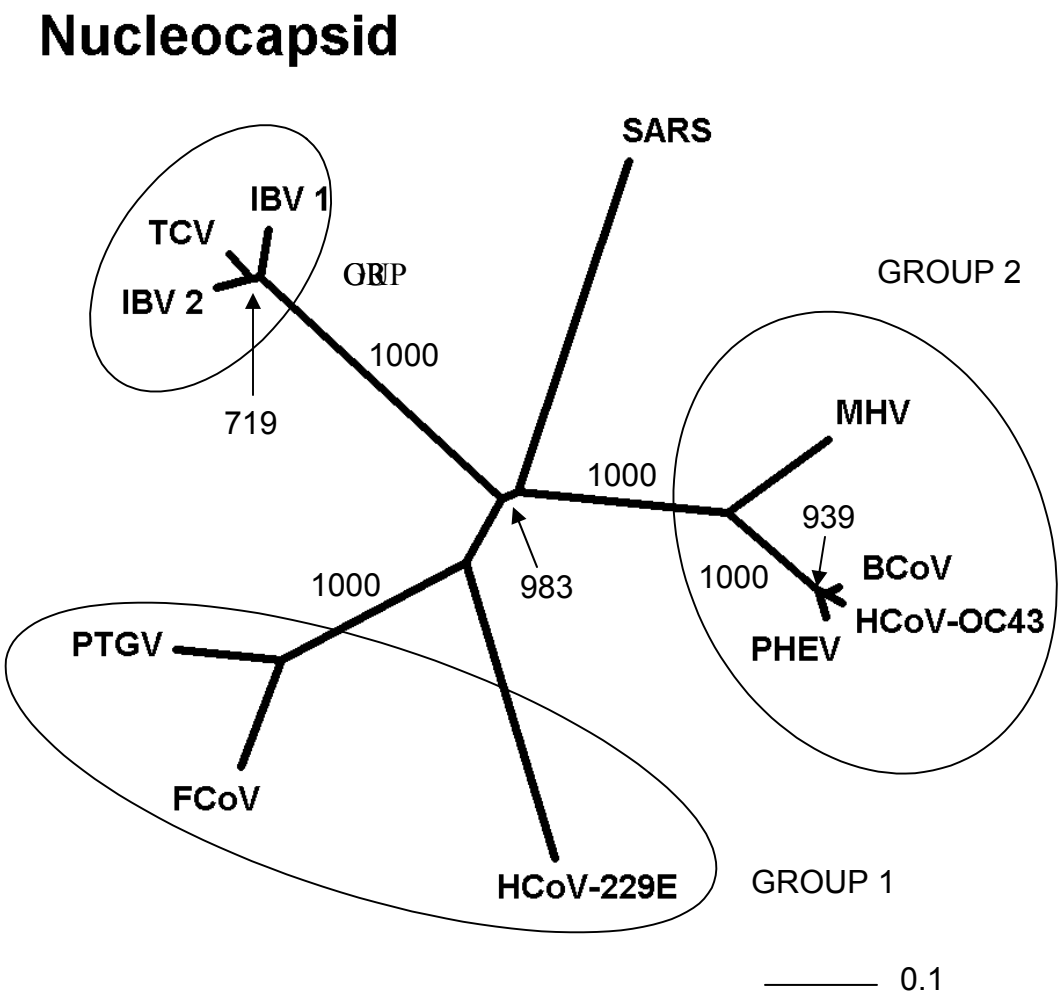
Replicase 1A



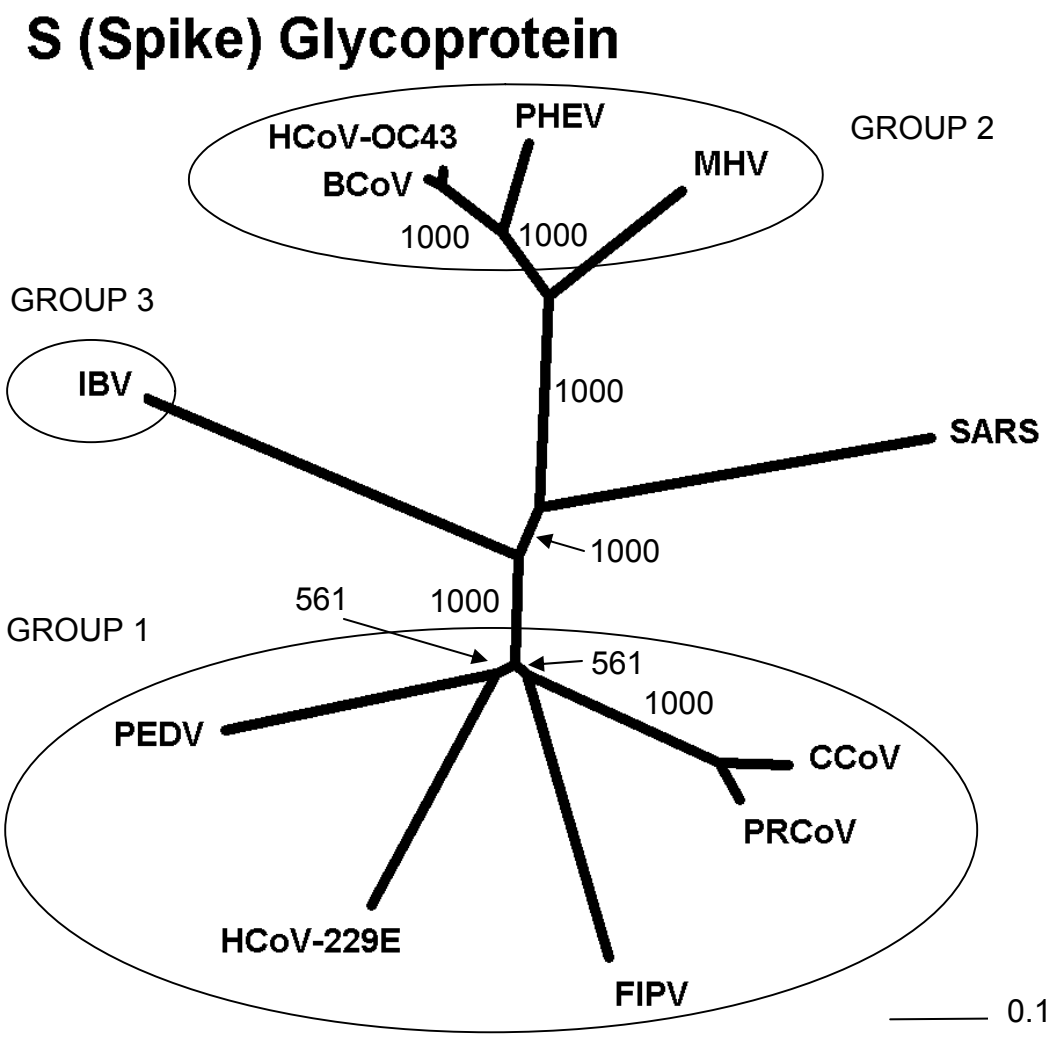
B


Membrane Glycoprotein
FCoV
TGEV
GROUP 1
1000
SARS
589
1000
BCoV
HCoV-OC43
PHEV
1000
447
GROUP 2
GROUP 3
IBV 1
IBV 2
0.1

C


Nucleocapsid
SARS
IBV 1
TCV
IBV 2
GROUP 2
1000
719
MHV
1000
939
1000
983
BCoV
HCoV-OC43
PHEV
1000
PTGV
FCoV
HCoV-229E
GROUP 1
0.1

D


S (Spike) Glycoprotein
HCoV-OC43
PHEV
BCoV
MHV
GROUP 2
1000
1000
GROUP 3
IBV
1000
SARS
1000
561
1000
GROUP 1
561
1000
PEDV
CCoV
PRCoV
HCoV-229E
FIPV
0.1

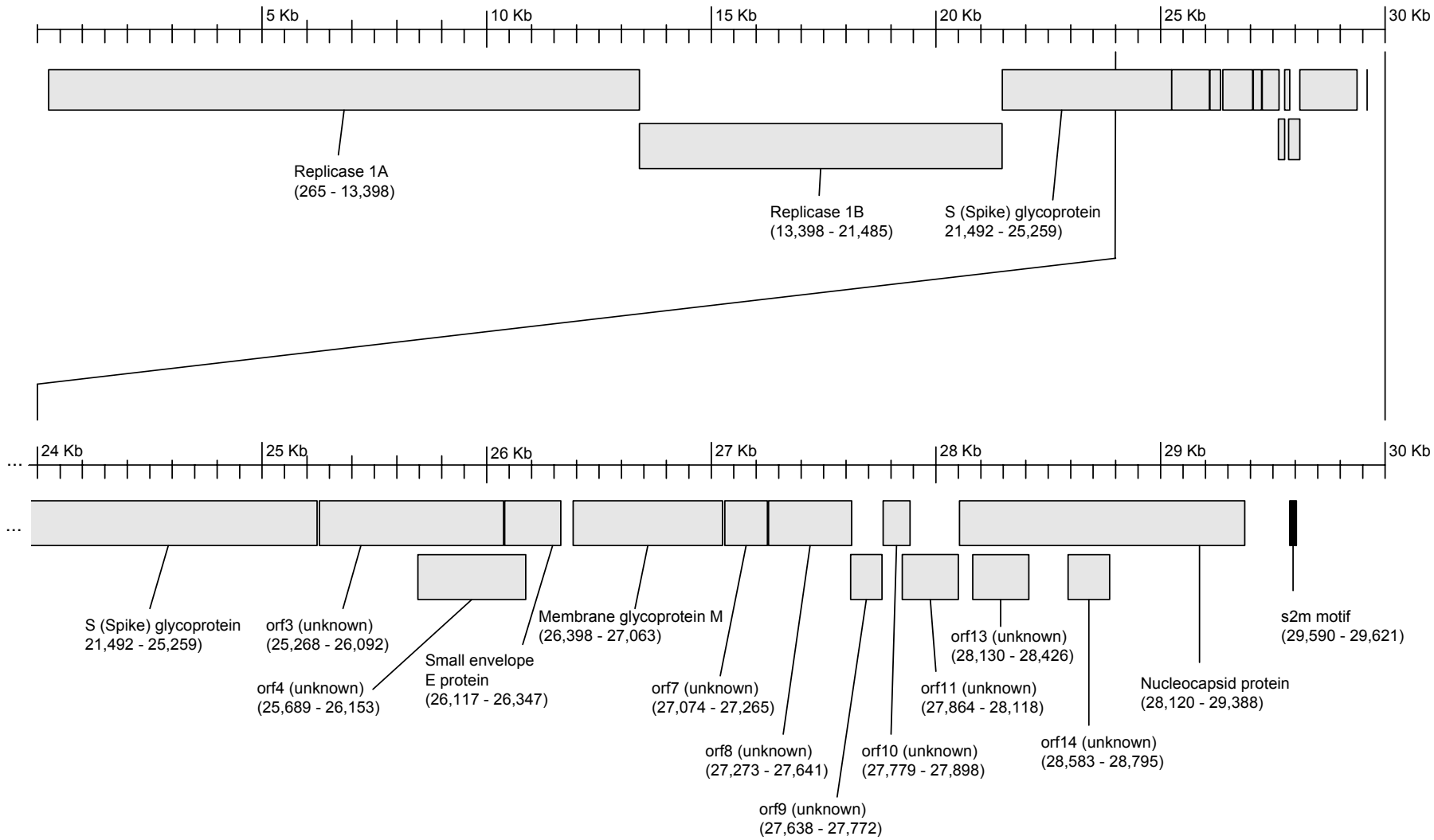

5 Kb
10 Kb
15 Kb
20 Kb
25 Kb
30 Kb
Replicase 1A
(265 - 13,398)
Replicase 1B
(13,398 - 21,485)
S (Spike) glycoprotein
21,492 - 25,259)
24 Kb
25 Kb
26 Kb
27 Kb
28 Kb
29 Kb
30 Kb
S (Spike) glycoprotein
21,492 - 25,259)
orf3 (unknown)
(25,268 - 26,092)
orf4 (unknown)
(25,689 - 26,153)
Small envelope
E protein
(26,117 - 26,347)
Membrane glycoprotein M
(26,398 - 27,063)
orf7 (unknown)
(27,074 - 27,265)
orf8 (unknown)
(27,273 - 27,641)
orf9 (unknown)
(27,638 - 27,772)
orf10 (unknown)
(27,779 - 27,898)
orf11 (unknown)
(27,864 - 28,118)
orf13 (unknown)
(28,130 - 28,426)
orf14 (unknown)
(28,583 - 28,795)
Nucleocapsid protein
(28,120 - 29,388)
s2m motif
(29,590 - 29,621)

Figure 2.