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

infections (*12*). Coronaviruses are divided into three serotypes: groups 1, 2, and 3 (*13*). 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 gene encoding hemagglutinin esterase (HE) that is 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 (*14*) from the Tor2 isolate (*15*). The RNA was converted to cDNA by means of a combined random-priming and oligo(dT) priming strategy (*14*). Size-selected cDNA products were cloned, and single sequence reads were generated from each end of the insert from randomly chosen clones. Sequences were assembled and the assembly was edited to produce a draft sequence of the viral genome on 12 April 2003 (*14*). Rapid amplification of cDNA ends [RACE (*14*)] was performed to capture the 5′ end of the viral genome. The SARS genomic sequence has been deposited into GenBank (accession number 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 U.S. Centers for Disease Control (CDC) (*16*). We report additional bases in the Tor2 sequence that correspond to the 3′ (encoded) polyadenylation tail. Eight base differences between the two sequences could represent sequencing errors, polymerase chain reaction (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′ (*6, 7*). Putative TRS sequences were determined through manual alignment of sequences upstream of potential initiating methionine codons (see below) with the region of the coronavirus genome sequence containing the leader sequence (Table 2). Candidate TRS sequences were scored as strong, weak, or absent on the basis of inspection of the alignments.

The 3′ untranslated region (3′UTR) sequence contains a 32–base pair region corresponding to the conserved s2m motif (*17*). 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

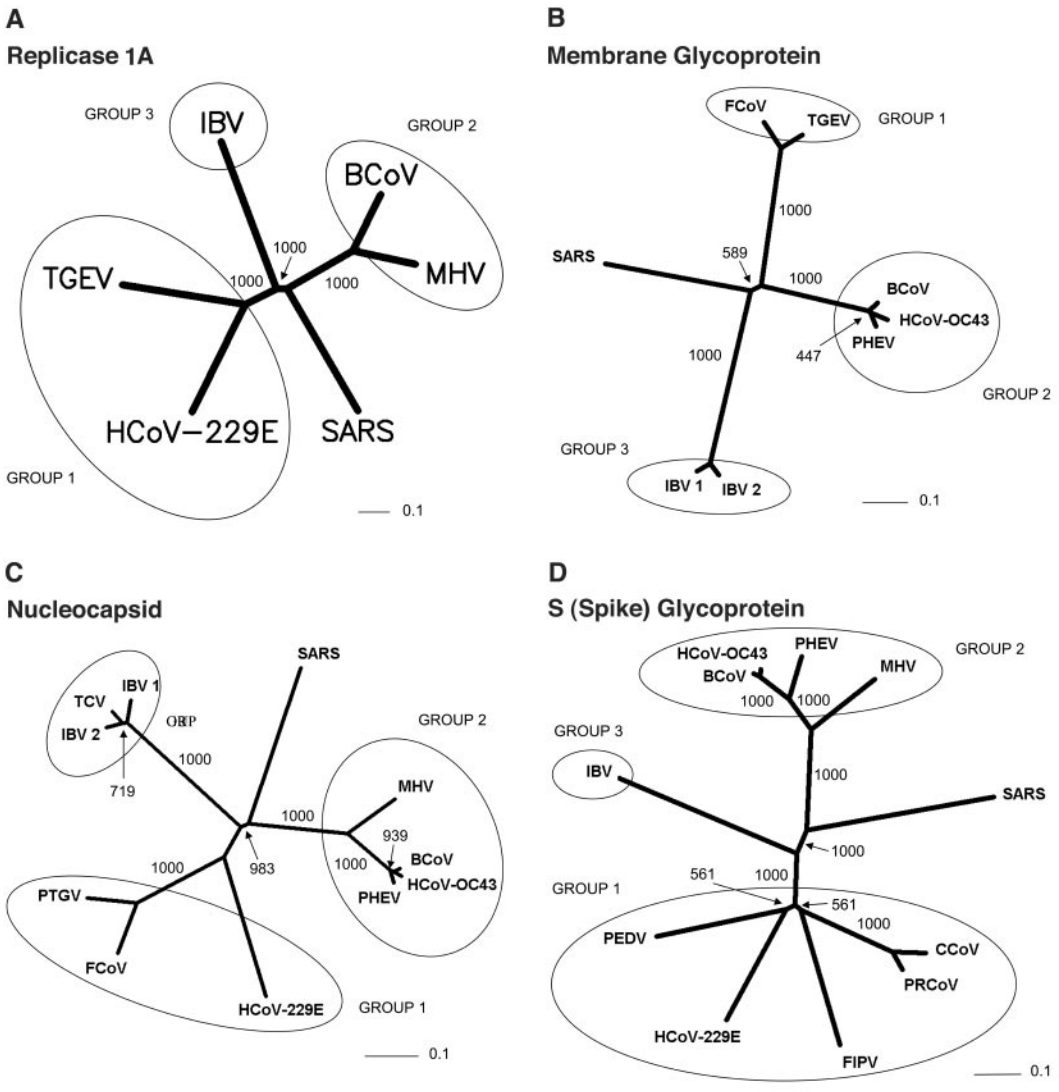

Fig. 1. Phylogenetic analysis of SARS proteins. Unrooted phylogenetic trees were generated by clustalw 1.74 (*33*) using the BLOSUM comparison matrix and a bootstrap analysis of 1000 iterations. Numbers indicate bootstrap replicates supporting each node. Phylogenetic trees were drawn with the Phylip Drawtree program 3.6a3 (*34*). Branch lengths indicate the number of substitutions per residue. GenBank accession numbers for protein sequences are as follows: (**A**) Replicase 1A: BCoV (bovine coronavirus), AAL40396; HCoV-229E (human coronavirus), NP_073555; MHV (mouse hepatitis virus), NP_045298; IBV (avian infectious bronchitis virus), CAC39113; TGEV (porcine transmissible gastroenteritis virus), NP_058423. (**B**) Membrane glycoprotein: PHEV (porcine hemagglutinating encephalomyelitis virus), AAL80035; BCoV, NP_150082; IBV 1, AAF35863; IBV 2, AAK83027; MHV, AAF36439; TGEV, NP_058427; HCoV-OC43, AAA45462; FCoV (feline coronavirus), BAC01160. (**C**) Nucleocapsid: MHV, P18446; BCoV, NP_150083; IBV 1, AAK27162; IBV 2, NP_040838; FCoV, CAA74230; PTGV (porcine transmissible gastroenteritis virus), AAM97563; HCoV-229E, NP_073556; HCoV-OC43, P33469; PHEV, AAL80036; TCV (turkey coronavirus), AAF23873. (**D**) S (Spike) protein: BCoV, AAL40400; MHV, P11225; HCoV-OC43, S44241; HCoV-229E, AAK32191; PHEV, AAL80031; PRCoV (porcine respiratory coronavirus), AAA46905; PEDV (porcine epidemic diarrhea virus), CAA80971; CCoV (canine coronavirus), S41453; FIPV (feline infectious peritonitis virus), BAA06805; IBV, AAO34396.

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 (*17*). 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. ORFs were determined initially through sequence similarity to known coronavirus proteins. This approach identified replicases 1a and 1b, the S protein, the small envelope (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*. (*16*) 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, because internal initiation from larger RNA transcripts is also able to facilitate translation (*18, 19*). Certain ORFs overlap (ORFs 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 ORFs 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 ORF (base pairs 265 to 13,398) and replicase 1b ORF (base pairs 13,398 to 21,485) 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 (*20*). As seen in other coronaviruses and as anticipated, a frame shift interrupts the protein-coding region and separates the 1a and 1b reading frames.

Table 1. Nucleotide base differences between the Tor2 sequence and the Urbani sequence [(*16*), www.cdc.gov/ncidod/sars/sequence.htm]. Boldface indicates a base difference resulting in an amino acid change (*32*); X indicates a nonconservative amino acid substitution.

Position*	Tor2		Urbani			Frame	Protein
	Base	Amino acid	Base	Amino acid			
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	ORF 3
26,857	T	**S**	C	**P**	X	1	M protein

*GenBank AY274119.3.

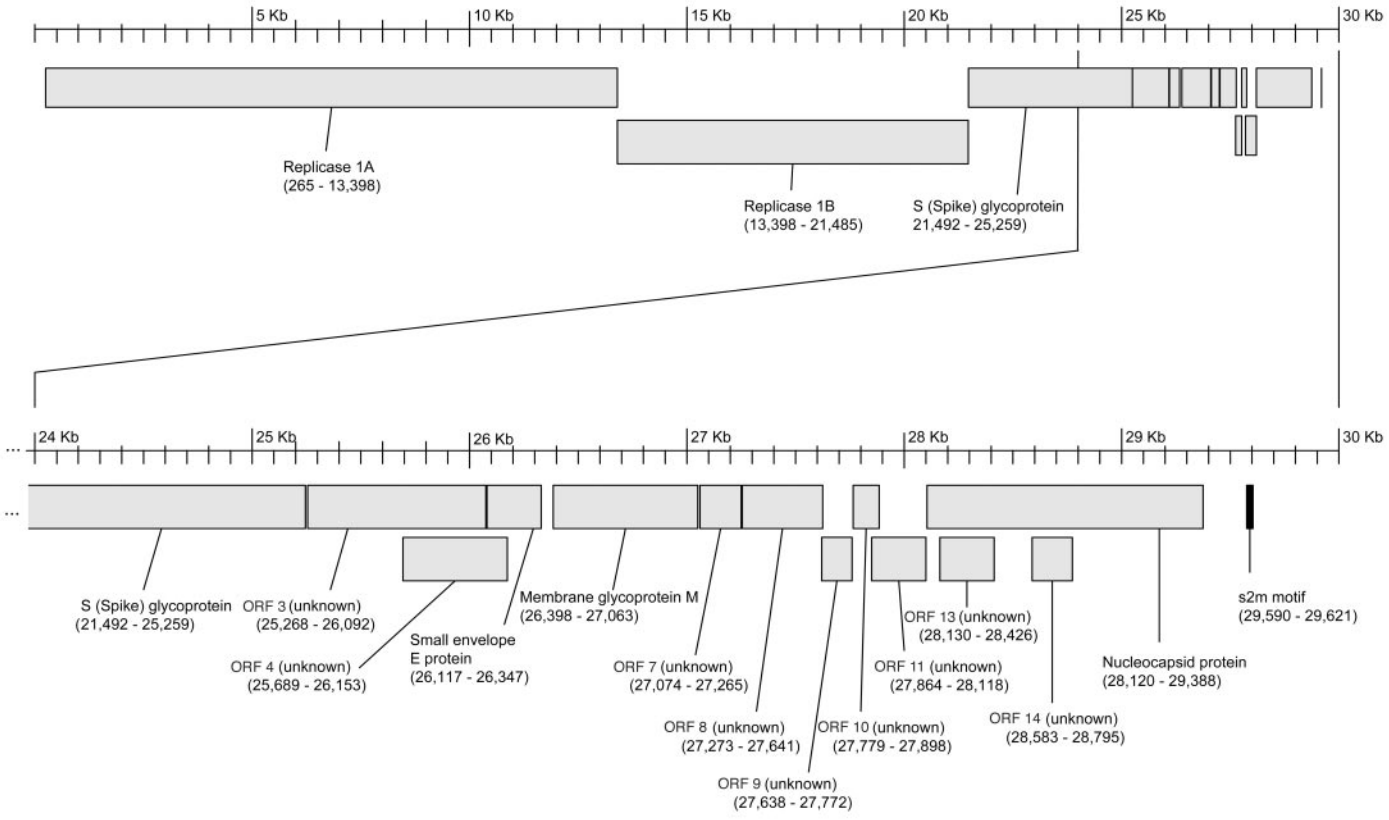

Fig. 2. Map of the predicted ORFs and s2m motif in the Tor2 SARS virus genome sequence.

The Spike (S) glycoprotein (Fig. 2; base pairs 21,492 to 25,259) encodes a surface projection glycoprotein precursor predicted to be 1255 amino acids in length. Mutations in the gene encoding the Spike protein have previously been correlated with altered pathogenesis and virulence in other coronaviruses (*5*). In some coronaviruses, the mature Spike protein is inserted in the viral envelope, with most 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 (*21*) reveals a high probability of a signal peptide (probability 0.996) with cleavage between residues 13 and 14. TMHMM (*22*) 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 to 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 to 546 are required for binding to the cellular receptor, aminopeptidase N (*23*). However, it is known that various coronaviruses use different receptors, and hence it is likely that different receptor binding sites are also used.

ORF 3 (Fig. 2; base pairs 25,268 to 26,092) encodes a predicted protein of 274 amino acids that lacks significant BLAST (*24*), FASTA (*25*), or PFAM (*26*) 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 (*27*) and TMHMM predict the existence of three transmembrane regions spanning approximately residues 34 to 56, 77 to 99, and 103 to 125. The most likely model from these analyses is that the C terminus and a large 149–amino acid N-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 (ProDom ID PD037277).

ORF 4 (Fig. 2; base pairs 25,689 to 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 that this protein is expressed from the ORF 3 mRNA using an internal ribosomal entry site. BLAST analyses fail to identify matching sequences. Analysis with TMpred weakly predicts a single transmembrane helix.

The gene encoding the small envelope (E) protein (Fig. 2; base pairs 26,117 to 26,347) yields 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 that the predicted protein is a member of the well-characterized NS3_EnvE protein family (*26*). InterProScan (*28, 29*) 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 reveals a similar transmembrane domain at positions 17 to 34, consistent with the known association of this protein with the viral envelope. TMHMM predicts a type II membrane protein with most of the hydrophilic domain (46 residues) and the C terminus located on the surface of the viral particle. In some coronaviruses such as porcine transmissible gastroenteritis virus (TGEV), the E protein is essential for virus replication (*30*). In contrast, in mouse hepatitis virus (MHV), although deletion of the gene encoding the E protein reduces virus replication by more than four orders of magnitude, the virus still can replicate (*31*).

The gene encoding the membrane (M) glycoprotein (Fig. 2; base pairs 26,398 to 27,063) yields a predicted protein of 221 amino acids. BLAST and FASTA analyses of the protein reveal 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 (*5*). Analysis of the amino acid sequence with SignalP predicts a signal sequence (probability 0.932) that is not likely cleaved. TMHMM and TMpred analyses indicate the presence of three transmembrane helices, located at approximately residues 15 to 37, 50 to 72, and 77 to 99, with the 121–amino acid hydrophilic domain on the inside of the virus

Table 2. Nucleotide position, associated ORF, 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 boldface in the putative leader sequence. Contiguous sequences identical to region of the leader sequence containing the core sequence are underlined. No putative TRSs were detected for ORFs 4, 13, and 14, although ORF 13 could share the TRS associated with the N protein.

Base	ORF	TRS sequence
60	Leader	UCU**CUAAAC**GAACUUUAAAAUCUGUG
21,479	S (Spike)	CAACUAAACGAAC**AUG**
25,252	ORF 3	CACAUAAACGAACUU**AUG**
26,104	Envelope	UGAGUACGAACUU**AUG**
26,341	M	GGUCUAAACGAACUAACU (40)**AUG**
27,001	ORF 7	AACUAUAAAUU (62)**AUG**
27,259	ORF 8	UCCAUAAAACGAAC**AUG**
27,590	ORF 9	UGCUCUA---GUAUUUUUAAUACUUUG (24)**AUG**
27,766	ORF 10	AGUCUAAACGAAC**AUG**
27,852	ORF 11	CUAAUAAACCUC**AUG**
28,099	Nucleocapsid	UAAAUAAACGAACAAAUUAAA**AUG**

Table 3. Features of the Tor2 genome sequence.

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. The right coordinate of ORF 1a is the end position of the ribosome slippage site (UUUAAC). The most likely frameshift site, based on alignment with other replicase proteins, is 13,392. †See text for details. ‡Corresponding ORFs from Rota *et al.* (*16*). N/R indicates the feature was not reported. §These ORFs overlap substantially or completely with others and may share TRSs. ‖N/A, not applicable.

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.

ORF 7 (Fig. 2; base pairs 27,074 to 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 transmembrane helix located between residues 3 and 22, with the N terminus located outside the viral particle. Similarly, ORF 8 (Fig. 2; base pairs 27,273 to 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 analyses also predict a transmembrane helix located approximately at residues 99 to 117. Together these data indicate that ORF 8 is likely to be a type I membrane protein, with the major hydrophilic domain of the protein (residues 16 to 98) and the N terminus 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.

ORF 9 (Fig. 2; base pairs 27,638 to 27,772) encodes a predicted protein of 44 amino acids. FASTA analysis of this sequence reveals 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) is also detected. The functional implications, if any, of these matches are unknown. TMpred predicts the existence of a single strong transmembrane helix, with little preference for alternate models in which the N terminus is located inside or outside the particle. Similarly, ORF 10 (Fig. 2; base pairs 27,779 to 27,898), encoding a predicted protein of 39 amino acids, exhibits no significant matches in BLAST and FASTA searches but is predicted to encode a transmembrane helix by TMpred, with the N terminus located within the viral particle. The region immediately upstream of ORF 10 exhibits a strong match to the TRS consensus (Table 2), providing support for the notion that a transcript initiates from this site. ORF 11 (Fig. 2; base pairs 27,864 to 28,118), encoding a predicted protein of 84 amino acids, exhibits only very short (9 or 10 residues) matches to a region of the human coronavirus S glycoprotein precursor (starting at residue 801). Analyses 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 gene encoding the nucleocapsid protein (Fig. 2; base pairs 28,120 to 29,388) yields a predicted protein of 422 amino acids. This protein aligns well with nucleocapsid proteins from other representative coronaviruses, although a short lysine-rich region (KTFPPTEPKKDKKKKTDEAQ) (*32*) appears to be unique to SARS. This region is suggestive of a nuclear localization signal, and although 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 that it may assist in RNA binding.

ORF 13 (Fig. 2; base pairs 28,130 to 28,426) encodes a predicted protein of 98 amino acids. BLAST analysis fails to identify similar sequences, and no transmembrane helices are predicted. ORF 14 (Fig. 2; base pairs 28,583 to 28,795) encodes a predicted protein of 70 amino acids. BLAST analysis fails to identify similar sequences. TMpred weakly predicts a single transmembrane helix.

Conclusions. We used genome sequencing to determine 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 (*16*). Although morphologically a coronavirus (*3*), 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 interviral recombination event between the known coronavirus groups as the origin 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 that antibodies to this virus are absent in people not infected with SARS-CoV (*3*), which implies that 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.

The 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, enabling 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 disease.

References and Notes

1. C. A. Donnelly *et al.*, *Lancet*; published online 7 May 2003 (http://image.thelancet.com/extras/03art4453web.pdf).
2. J. S. M. Peiris *et al.*, *Lancet*; published online 8 April 2003 (http://image.thelancet.com/extras/03art3477web.pdf).
3. T. G. Ksiazek *et al.*, *N. Engl. J. Med.*; published online 10 April 2003 (10.1056/NEJMoa030781).
4. R. Munch, *Microbes Infect.* **5**, 69 (2003).
5. B. N. Fields, D. M. Knipe, P. M. Howley, D. E. Griffin, *Fields Virology* (Lippincott Williams & Wilkins, Philadelphia, ed. 4, 2001).
6. M. M. C. Lai, D. Cavanagh, *Adv. Virus Res.* **48**, 1 (1997).
7. S. G. Sawicki, D. L. Sawicki, *Adv. Exp. Med. Biol.* **440**, 215 (1998).
8. D. L. Sawicki *et al.*, *J. Gen. Virol.* **82**, 386 (2001).
9. S. G. Sawicki, D. L. Sawicki, *J. Virol.* **64**, 1050 (1990).
10. M. Schaad, R. S. J. Baric, *J. Virol.* **68**, 8169 (1994).
11. P. B. Sethna *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5626 (1989).
12. S. H. Myint, in *The Coronaviridae*, S. G. Siddell, Ed. (Plenum, New York, 1995), pp. 389–401.
13. 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.
14. Information on materials and methods is available on *Science* Online.
15. S. M. Poutanen *et al.*, *N. Engl. J. Med.*; published online 31 March 2003 (10.1056/NEJMoa030634).
16. P. A. Rota *et al.*, *Science* **300**, 1394 (2003); published online 1 May 2003 (10.1126/science.1085952).
17. C. M. Jonassen, T. O. Jonassen, B. Grinde, *J. Gen. Virol.* **79**, 715 (1998).
18. W. Lapps, B. G. Hogue, D. A. Brian, *Virology* **157**, 47 (1987).
19. R. Krishnan, R. Y. Chang, D. A. Brian, *Virology* **218**, 400 (1996).
20. J. Ziebuhr, E. J. Snijder, A. E. Gorbalenya, *J. Gen. Virol.* **81**, 853 (2000).
21. H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, *Protein Eng.* **10**, 1 (1997).
22. E. L. Sonnhammer, G. von Heijne, A. Krogh, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175 (1998).
23. J. C. Tsai, B. D. Zelus, K. V. Holmes, S. R. Weiss, *J. Virol.* **77**, 841 (2003).
24. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
25. W. R. Pearson, D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (1988).
26. A. Bateman *et al.*, *Nucleic Acids Res.* **30**, 276 (2002).
27. K. Hofman, W. Stoffel, *Biol. Chem. Hoppe-Seyler* **374**, 166 (1993).
28. R. Apweiler *et al.*, *Nucleic Acids Res.* **29**, 37 (2001).
29. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
30. J. Ortego *et al.*, *J. Virol.* **76**, 11518 (2002).
31. L. Kuo *et al.*, paper presented at the annual meeting

of the American Society for Virology, Lexington, KY, 20 to 24 July 2002.
32. 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.
33. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
34. J. Felsenstein, *PHYLIP (Phylogeny Inference Package)* version 3.5c (1993). Distributed by the author, Department of Genetics, University of Washington, Seattle.
35. We thank all the staff at the BCCA Genome Sciences Centre for helping to facilitate the rapid sequencing of the SARS-CoV genome; R. Tellier (Hospital for Sick Children) for information on primer sequences that amplify a 216–base pair region of the Pol gene; I. Sadowski (Department of Biochemistry and Molecular Biology) and J. Hobbs and his staff (Nucleic Acid and Protein Services Unit) of the University of British Columbia for rapid synthesis of PCR primers; F. Ouellette (University of British Columbia Bioinformatics Centre) for advice and assistance; the staff at the National Center for Biotechnology Information for rapidly processing and making available our sequence data; and anonymous reviewers for their useful suggestions. The BCCA Genome Sciences Centre is supported by the British Columbia Cancer Foundation, Genome Canada/Genome British Columbia, Western Economic Diversification, Canada Foundation for Innovation, British Columbia Knowledge Development Fund, Canadian Institutes of Health Research, Michael Smith Foundation for Health Research, and 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.

Stress-Induced Mutagenesis in Bacteria

Ivana Bjedov,[1]* Olivier Tenaillon,[2]* Bénédicte Gérard,[2]* Valeria Souza,[3] Erick Denamur,[2] Miroslav Radman,[1] François Taddei,[1] Ivan Matic[1]†

The evolutionary significance of stress-induced mutagenesis was evaluated by studying mutagenesis in aging colonies (MAC) of *Escherichia coli* natural isolates. A large fraction of isolates exhibited a strong MAC, and the high MAC variability reflected the diversity of selective pressures in ecological niches. MAC depends on starvation, oxygen, and RpoS and adenosine 3′,5′-monophosphate regulons; thus it may be a by-product of genetic strategies for improving survival under stress. MAC could also be selected through beneficial mutations that it generates, as shown by computer modeling and the patterns of stress-inducible and constitutive mutagenesis. We suggest that irrespective of the causes of their emergence, stress-induced mutations participate in adaptive evolution.

Bacteria are champions of evolutionary success; they grow in practically all ecological niches. Evolutionary success depends on phenotypic selection, which in turn depends on available genetic variability. The genetic variability is produced primarily by mutagenesis and secondarily by recombination, which shuffles preexisting mutations. Molecular mechanisms controlling mutation rates are themselves indirectly subject to natural selection through genetic modifications they produce (second-order selection) (*1*, *2*). The linkage between selected mutations and the alleles responsible for their generation is particularly high in bacteria because their gene-transfer and recombination rates are generally low. Consequently, when adaptation is limited by the supply of mutations, selection was shown to favor strains having constitutively increased mutation rates. Such strains display high mutation rates owing to the loss of genetic fidelity functions, e.g., mutational inactivation of the mismatch repair system results in a 10^2- to 10^3-fold increase in mutagenesis (*3*). The selection of constitutive mutators and their role in adaptive evolution of bacteria has been supported by in vivo and in vitro experimental evolution, computer modeling, molecular evolution, and studies of natural isolates (*1*, *2*).

Mutation rates in bacteria can also be increased by stress-induced reversible activation of some gene functions, which results in a transient mutator phenotype, the SOS response being a paradigm of such a process (*4*). However, the evolutionary significance of stress-inducible mutagenesis in bacterial evolution remains a subject of intense debate (*5, 6*). While some argue that it is a consequence of a genetically programmed evolutionary strategy which, by increasing mutagenesis, increases the probability of generation of adaptive variants, others argue that mutations arise in stressed bacteria only as an accidental consequence of accumulation and/or processing of DNA lesions. However, these hypotheses are based on results obtained with laboratory strains. It is difficult to assess the evolutionary significance of any phenomenon without knowing its frequency and ecological distribution in natural populations, as well as their physiological and genetic determinants. With this premise, we have studied stress-induced mutagenesis phenotypes among 787 worldwide natural isolates of *Escherichia coli* from diverse ecological niches: commensal and pathogenic isolates from a variety of hosts and isolates from air, water, and sediments (*7*). To mimic stress conditions commonly encountered by bacteria in natural environments, we used progressive starvation following an exponential growth phase occurring in colonies. We chose colonies, instead of liquid cultures, because the primary natural *E. coli* habitat is the gut of warm-blooded animals, where it can be found in tightly packed communities. In secondary environments, like soil and water, bacterial cells also tend to aggregate and form (micro)colonies and biofilms.

Diversity of constitutive and colony-aging induced mutation rates among natural isolates of *E. coli*. To estimate mutagenesis in aging colonies (MAC) of natural isolates of *E. coli*, we measured the frequency of mutations conferring resistance to rifampicin (Rif^R) in 1-day- (D1) and 7-day- (D7) old colonies (*7*). For all strains, the median values of the frequency of Rif^R mutations were 5.8×10^{-9} on D1 and 4.03×10^{-8} on D7 (Fig. 1, A and B). Thus, the frequency of mutations increased on average sevenfold between D1 and D7 [(Fig. 1C) Mann-Whitney: $P < 0.0001$], while the median number of colony-forming units (CFU) increased 1.2-fold. The *E. coli* K12 MG1655 laboratory strain showed a 5.5-fold increase in frequency of Rif^R mutagenesis and a 1.7-fold increase in CFU. The increase in CFU from D1 to D7 was not correlated with the increase in the D7/D1 ratio of mutation frequency.

Strains having D1 mutation frequencies >10-fold or >100-fold higher than the median D1 mutation frequency of all the strains represented 3.3 and 1.4% of isolates, respectively, which corresponds to previous reports on the frequency of constitutive mutators in natural *E. coli* populations (*8–10*). The D7/D1 mutation frequency ratio showed that 40% of strains had more than a 10-fold, and 13% more than a

[1]INSERM U571, Faculté de Médecine Necker-Enfants Malades, Université Paris V, 156 rue Vaugirard, 75730 Paris Cedex 15, France. [2]INSERM E0339, Faculté de Médecine Xavier Bichat, Université Paris VII, 16 rue Henri Huchard, 75870 Paris Cedex 18, France. [3]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Apartado Postal 70-275, México D.F. 04510, México.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: matic@necker.fr