

Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome

Paul A. Rota,^{1*} M. Steven Oberste,¹ Stephan S. Monroe,¹ W. Allan Nix,¹ Ray Campagnoli,¹ Joseph P. Icenogle,¹ Silvia Peñaranda,¹ Bettina Bankamp,¹ Kaija Maher,¹ Min-hsin Chen,¹ Suxiong Tong,¹ Azaibi Tamin,¹ Luis Lowe,¹ Michael Frace,¹ Joseph L. DeRisi,² Qi Chen,¹ David Wang,² Dean D. Erdman,¹ Teresa C. T. Peret,¹ Cara Burns,¹ Thomas G. Ksiazek,¹ Pierre E. Rollin,¹ Anthony Sanchez,¹ Stephanie Liffick,¹ Brian Holloway,¹ Josef Limor,¹ Karen McCaustland,¹ Melissa Olsen-Rasmussen,¹ Ron Fouchier,³ Stephan Günther,⁴ Albert D. M. E. Osterhaus,³ Christian Drosten,⁴ Mark A. Pallansch,¹ Larry J. Anderson,¹ William J. Bellini¹

¹National Center for Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA. ²Departments of Biochemistry and Biophysics, University of California–San Francisco, San Francisco, CA 94143, USA. ³Department of Virology, Erasmus University, Rotterdam, 3000 DR, The Netherlands. ⁴Department of Virology, Bernhard-Nocht Institute for Tropical Medicine, 20359 Hamburg, Germany.

*To whom correspondence should be addressed. E-mail: prota@cdc.gov

In March 2003, a novel coronavirus (SARS-CoV) was discovered in association with cases of severe acute respiratory syndrome (SARS). The sequence of the complete genome of SARS-CoV was determined, and the initial characterization of the viral genome is presented in this report. The genome of SARS-CoV is 29,727 nucleotides in length, has 11 open reading frames, and the genome organization is similar to that of other coronaviruses. Phylogenetic analyses and sequence comparisons showed that SARS-CoV is not closely related to any of the previously characterized coronaviruses.

Several hundred cases of severe, atypical pneumonia of unknown etiology were reported in Guangdong Province of the People's Republic of China beginning in late 2002. After similar cases were detected in patients in Hong Kong, Vietnam, and Canada during February and March 2003, the World Health Organization (WHO) issued a global alert for the illness, designated "severe acute respiratory syndrome" (SARS). In mid-March 2003, SARS was recognized in healthcare workers and household members who had cared for patients with severe respiratory illness in Hong Kong and Vietnam. Many of these cases could be traced through multiple chains of transmission to a healthcare worker from Guangdong Province who visited Hong Kong, where he was hospitalized with pneumonia and died. By late April 2003, over 4300 SARS cases and 250 SARS-related deaths were reported to WHO from over 25 countries around the world. Most of these cases occurred after exposure to SARS patients in household or healthcare settings. The incubation period for the disease is usually from 2 to 7 days. Infection is usually characterized by fever, which is followed a few days later by a dry, non-productive cough, and shortness of breath. Death

from progressive respiratory failure occurs in about 3% to nearly 10% of cases (1–4).

In response to this outbreak, WHO coordinated an international collaboration that included clinical, epidemiologic, and laboratory investigations, and initiated efforts to control the spread of SARS. Attempts to identify the etiology of the SARS outbreak were successful during the third week of March 2003, when laboratories in the United States, Canada, Germany, and Hong Kong isolated a novel coronavirus (SARS-CoV) from SARS patients. Unlike other human coronaviruses, it was possible to isolate SARS-CoV in Vero cells. Evidence of SARS-CoV infection has now been documented in SARS patients throughout the world. SARS-CoV RNA has frequently been detected in respiratory specimens, and convalescent-phase serum specimens from SARS patients contain antibodies that react with SARS-CoV. There is strong evidence that this new virus is etiologically linked to the outbreak of SARS (5–7).

The coronaviruses (order *Nidovirales*, family *Coronaviridae*, genus *Coronavirus*) are a diverse group of large, enveloped, positive-stranded RNA viruses that cause respiratory and enteric diseases in humans and other animals. At ~30,000 nucleotides, their genome is the largest found in any of the RNA viruses. There are three groups of coronaviruses; groups 1 and 2 contain mammalian viruses, while group 3 contains only avian viruses. Within each group, coronaviruses are classified into distinct species by host range, antigenic relationships, and genomic organization. Coronaviruses typically have narrow host ranges and are fastidious in cell culture. The viruses can cause severe disease in many animals, and several viruses, including infectious bronchitis virus, feline infectious peritonitis virus, and transmissible gastroenteritis virus, are significant veterinary

pathogens. Human coronaviruses (HCoVs) are found in both group 1 (HCoV-229E) and group 2 (HCoV-OC43) and are responsible for ~30% of mild upper respiratory tract illnesses (8–10).

Sequence analysis of a limited region of the replicase (rep) gene suggested that SARS-CoV was distinct from all other coronaviruses (5, 7). In this report, we compared the sequence of the entire genome of SARS-CoV (Urbani strain) to the genomic sequences of other coronaviruses.

Genome organization. The sequence of the entire genome of SARS-CoV (GenBank accession number AY278741) was obtained by several approaches (11). During completion of this manuscript, other laboratories determined the genomic sequences of three additional strains of SARS-CoV. These nucleotide sequences vary at only 24 positions (table S3).

The genome of SARS-CoV is a 29,727-nucleotide, polyadenylated RNA, and 41% of the residues are G or C (range for published coronavirus complete genome sequences, 37% to 42%). The genomic organization is typical of coronaviruses, with the characteristic gene order (5′-replicase [rep], spike [S], envelope [E], membrane [M], nucleocapsid [N]-3′) and short untranslated regions at both termini (Fig. 1A, table S1). The SARS-CoV rep gene, which comprises approximately two-thirds of the genome, is predicted to encode two polypeptides (encoded by ORF1a and ORF1b) that undergo co-translational proteolytic processing. There are four open reading frames (ORFs) downstream of rep that are predicted to encode the structural proteins, S, E, M, and N, which are common to all known coronaviruses. The hemagglutinin-esterase gene, which is present between ORF1b and S in group 2 and some group 3 coronaviruses (8), was not found.

Coronaviruses also encode a number of non-structural proteins that are located between S and E, between M and N, or downstream of N. These non-structural proteins, which vary widely among the different coronavirus species, are of unknown function and are dispensable for virus replication (8). The genome of SARS-CoV contains ORFs for five potential non-structural proteins of greater than 50 amino acids in these intergenic regions (Fig. 1B, Table 1, table S1). Two overlapping ORFs encoding predicted proteins of 274 and 154 amino acids (termed X1 and X2, respectively) are located between S and E. Three additional potential non-structural genes, X3, X4, and X5 (encoding proteins of 63, 122, and 84 amino acids, respectively), are located between M and N. In addition to the five ORFs encoding predicted nonstructural proteins described above, there are also two smaller ORFs between M and N, encoding predicted proteins of less than 50 amino acids (Table 1). Searches of the GenBank database (BLAST and FastA) indicated that there is no significant sequence similarity between these potential non-structural proteins of SARS-CoV and any other proteins

(12). Note that there are ORFs encoding predicted proteins of greater than 50 amino acids in the structural genes of SARS-CoV (e.g. N, S, Rep). Many short ORFs are present in the structural genes. They are unlikely to be expressed and, for simplicity, they are not shown in Fig. 1.

The coronavirus rep gene products are translated from genomic RNA, but the remaining viral proteins are translated from subgenomic mRNAs that form a 3′-coterminal nested set, each with a 5′-end derived from the genomic 5′-leader sequence. The coronavirus subgenomic mRNAs are synthesized through a discontinuous transcription process, the mechanism of which has not been unequivocally established (8, 13). The SARS-CoV leader sequence was mapped by comparing the sequence of 5′-RACE (rapid amplification of cDNA ends) (11) products synthesized from the N gene mRNA with those synthesized from genomic RNA. A sequence, AAACGAAC (genomic nucleotides 65-72), was identified immediately upstream of the site where the N gene mRNA and genomic sequences diverged. This sequence was also present upstream of ORF1a and immediately upstream of five other ORFs (Fig. 1, A and B, table S1), suggesting that it functions as the conserved core of the transcription regulating sequences (TRS). The nucleotides required for TRS function must be identified experimentally.

The favored model for production of subgenomic mRNAs of coronaviruses proposes that discontinuous transcription occurs during synthesis of the negative strand (13). Subgenomic negative strands containing a complementary copy of the leader sequence at their 3′ termini serve as templates for synthesis of subgenomic mRNAs. In addition to the site the leader, the TRS conserved core sequence appears six times in the remainder of the genome. The positions of the TRS in the genome of SARS-CoV predict that subgenomic mRNAs of 8.3, 4.5, 3.4, 2.5, 2.0, and 1.7 kb, not including the poly(A) tail, should be produced (Fig. 1, A and B, table S1). At least five subgenomic mRNAs were detected by Northern hybridization of RNA from SARS-CoV-infected cells, using a probe derived from the 3′-untranslated region (Fig. 1C). The calculated sizes of the five predominant bands correspond to the sizes of five of the predicted subgenomic mRNAs of SARS-CoV; we cannot exclude the possibility that other, low-abundance mRNAs are present. Full-length genomic RNA was not detected, likely because it is the least prevalent viral RNA in infected cells (8). The predicted 2.0-kb transcript was also not detected, suggesting that the consensus TRS at nt 27,771-27,778 is not used or that it is a low-abundance mRNA. By analogy with other coronaviruses (8), the 8.3-kb and 1.7-kb subgenomic mRNAs are predicted to be monocistronic, directing translation of S and N, respectively, whereas multiple proteins could be translated from the 4.5-kb (X1, X2, and E), 3.4-kb (M and X3), and 2.5-kb (X4 and X5) mRNAs. A consensus TRS is not found

directly upstream of the ORF encoding the predicted E protein (14), and a monocistronic mRNA that would be predicted to code for E could not be clearly identified by Northern blot analysis. It is possible that the 3.4-kb band contains more than one mRNA species that are not resolved in the gel or that the monocistronic mRNA for E is a low-abundance message. Also, in some coronaviruses, the E protein is translated from the second ORF on a polycistronic mRNA (15, 16).

Phylogenetic analyses of the sequence of SARS-CoV.

To determine the relationship between SARS-CoV and the previously characterized coronaviruses, we compared the predicted amino acid sequences for three well-defined enzymatic proteins encoded by the rep gene and the four major structural proteins of SARS-CoV with those from representative viruses for each of the species of coronavirus for which complete genomic sequence information was available (Fig. 2). The topologies of the resulting phylograms are remarkably similar (Fig. 2A). For each protein analyzed, the species formed monophyletic clusters consistent with the established taxonomic groups. In all cases, SARS-CoV sequences segregated into a fourth, well-resolved branch. These clusters were supported by bootstrap values above 90% (1000 replicates, data not shown). Consistent with pairwise comparisons between the previously characterized coronavirus species (Fig. 2B), there was greater sequence conservation in the enzymatic proteins (3CL^{pro}, polymerase [POL], and helicase [HEL]) than among the structural proteins (S, E, M, and N). These results indicate that SARS-CoV is not closely related to any of the previously characterized coronaviruses and forms a distinct group within the genus *Coronavirus*. The SARS-CoV is approximately equidistant from all previously characterized coronaviruses, just as the existing groups are from one another. Detailed pairwise comparison by dot plot analysis identified many regions of amino acid conservation within each protein (fig. S1), but the overall level of similarity between SARS-CoV and the other coronaviruses was low (Fig. 2B). No evidence for recombination was detected when the predicted protein sequences were analyzed using SimPlot (16a, data not shown).

Predicted replicase gene products of SARS-CoV.

Coronaviruses encode a chymotrypsin-like protease, 3CL^{pro}, that is analogous to the main picornaviral protease, 3C^{pro} (17). They also encode one (group 3) or two (groups 1 and 2) papain-like proteases, termed PLP1^{pro} and PLP2^{pro}, which are analogous to the foot and mouth disease virus leader protease, L^{pro}. Overall, gene products of ORF1a are poorly conserved among different coronaviruses, except for these protease sequences (fig. S1). The predicted gene product of ORF1a of SARS-CoV appears to contain only one PLP^{pro} domain at amino acids 1632 to 1847. The 3CL^{pro} catalytic histidine and

cysteine residues are fully conserved among all coronaviruses (SARS-CoV amino acids His₃₂₈₁ and Cys₃₃₈₅), but coronaviruses appear to lack the conserved catalytic acidic residue that is characteristic of other 3C-like proteases (17). The coronavirus replicase polyprotein is synthesized by a -1 ribosomal frameshift at a conserved "slippery" site (UUUAAAC) immediately upstream of a pseudoknot structure in the overlap of ORFs1a and 1b. This polyprotein is autocatalytically processed to yield the mature viral proteases (PLP^{pro} and 3CL^{pro}), the RNA-dependent RNA polymerase (POL), the RNA helicase (HEL), and other proteins whose functions have not been well characterized. The predicted ribosomal frameshift at the SARS-CoV slippery site (nt 13,392-13,398) would result in translation of 7073-amino acids from a single start site.

Analysis of the predicted structural proteins of SARS-CoV. The structural proteins of coronaviruses (S, E, M, and N) function during host cell entry and virion morphogenesis and release (18). During virion assembly, N binds to a defined packaging signal on viral RNA, leading to the formation of the helical nucleocapsid. M is localized at specialized intracellular membrane structures, and interactions between the M and E proteins and nucleocapsids result in budding through the membrane. In some group 2 coronaviruses, the carboxyl terminus of M interacts with the nucleocapsid to form a core structure (19). The S protein is incorporated into the viral envelope, again by interaction with M, and mature virions are released from smooth vesicles (20). Bands corresponding to the predicted N and S proteins of SARS-CoV were visible in preparations of purified virions that were analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis; however, the assignment of other proteins in virions awaits the availability of specific antibodies to identify these viral proteins (fig. S4).

The S proteins of coronaviruses are large, type I membrane glycoproteins that are responsible for both binding to receptors on host cells and for membrane fusion. The S proteins of some coronaviruses are cleaved into S1 and S2 subunits. S proteins also contain important virus-neutralizing epitopes, and amino acid changes in the S proteins can dramatically affect the virulence and in vitro host cell tropism of the virus (21, 22). Because of the low level of similarity (20-27% pairwise amino acid identity) between the predicted amino acid sequence of the S protein of SARS-CoV and the S proteins of other coronaviruses (Fig. 2B, fig. S1A), comparison of primary amino acid sequences does not provide insight into the receptor binding specificity or antigenic properties of SARS-CoV.

The S protein of SARS-CoV has 23 potential N-linked glycosylation sites (table S2). Functional motifs at the amino and carboxyl termini of the S protein that are conserved among the coronaviruses are also present in the predicted

SARS-CoV S protein, although the S2 domain is more conserved than the S1 domain. The amino terminus of the SARS-CoV S protein contains a short type I signal sequence composed of hydrophobic amino acids that are presumably removed during cotranslational transport through the endoplasmic reticulum. The carboxyl terminus, consisting of a transmembrane domain and a cytoplasmic tail rich in cysteine residues, is highly conserved in SARS-CoV (Fig. 3). At 52 amino acids in length, the SARS-CoV S protein is predicted to have the shortest transmembrane domain and cytoplasmic tail of any coronavirus analyzed (Fig. 3) (range, 61-74 amino acids).

The current paradigm of protein-mediated membrane fusion proposes the collapse of alpha-amphipathic regions in the carboxyl half of the coronavirus S protein into coiled-coils, thus bringing a fusion peptide toward the transmembrane domain, resulting in cellular and viral membrane fusion. Two or three alpha-amphipathic regions are predicted for the carboxyl half of coronavirus S proteins. An alpha-amphipathic region of 116 amino acids was predicted with high confidence at positions 884-999 of the SARS-CoV S protein (fig. S2). Syncytia formation, however, is not a prominent feature of SARS-CoV infection of Vero cells (5). The SARS-CoV S protein lacks the basic amino acid cleavage site found in group 2 and group 3 coronaviruses (23), suggesting that the SARS-CoV S protein is probably not cleaved into S1 and S2 subunits.

While overall sequence conservation is low (Fig. 2B), the predicted E, M, and N proteins of SARS-CoV contain conserved motifs that are found in other coronaviruses. Consistent with the E proteins of other coronaviruses, the predicted E protein of the SARS-CoV contains a hydrophobic domain (residues 12-37) flanked by charged residues and followed by a cysteine-rich region. The amino terminal domains of coronavirus M proteins are exposed on the viral surface, while the carboxyl terminus is inside the viral membrane. Most coronavirus M proteins, including the predicted M protein of SARS-CoV, contain three hydrophobic transmembrane domains in the amino-terminal half of the protein, although some viruses have four. A highly conserved amino acid sequence, SwWSFNPE, immediately following the third hydrophobic domain is SMWSFNPE in the SARS-CoV M protein. The M proteins of coronaviruses are invariably glycosylated near the amino terminus. Group 1 and group 3 coronaviruses are N-glycosylated, while those of group 2 viruses are O-glycosylated (24, 25). The predicted M protein of SARS-CoV has an NGT at near its amino terminus, suggesting that this protein is N-glycosylated at position 4.

The predicted N protein of SARS-CoV is a highly charged, basic protein of 422 amino acids (range for other coronaviruses, 377 to 454) with seven successive hydrophobic residues near the middle of the protein.

Although the overall amino acid sequence homology among coronavirus N proteins is low (Fig. 2B), a highly conserved motif (FY YLGTGP) occurs in the amino-terminal half of all coronavirus N proteins, including that of SARS-CoV. Other conserved residues occur near this highly conserved motif (fig. S3).

Conclusion. The completion of the genomic sequence of SARS-CoV provides a first look at the molecular characteristics of this virus and clearly demonstrates that this virus has features typical of a coronavirus, while it also has features that distinguish it from all previously sequenced coronaviruses. Relative to other coronaviruses, no significant major genomic rearrangements or any examples of large insertions or deletions in the genes coding for the rep, S, E, M, or N proteins were found. Like some other coronaviruses, SARS-CoV has several small non-structural ORFs that are found between the S and E genes and between the M and N genes. SARS-CoV represents a novel virus that is phylogenetically distinct from other characterized coronaviruses. The genetic distance between the SARS-CoV and any other coronavirus in all gene regions implies that no large part of the SARS-CoV genome was derived from other known viruses. The SARS-CoV genomic sequence does not provide obvious clues concerning the potential animal origins of this pathogen.

The genome of SARS-CoV has several unique features that could be of biological significance. The short anchor of the S protein, the specific number and location of small ORFs, and the presence of only one copy of the PLP^{pro} provide a combination of genetic features that readily differentiate this virus from previously described coronaviruses. Of course, the significance of any of these features remains to be determined experimentally.

Successful control of the global SARS epidemic will require the development of vaccines and antiviral compounds that effectively prevent or treat this disease as well as rapid and sensitive diagnostic tests to monitor its spread. The availability of complete genomic sequences (table S3) (26) of SARS-CoV in just a few weeks after the discovery of the virus should have an immediate impact on disease control efforts by making it possible to develop improved diagnostic tests, vaccines, and antiviral agents. The sequence information will also make it possible to identify the origin and natural reservoir of this virus and to contribute to studies of the immune response to this virus and the pathogenesis of SARS-CoV-related disease. The stage is set for the international scientific community to respond and to rapidly develop the tools to control this emerging infectious disease.

References and Notes

1. S. M. Poutanen *et al.*, *N. Engl. J. Med.* (2003). Available April 17 at <http://nejm.org/earlyrelease/sars.asp#4-2>.

2. N. Lee *et al.*, *N. Engl. J. Med.* (2003). Available April 17 at <http://nejm.org/earlyrelease/sars.asp#4-2>.
3. K. W. Tsang *et al.*, *N. Engl. J. Med.* (2003). Available April 17 at <http://nejm.org/earlyrelease/sars.asp#4-2>.
4. Centers for Disease Control and Prevention. *M.M.W.R.* **52**, 357 (2003).
5. T. G. Ksiazek *et al.*, *N. Engl. J. Med.* **348**, 1947 (2003).
6. J. S. Peiris *et al.*, *Lancet* **361**, 1319 (2003).
7. C. Drosten *et al.*, *N. Engl. J. Med.* (2003). Available April 17 at <http://nejm.org/earlyrelease/sars.asp#4-2>.
8. M. M. C. Lai, K. V. Holmes, in *Fields Virology*, D. M. Knipe, P. M. Howley, Eds. (Lippincott Williams and Wilkins, New York, 4th edition, 2001), Ch. 35.
9. L. Enjuanes *et al.*, in *Virus Taxonomy*, Eds. M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B. Carstens, M. K. Estes, S. M. Lemon, M. A. Mayo, D. J. McGeoch, C. R. Pringle, R. B. Wickner, Eds. (Academic Press, New York, 2000), pp. 835–849.
10. K. V. Holmes, in *Fields Virology*, D. M. Knipe, P. M. Howley, Eds. (Lippincott Williams and Wilkins, New York, 4th edition, 2001), Ch. 36.
11. Materials and methods are available as supporting material on *Science Online*.
12. Although the match was not statistically significant, the carboxy half of potential protein X1 contains a region of similarity with calcium-transporting ATPases.
13. G. S. Sawicki, D. L. Sawicki. *Adv. Exp. Med. Biol.*, **440**, 215 (1998).
14. The sequence immediately upstream of the ORF coding for the predicted E protein is GTACGAAC and differs from the sequence of the consensus TRS at the first two positions.
15. D. X. Liu, S. C. Inglis, *J. Virol.* **66**, 6143 (1992).
16. V. Thiel, S. G. Siddell, *J. Gen. Virol.* **75**, 3041 (1994).
- 16a. K. S. Lole *et al.*, *J. Virol.* **73**, 152 (1999).
17. J. Ziebuhr, E. J. Snijder, A. E. Gorbalenya, *J. Gen. Virol.* **81**, 853 (2000).
18. S. G. Siddell, Ed., *The Coronaviridae* (Plenum Press, New York, 1995).
19. D. Escors, J. Ortego, H. Laude, L. Enjuanes, *J. Virol.* **75**, 1312 (2001).
20. H. Garoff, R. Hewson, D.-J. E. Opstelten, *Microbiol. Mol. Biol. Rev.* **62**, 1171 (1998).
21. C. M. Sanchez *et al.*, *J. Virol.* **73**, 7607 (1999).
22. I. Leparc-Goffart *et al.*, *J. Virol.* **72**, 9628 (1998).
23. Cleavage sites in the S proteins of coronaviruses are RRFRR, RRSRR, RRSRR, RSRR, RARS, RARR in infectious bronchitis virus, bovine coronavirus, human coronavirus OC43, porcine hemagglutinating encephalomyelitis virus, mouse hepatitis virus, and rat coronavirus, respectively.
24. C. A. M. de Haan *et al.*, *Virus Res.* **82**, 77 (2002).
25. C. A. M. de Haan, L. Kuo, P. S. Masters, H. Vennema, P. J. M. Rottier, *J. Virol.* **72**, 6838 (1998).
26. As of this writing, complete genomic sequences of three additional SARS-CoV isolates were available on GenBank (Tor-2 strain, Canada, accession no. ay274119; CUHK-W1 isolate, Hong Kong, accession no. ay278554; HKU-39849 isolate, Hong Kong, accession no. ay278491). A comparison of these sequences to the sequence described in this paper is shown in table S3.
27. The authors thank the WHO SARS Aetiology Laboratory Investigation Group (Bernhard-Nocht Institute, Hamburg, Germany; Erasmus Universiteit, National Influenza Centre, Rotterdam, The Netherlands; Federal Microbiology Laboratories for Health Canada, Winnipeg, Canada; Institut für Virologie, Marburg Germany; Frankfurt A. M. University Hospital, Klinikum der Johann Wolfgang Goethe-Universität, Frankfurt, Germany; Chinese Center for Disease Control, Beijing, China; PHLS Central Public Health Laboratory, London; Prince of Wales Hospital, Hong Kong; National Institute of Infectious Disease, Tokyo, Japan; The Chinese University of Hong Kong, Hong Kong; Government Virus Unit, Hong Kong; Queen Mary Hospital, Hong Kong; and Institute Pasteur, Paris, France) for the open collaboration and sharing of information; CDC Laboratory Partners Group for support and suggestions; the Coronavirology Partners Group (S. C. Baker, R. Baric, D. A. Brian, D. Cavanagh, M. R. Denison, M. S. Diamond, B. G. Hogue, K. V. Holmes, J. Leibowitz, S. Perlman, L. J. Saif, L. Sturman, and S. R. Weiss) for many helpful reagents, guidance and discussion; B. W.J. Mahy for advice, and discussions and for organizing the Laboratory Partners Conferences; S. Emery for technical support; J. Osborne and S. Sammons for help with the figures; and C. Chesley for editorial assistance. M-h.C. is supported by CDC/Georgia State University interagency agreement.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1085952/DC1

Materials and Methods

Figs. S1 to S4

Tables S1 to S3

References

21 April 2003; accepted 30 April 2003

Published online 1 May 2003; 10.1126/science.1085952

Include this information when citing this paper.

Fig. 1. Genome organization and mRNA mapping of SARS-CoV. (A) Overall organization of the 29,727-nt SARS-CoV genomic RNA (arrow). The 72-nt leader sequence is represented as a small square. Predicted open reading frames (ORFs) 1a and 1b, encoding the nonstructural polyproteins,

and those encoding the S, E, M, and N structural proteins are indicated. Vertical position of the boxes indicates the phase of the reading frame. **(B)** Expanded view of the structural protein coding region and predicted mRNA transcripts. Known structural protein coding regions (dark gray boxes) and reading frames X1-X5, encoding potential nonstructural proteins greater than 50 amino acids (light gray boxes), are indicated. Lengths and map locations of the 3'-coterminal mRNAs, as predicted by identification of conserved transcriptional regulatory sequences, are indicated. **(C)** Northern blot analysis of SARS-CoV mRNAs. Poly(A)⁺ RNA was separated on a formaldehyde-agarose gel, transferred to a nylon membrane, and hybridized with a digoxigenin-labeled riboprobe overlapping the 3' untranslated region. Signals were visualized by chemiluminescence. Sizes of the SARS-CoV mRNAs were calculated by interpolation from a log-linear fit of those of the molecular mass marker. Lane 1, SARS-CoV mRNA; lane 2, Vero E6 cell mRNA; lane 3, molecular mass marker, sizes in kB.

Fig. 2. Phylogenetic analysis and pairwise identities of coronavirus proteins. Predicted amino acid sequences of SARS-CoV proteins were compared with those from reference viruses representing each species in the three groups of coronaviruses for which complete genomic sequence information was available (Group 1: human coronavirus 229E [HCoV-229E], af304460; porcine epidemic diarrhea virus [PEDV], af353511; transmissible gastroenteritis virus [TGE], aj271965. Group 2: bovine coronavirus [BCoV], af220295; murine hepatitis virus [MHV], af201929. Group 3: infectious bronchitis virus [IBV], m95169). Sequences for representative strains of other coronavirus species, for which partial sequence information was available, were included for some of the structural protein comparisons (Group 1: canine coronavirus [CCoV], d13096; feline coronavirus [FCoV], ay204704; porcine respiratory coronavirus [PRCoV], z24675. Group 2: human coronavirus OC43 [HCoV-OC43], m76373, 114643, m93390; porcine hemagglutinating encephalomyelitis virus [HEV], ay078417; rat coronavirus [RtCoV], af207551). **(A)** Sequence alignments and neighbor-joining trees were generated by using ClustalX 1.83 with the Gonnet protein comparison matrix. The resulting trees were adjusted for final output using tree2ool 2.0.1. **(B)** Uncorrected pairwise distances were calculated from the aligned sequences by using the Distances program from the Wisconsin Sequence Analysis Package, version 10.2 (Accelrys, Burlington, MA). Distances were converted to percent identity by subtracting from 100.

Fig. 3. Conserved motifs in coronavirus S proteins. Alignment of the carboxy-terminal region of the SARS-CoV and reference coronavirus S proteins was generated using Clustalx 1.83. Residues that match the SARS-CoV sequence

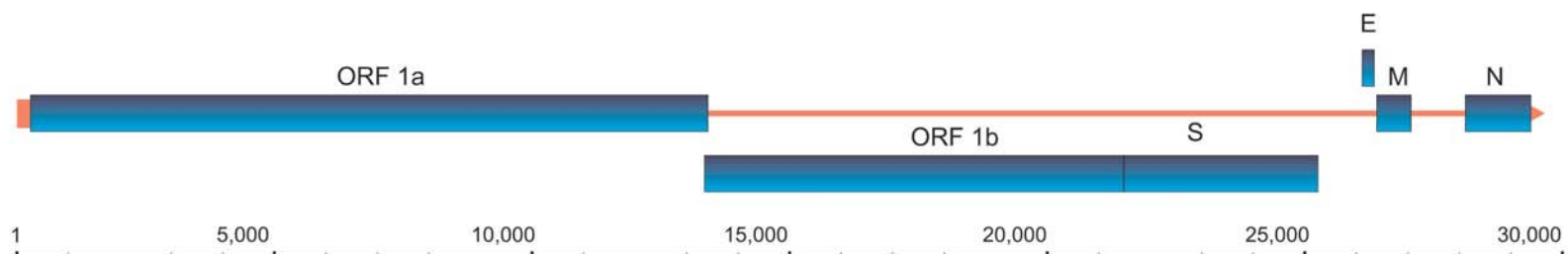
exactly are boxed. The membrane spanning domain and cytoplasmic tails are delineated with arrows. The amino acid sequence, Y(V/I)KWPW(Y/W)VWL, is a conserved motif in all three coronavirus groups. The cysteine-rich region, which overlaps the membrane spanning region and the cytoplasmic region, is also found in all coronavirus groups.

Table 1. Classification of ORFs encoding potential nonstructural proteins of SARS-CoV.*

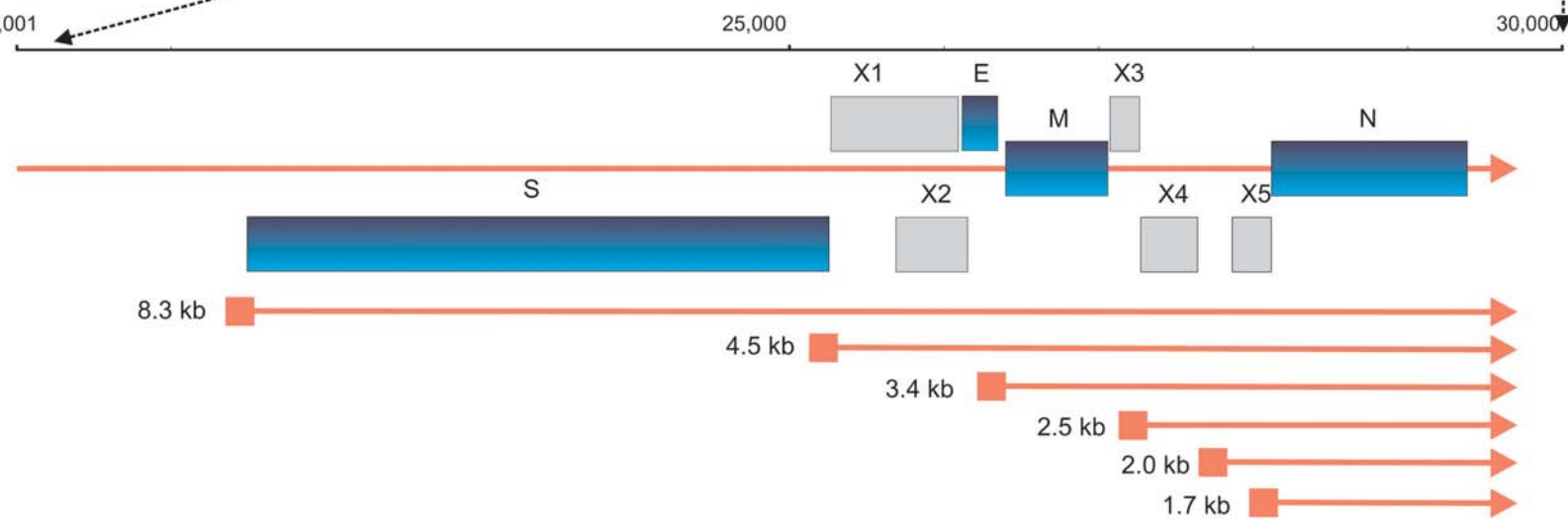
Genome location†	Protein (aa)	This report‡	Marra <i>et al.</i> §
25,268-26,089	274	X1	orf3
25,689-26,150	154	X2	orf4
27,074-27,262	63	X3	orf7
27,273-27,638	122	X4	orf8
27,638-27,769	44	<50 aa	orf9
27,779-27,895	39	<50 aa	orf10
27,864-28,115	84	X5	orf11
28,130-28,423	98	See text	orf13
28,583-28,792	70	See text	orf14

*Table shows the differences in nomenclature used to describe ORFs encoding potential nonstructural proteins of SARS-CoV in this report and the report by Marra *et al.* It should be noted that these differences are in nomenclature only and that the seven nucleotide sequence differences between these strains do not change the position or number of ORFs (table S2). Note that since the complete SARS-CoV sequences have been available for only a few weeks and will likely be analyzed in great detail in the upcoming months, any nomenclature proposed at this time should be considered preliminary. The nomenclature used for the nonstructural proteins (X1 through X5) is expected to be clarified once experiments on the transcriptional expression of the SARS-CoV genome are reported. †Based on the sequence of the Urbani strain of SARS-CoV (GenBank accession no. AY278741.1). ‡In this report, the ORFs encoding the predicted nonstructural proteins are designated as X1 through X5 and numbered sequentially beginning at the 5' terminus of the genome. Only ORFs encoding for predicted proteins of greater than 50 amino acids are included in Fig. 1B. The locations and sizes of the ORFs encoding the predicted replicase protein, structural proteins and nonstructural proteins are shown in table S2. §In Marra *et al.*, all of the ORFs, including those encoding predicted replicase protein and structural proteins, are numbered sequentially from the 5' terminus of the genome. Table shows only those ORFs encoding predicted nonstructural proteins. ||These ORFs overlap the coding region of the N protein.

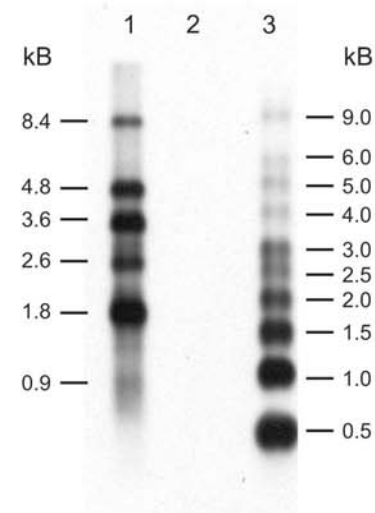
A

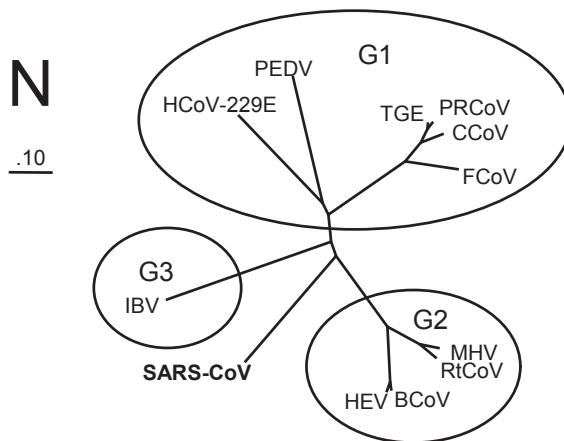
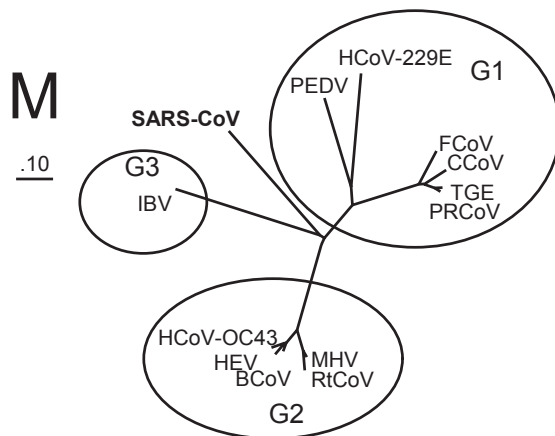
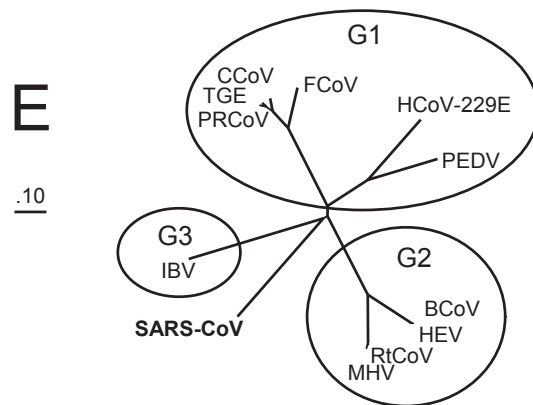
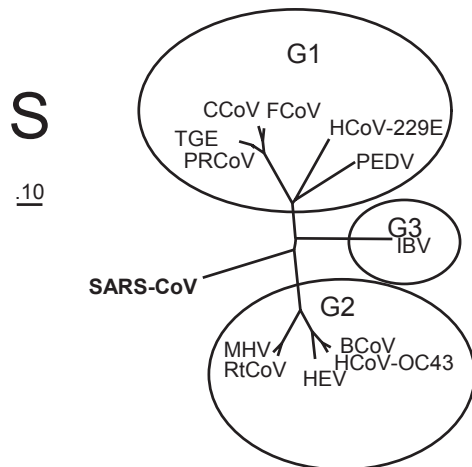
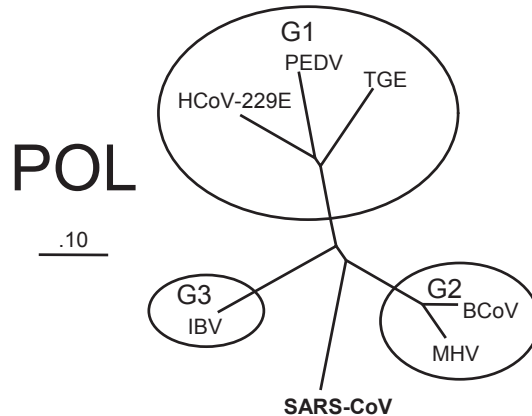
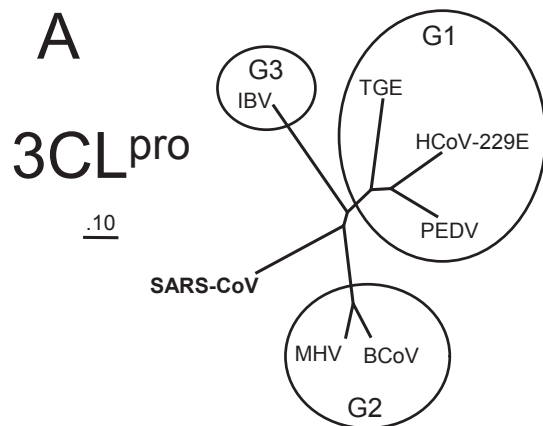


B



C





B

Group	Virus	Pairwise Amino Acid Identity (Percent)						
		3CLPRO	POL	HEL	S	E	M	N
G1	HCoV-229E	40.1	58.8	59.7	23.9	22.7	28.8	23.0
	PEDV	44.4	59.5	61.7	21.7	17.6	31.8	22.6
	TGEV	44.0	59.4	61.2	20.6	22.4	30.0	25.6
G2	BCoV	48.8	66.3	68.3	27.1	20.0	39.7	31.9
	MHV	49.2	66.5	67.3	26.5	21.1	39.0	33.0
G3	IBV	41.3	62.5	58.6	21.8	18.4	27.2	24.0

Virus		Predicted Protein Length (aa)						
SARS-CoV		306	932	601	1255	76	221	422
CoV Range		302-307	923-940	506-600	1173-1452	76-108	225-262	377-454

Membrane Spanning

BCoV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	F	A	G	V	A	M	L	V	L	L	F	F	I	C	C	C	T	G	C	G	T	S	C	F	K	I	C	-	-	-	G	G	C	-
HCoV-OC43	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	C	L	A	G	V	A	M	L	V	L	L	F	F	I	C	C	C	T	G	C	G	T	S	C	F	K	K	C	-	-	-	G	G	C	-
HEV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	A	G	V	A	M	L	V	L	L	F	F	I	C	C	C	T	G	C	G	T	S	C	F	K	K	C	-	-	-	G	G	C	-
MHV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	A	G	V	A	V	C	V	L	L	F	F	I	C	C	C	T	G	C	G	S	C	C	F	K	K	C	-	-	-	G	N	C	-
RtCoV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	A	G	V	A	V	C	V	L	L	F	F	I	C	C	C	T	G	C	G	S	C	C	F	K	K	C	-	-	-	G	N	C	-
SARS-CoV	Y	I	K	W	P	W	Y	V	W	L	G	F	-	I	A	G	L	I	A	I	V	M	V	T	I	L	L	-	-	C	C	M	T	S	C	-	C	S	C	L	K	-	-	-	-	G	A	C	S	
PRCoV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	V	V	I	F	C	I	P	L	L	L	F	C	C	C	S	T	G	C	-	C	G	C	I	G	C	L	-	-	-	G	S	C	-
TGE	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	V	V	I	F	C	I	P	L	L	L	F	C	C	C	S	T	G	C	-	C	G	C	I	G	C	L	-	-	-	G	S	C	-
CCoV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	V	V	I	F	C	I	P	L	L	L	F	C	C	C	S	T	G	C	-	C	G	C	I	G	C	L	-	-	-	G	S	C	-
FCoV	Y	V	K	W	P	W	Y	V	W	L	-	-	-	L	I	G	L	V	V	V	F	C	I	P	L	L	L	F	C	C	F	S	T	G	C	-	C	G	C	I	G	C	L	-	-	-	G	S	C	-
PEDV	Y	I	K	W	P	W	W	V	W	L	-	-	-	I	I	V	I	V	L	I	F	V	V	S	L	L	V	F	C	C	I	S	T	G	C	-	C	G	C	C	G	C	C	-	-	-	G	A	C	-
HCoV-229E	Y	I	K	W	P	W	W	V	W	L	-	-	-	C	I	S	V	V	L	I	F	V	V	S	M	L	L	L	C	C	C	S	T	G	C	-	C	G	F	F	S	C	F	-	-	-	A	S	S	-
IBV	Y	I	K	W	P	W	Y	V	W	L	A	I	A	F	A	T	I	I	F	I	L	I	L	G	W	V	F	F	M	T	G	C	C	G	C	C	C	G	C	F	G	I	M	P	L	M	S	K	C	G

B CoV	C	D	D	-	-	Y	T	G	H	Q	E	L	V	I	K	-	-	-	T	S	-	H	D	D
HCoV-OC43	C	D	D	-	-	Y	T	G	Y	Q	E	L	V	I	K	-	-	-	T	S	-	H	D	D
HEV	C	D	D	-	-	Y	T	G	H	Q	E	F	V	I	K	-	-	-	T	S	-	H	D	D
MHV	C	D	E	-	-	C	G	G	H	Q	D	S	I	V	I	H	N	I	S	S	-	H	E	D
RtCoV	C	D	E	-	-	Y	G	G	R	Q	A	G	I	V	I	H	N	I	S	S	-	H	E	D
SARS-CoV	C	G	S	-	-	C	C	K	F	D	E	-	-	-	-	-	-	-	-	-	-	D	D	
PRCoV	C	H	S	-	-	I	F	S	R	R	Q	F	E	N	Y	E	P	I	E	K	V	H	V	H
TGE	C	H	S	-	-	I	C	S	R	R	Q	F	E	N	Y	E	P	I	E	K	V	H	V	H
CCoV	C	H	S	-	-	I	C	S	R	G	Q	F	E	S	Y	E	P	I	E	K	V	H	V	H
FCoV	C	H	S	-	-	I	C	S	R	R	Q	F	E	N	Y	E	P	I	E	K	V	H	V	H
PEDV	F	S	G	-	-	C	C	R	G	P	R	L	Q	P	Y	E	A	F	E	K	V	H	V	Q
HCoV-229E	I	R	G	-	-	C	C	E	S	T	K	L	-	P	Y	Y	D	V	E	K	I	H	I	Q
IBV	K	K	S	S	Y	Y	T	T	F	D	N	D	V	V	T	E	Q	Y	R	P	K	K	S	V

Cytoplasmic Tail