# 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. Bellini[1]

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 and has 11 open reading frames, and its 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 health care 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 health care 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 health care 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 nonproductive 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*).

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

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

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 (nt), 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, whereas 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 important 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 compare 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 (the range for published complete coronavirus genome sequences is 37 to 42%). The genomic organization is typical of coronaviruses, having the characteristic gene order [5′-*replicase* (*rep*), *spike* (S), *envelope* (E), *membrane* (M), and *nucleocapsid* (N)-3′] and short untranslated regions at both termini (Fig. 1A and table S1). The SARS-CoV *rep* gene, which comprises approximately two-thirds of the genome, is predicted to encode two polyproteins (encoded by ORF1a and ORF1b) that undergo cotranslational 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 gene encoding hemagglutinin-esterase, 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 nonstructural proteins that are located between S

and E, between M and N, or downstream of N. These nonstructural 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 nonstructural proteins that are more than 50 amino acids long in these intergenic regions (Fig. 1B, Table 1, and 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 nonstructural 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 the 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 (with BLAST and FastA) indicated that there is no significant sequence similarity between these potential nonstructural proteins of SARS-CoV and any other proteins (*12*). Note that there are ORFs encoding predicted proteins more than 50 amino acids long in the structural genes of SARS-CoV (such as N, S, and *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 to 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, and table S1), suggesting that it functions as the conserved core of the transcription-regulating sequences (TRSs). 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 at the 5′ terminus of the genome, 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, and 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, probably because it is the least prevalent viral RNA in infected cells (*8*). The predicted 2.0-kb transcript was also not detected, which suggests that the consensus TRS at nt 27,771 to 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 contained more than one mRNA species that were not resolved in the gel or that the monocistronic mRNA for E is a low-abundance message.

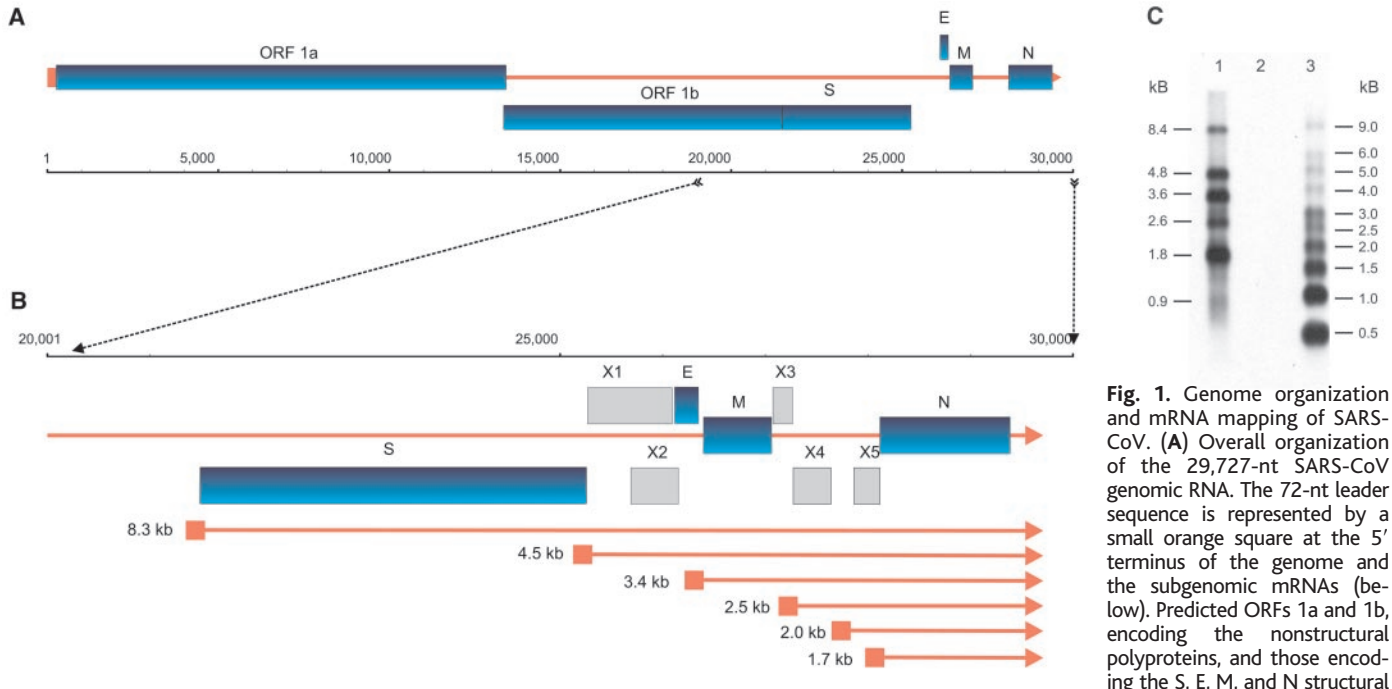
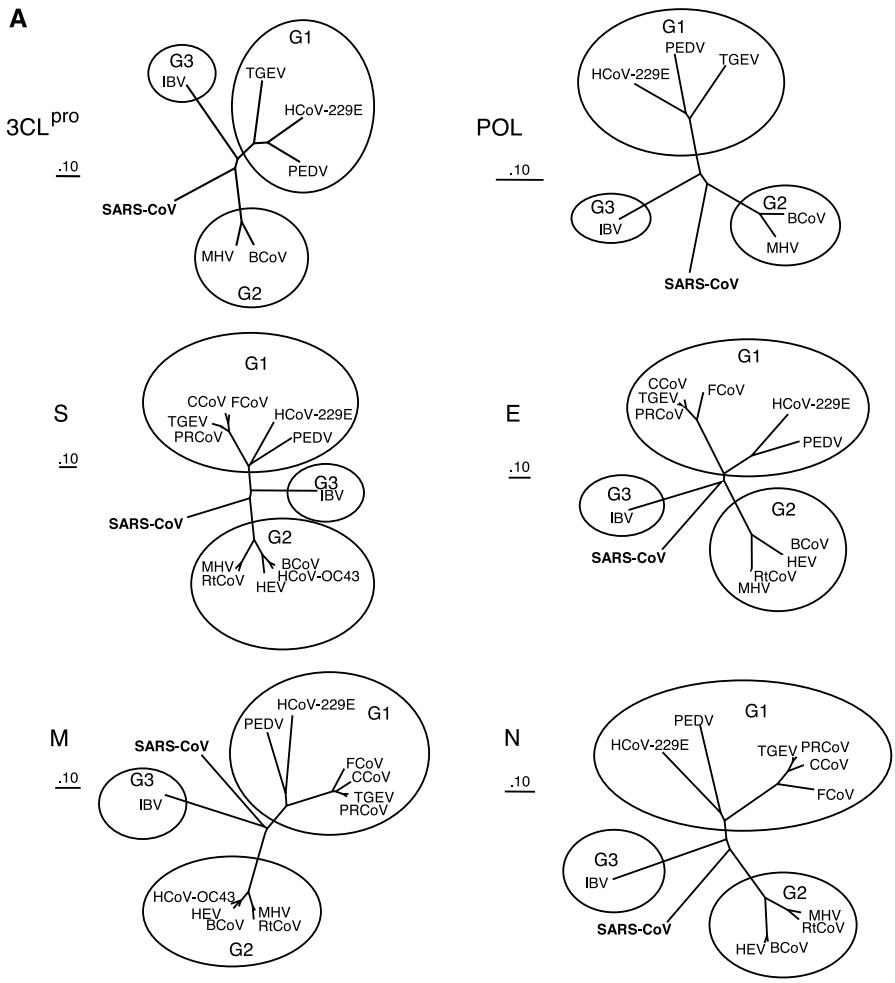



**Fig. 1.** Genome organization and mRNA mapping of SARS-CoV. (**A**) Overall organization of the 29,727-nt SARS-CoV genomic RNA. The 72-nt leader sequence is represented by a small orange square at the 5′ terminus of the genome and the subgenomic mRNAs (below). Predicted ORFs 1a and 1b, encoding the nonstructural polyproteins, and those encoding the S, E, M, and N structural proteins are indicated. The 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 (blue boxes) and reading frames X1 to X5, encoding potential nonstructural proteins longer than 50 amino acids (gray boxes), are indicated. Lengths and map locations of the 3′-coterminal mRNAs, as predicted by identification of conserved transcription-regulating 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 kilobases).




**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 |

**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(G1): human coronavirus 229E (HCoV-229E), af304460; porcine epidemic diarrhea virus (PEDV), af353511; transmissible gastroenteritis virus (TGEV), aj271965. Group 2 (G2): bovine coronavirus (BCoV), af220295; murine hepatitis virus (MHV), af201929. Group 3 (G3): 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, l14643, m93390; porcine hemagglutinating encephalomyelitis virus (HEV), ay078417; rat coronavirus (RtCoV), af207551]. (**A**) Sequence alignments and neighbor-joining trees were generated by the use of ClustalX 1.83 with the Gonnet protein comparison matrix. The resulting trees were adjusted for final output with treetool 2.0.1. (**B**) Uncorrected pairwise distances were calculated from the aligned sequences with 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. aa, amino acid.

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 (*17*)]. 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*. SARS-CoV is approximately equidistant from all previously characterized coronaviruses, just as the existing groups are from one another. Detailed pairwise comparison by dotplot 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 with the program SimPlot (*17*, *18*).

**Predicted replicase gene products of SARS-CoV.** Coronaviruses encode a chymotrypsin-like protease, 3CL$^{pro}$, that is analogous to the main picornaviral protease 3C$^{pro}$ (*19*). 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 His$^{3281}$ and Cys$^{3385}$), but coronaviruses appear to lack the conserved catalytic acidic residue that is characteristic of other 3C-like proteases (*19*). 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 ORF1a and ORF1b. This polyprotein is autocatalytically processed to yield the mature viral proteases $PLP^{pro}$ and $3CL^{pro}$, the RNA-dependent polymerase (POL), the RNA helicase (HEL), and other proteins whose functions have not been well characterized. The predicted ribosomal frame shift at the SARS-CoV slippery site (nt 13,392 to 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 (*20*). 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 C terminus of M interacts with the nucleocapsid to form a core structure (*21*). The S protein is incorporated into the viral envelope, again by interaction with M, and mature virions are released from smooth vesicles (*22*). Bands corresponding to the predicted N and S proteins of SARS-CoV were visible in preparations of purified virions that were analyzed by SDS–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 both for 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 (*23*, *24*). Because of the low level of similarity (20 to 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 and fig. S1A), the 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 (N) and carboxyl (C) 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 N 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 C 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 to 74 amino acids).

The current paradigm of protein-mediated membrane fusion proposes the collapse of alpha-amphipathic regions in the C 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 C half of coronavirus S proteins. An alpha-amphipathic region of 116 amino acids was predicted with high confidence at positions 884 to 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 (*25*), suggesting that the SARS-CoV S protein is probably not cleaved into S1 and S2 subunits.

Although 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 SARS-CoV contains a hydrophobic domain (residues 12 to 37) flanked by charged residues and followed by a cysteine-rich region. The N-terminal domains of coronavirus M proteins are exposed on the viral surface, whereas the C 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 N-terminal half of the protein, although some viruses have four. A highly conserved amino acid sequence [SwWSFNPE (*26*)], immediately following the third hydrophobic domain, is SMWSFNPE in the SARS-CoV M protein. The M proteins of coronaviruses are invariably glycosylated near the N terminus. Group 1 and group 3 coronaviruses are N-glycosylated, whereas those of group 2

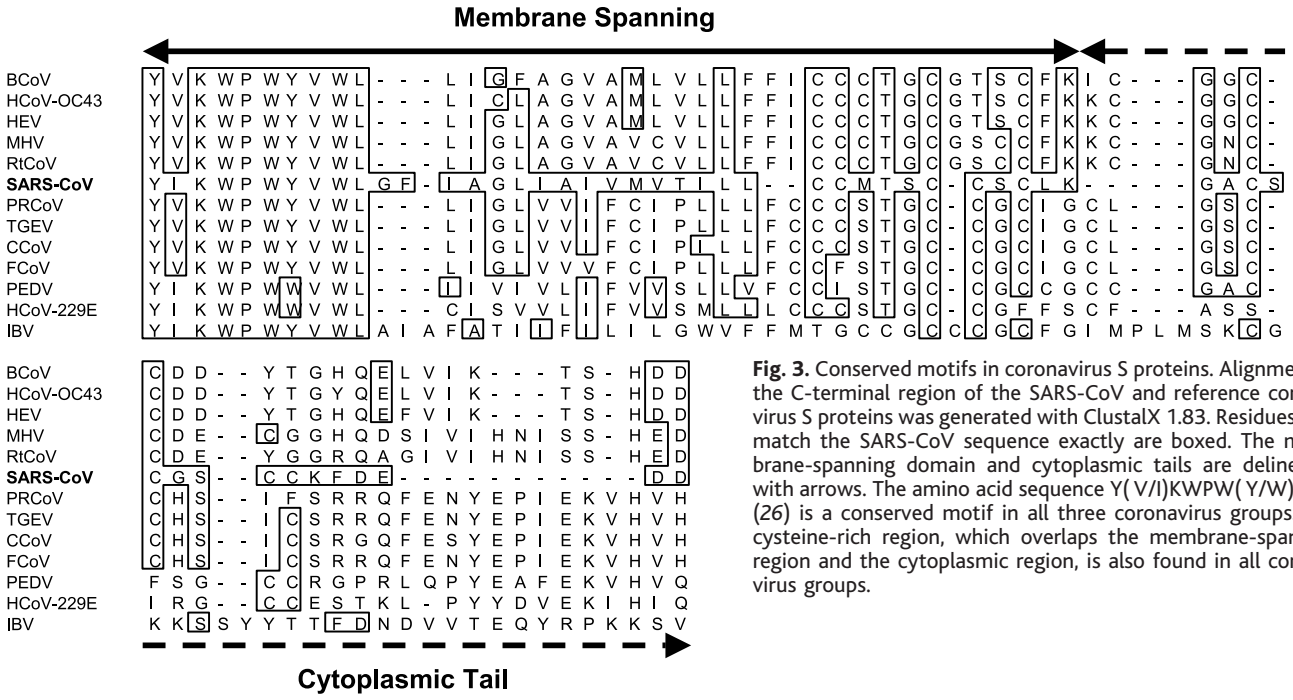



**Fig. 3.** Conserved motifs in coronavirus S proteins. Alignment of the C-terminal region of the SARS-CoV and reference coronavirus S proteins was generated with ClustalX 1.83. Residues that match the SARS-CoV sequence exactly are boxed. The membrane-spanning domain and cytoplasmic tails are delineated with arrows. The amino acid sequence Y( V/I)KWPW( Y/W)VWL (*26*) 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. [The table shows the differences in nomenclature used to describe ORFs encoding potential nonstructural proteins of SARS-CoV in this report and in the report by Marra *et al*. (*30*). These differences are in nomenclature only, and the seven nt sequence differences between these strains do not change the position or number of ORFs (table S2). Because the complete SARS-CoV sequences have been available for only a few weeks and will probably 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 to X5 is expected to be clarified once experiments on the transcriptional expression of the SARS-CoV genome are reported.]

| Genome location (nt)* | Protein (number of amino acids) | This report† | Marra *et al*. (*30*)‡ |
|---|---|---|---|
| 25,268 to 26,089 | 274 | X1 | ORF3 |
| 25,689 to 26,150 | 154 | X2 | ORF4 |
| 27,074 to 27,262 | 63 | X3 | ORF7 |
| 27,273 to 27,638 | 122 | X4 | ORF8 |
| 27,638 to 27,769 | 44 | <50 amino acids | ORF9 |
| 27,779 to 27,895 | 39 | <50 amino acids | ORF10 |
| 27,864 to 28,115 | 84 | X5 | ORF11 |
| 28,130 to 28,423§ | 98 | See text | ORF13 |
| 28,583 to 28,792§ | 70 | See text | ORF14 |

*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 to X5 and are numbered sequentially beginning at the 5′ terminus of the genome. Only ORFs encoding for predicted proteins longer 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*. (*30*), all of the ORFs, including those encoding the predicted replicase protein and structural proteins, are numbered sequentially from the 5′ terminus of the genome. This table shows only ORFs encoding predicted nonstructural proteins. §These ORFs overlap the coding region of the N protein.

viruses are O-glycosylated (*27*, *28*). The predicted M protein of SARS-CoV has an NGT near its N 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 [FYYLGTGP (*26*)] occurs in the N-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 replicase, S, E, M, or N proteins were found. Like some other coronaviruses, SARS-CoV has several small nonstructural ORFs that are found between the genes for S and E and between the genes for M and N. SARS-CoV is a novel virus that is phylogenetically distinct from other characterized coronaviruses. The genetic distance between 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) (*29*) 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.*, available 17 April 2003 at http://nejm.org/earlyrelease/sars.asp#4-2.
2. N. Lee *et al.*, *N. Engl. J. Med.*, available 17 April 2003 at http://nejm.org/earlyrelease/sars.asp#4-2.
3. K. W. Tsang *et al.*, *N. Engl. J. Med.*, available 17 April 2003 at http://nejm.org/earlyrelease/sars.asp#4-2.
4. Centers for Disease Control and Prevention, *Morb. Mortal. Wkly. Rep.* **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.*, available 17 April 2003 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 & Wilkins, New York, ed. 4, 2001), chap. 35.
9. L. Enjuanes *et al*., in *Virus Taxonomy*, M. H. V. van Regenmortal *et al*., 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 & Wilkins, New York, ed. 4, 2001), chap. 36.
11. Materials and methods are available as supporting material on *Science* Online.
12. Although the match was not statistically significant, the C half of potential protein X1 contains a region of similarity with calcium-transporting adenosine triphosphatases.
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).
17. P. Rota *et al*., data not shown.
18. K. S. Lole *et al*., *J. Virol*. **73**, 152 (1999).
19. J. Ziebuhr, E. J. Snijder, A. E. Gorbalenya, *J. Gen. Virol*. **81,** 853 (2000).
20. S. G. Siddell, Ed., *The Coronaviridae* (Plenum, New York, 1995).
21. D. Escors, J. Ortego, H. Laude, L. Enjuanes, *J. Virol*. **75**, 1312 (2001).
22. H. Garoff, R Hewson, D.-J. E. Opstelten, *Microbiol. Mol. Biol. Rev*. **62,** 1171 (1998).
23. C. M. Sanchez *et al., J. Virol*. **73**, 7607 (1999).
24. I. Leparc-Goffart *et al., J. Virol*. **72**, 9628 (1998).
25. Cleavage sites in the S proteins of coronaviruses are RRFRR, RRSRR, RRSRR, RSRR, RARS, and RARR (*26*) in infectious bronchitis virus, bovine coronavirus, human coronavirus OC43, porcine hemagglutinating encephalomyelitis virus, mouse hepatitis virus, and rat coronavirus, respectively.
26. Single-letter abbreviations for the amino acid residues are as follows: 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.
27. C. A. M. de Haan *et al., Virus Res*. **82**, 77 (2002).
28. C. A. M. de Haan, L. Kuo, P. S. Masters, H. Vennema, P. J. M. Rottier, *J. Virol*. **72**, 6838 (1998).
29. As of this writing, complete genomic sequences of three additional SARS-CoV isolates were available at GenBank (Tor-2 strain, Canada, accession no. ay274119; CUHK-W1 isolate, Hong Kong, accession no. ay278554; and 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.
30. M. A. Marra *et al*., *Science* **300**, 1399 (2003); published online 1 May 2003 (10.1126.science.1085953).
31. The authors thank the WHO SARS Aetiology Laboratory Investigation Group (Bernhard-Nocht Institute, Hamburg, Germany; Erasmus Universiteit, National Influenza Centre, Rotterdam, 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; Public Health Laboratory Service 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; Centers for Disease Control (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 a 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

18 April 2003; accepted 30 April 2003
Published online 1 May 2003;
10.1126/science.1085952
Include this information when citing this paper.

# 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] Martin Czub,[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]**

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 polymerase chain reaction 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%, although a recent report suggests this rate can be as high as 43 to 55% in people older than 60 years (*1*). A number of laboratories worldwide have undertaken the identification of the causative agent (*2, 3*). 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 was named publicly by the World Health Organization and member laboratories as the "SARS virus" (WHO press release, 16 April 2003) after tests of causation according to Koch's postulates, including monkey inoculation (*4*). This virus, which we refer to as SARS-HCoV, was purified, and its RNA genome was extracted and sent to the British Columbia Centre for Disease Control in Vancouver for genome sequencing by the BCCA Genome Sciences Centre.

[1]British Columbia Cancer Agency (BCCA) 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

The coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cells (*5*). They are distinguished by the presence of a single-stranded plus-sense RNA genome about 30 kb in length that has a 5′ cap structure and 3′ polyadenylation 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 adenosine triphosphatase (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 (TRSs) at the 5′ end of each gene represent signals that regulate the discontinuous transcription of subgenomic mRNAs. 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 (*6, 7*). The discovery of transcriptionally active, subgenomic-size minus strands containing the antileader sequence and of transcription intermediates active in the synthesis of mRNAs (*8–11*) favors the model of discontinuous transcription during the minus strand synthesis (*7*).

The viral membrane proteins, including the major proteins S (Spike) and M (membrane), are inserted into the endoplasmic reticulum (ER) Golgi intermediate compartment while full-length replicated RNA plus 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 lumen of the ER. The virus then migrates through the Golgi complex and eventually exits the cell, likely by exocytosis (*5*). 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. Human coronaviruses identified in the 1960s (including the prototype viruses HCoV-OC43 and HCoV-229E) are responsible for up to 30% of respiratory