Coronaviruses with Special Emphasis on First Insights Concerning SARS
ed. by A. Schmidt, M.H. Wolff and O. Weber

Genome organization and structural aspects of the SARS-related virus

Caroline R. Astell, Robert A. Holt, Steven J. M. Jones and Marco A. Marra

Genome Sciences Centre, British Columbia Cancer Agency, Suite 100–570 West 7th Ave., Vancouver, B.C., Canada V5Z 4S6

Background

The first appearance of severe acute respiratory syndrome (SARS) occurred in Guangdong province in southern China with the earliest cases dating from November 16, 2002. By February 14, 2003 the WHO reported a total of 305 cases of acute respiratory syndrome of unknown etiology in Guangdong province (WHO WER 7/2003). At the time of writing this review, scientists in the Guangdong province of China believe that the SARS virus may have shifted to human hosts by at least five independent events, however sequence information to support this conclusion on these five isolates is still incomplete. The SARS virus was spread to Hong Kong by a physician from Guangdong when he traveled to Hong Kong and stayed at the Metropole Hotel on Feb 21. From there, world-wide dissemination of the virus to Vietnam, Singapore, Taiwan and Canada plus other locations occurred.

The SARS agent was initially believed to be an influenza virus, possibly an avian influenza virus, parainfluenza virus (metapneumovirus) or a bacterium, *Chlamydia pneumoniae* However, by March 19, the WHO reported that these agents were unlikely to be the cause and hence suggested that a new agent was responsible. Within a few days four groups obtained evidence that a coronavirus-like agent might be the causative agent using PCR primers for known coronaviruses to amplify short fragments of DNA which were sequenced ([1–3]; R. Tellier, personal communication). In addition de Risi's group at UCSF used a DNA microarray-based assay to detect viral sequences from samples cultured in Vero6 cells. These results were made available over the internet and have now appeared in print [4]. This group also recovered a ~ 1kb fragment of viral cDNA which was purified from contaminating cellular cDNAs by selecting the fragment on a DNA microarray and sent the DNA to the Washington University Genome Sequence Centre. The sequence was determined and the results indicated that it was most closely related to coronaviruses.

Hence all of these results suggested a novel coronavirus-like agent might be the cause of SARS.

With the world-wide scientific community uncertain of the causative agent of SARS and the fact that SARS had now been transported to Canada by travelers from Asia, the Genome Sciences Centre (GSC) at the BC Cancer Agency decided on March 27 to apply it's high-throughput DNA sequencing capabilities to fully characterize the genome of the "new" virus. Colleagues at the British Columbia Centre for Disease Control (BC CDC) as well as the National Microbiology Laboratory (NML) were contacted. The NML in Winnipeg, Canada's National Reference Laboratory which has a level 4 biosafety facility, had already received patient samples from Toronto and Vancouver and had succeeded in growing the virus in Vero6 cells. The NML purified the virus and were able to send to the GSC a sample of highly purified viral RNA (150 ng) which simplified considerably our ability to rapidly sequence the genome. This sample was obtained from the second victim of SARS in Toronto and was termed Tor2. As soon as the virus sample arrived at the BC CDC (approximately 5 PM on April 6) our rapid sequencing effort began.

The initial strategy for sequencing the genome was to use primers designed from homologous regions of all known coronaviruses to amplify overlapping fragments from the genome and then sequence these fragments. However, we repeated experiments to amplify a small region of the replicase gene using pancoronavirus primers, sequenced this fragment and confirmed that the sequence was related about 80 % at the protein level but only 50% at the nucleic acid level. This relatively low level of homology at the nucleic acid level caused us to shift our initial strategy to that of using combined oligo-dT and random primers to create a library of cDNA clones in two different vectors that could be submitted to our high-throughput sequencing pipeline.

RT-PCR of SARS-CoV genome and construction of genomic libraries

The manipulation of full-length viral RNA was carried out in a level 2 biosafety laboratory. The strategy used to sequence the SARS genome was to construct a library of DNA fragments spanning the ~30,000 base genome. Purified viral RNA (55 ng) was used in an oligo-dT and random primers RT-PCR reaction using the SuperScript Choice System for cDNA synthesis (Invitrogen Canada Inc., Burlington, Ontario, Canada). At this stage, the samples were handled in a level 1 laboratory. EcoR I linkers were added to the DNA fragments and the products analyzed on an agarose gel. As expected, the products were barely visible and showed a range of sizes from ~ 400 bp–3000 bp. The fragments were resolved on a low melting point agarose gel and fragments of 1000–4000 bp were recovered. These frag-

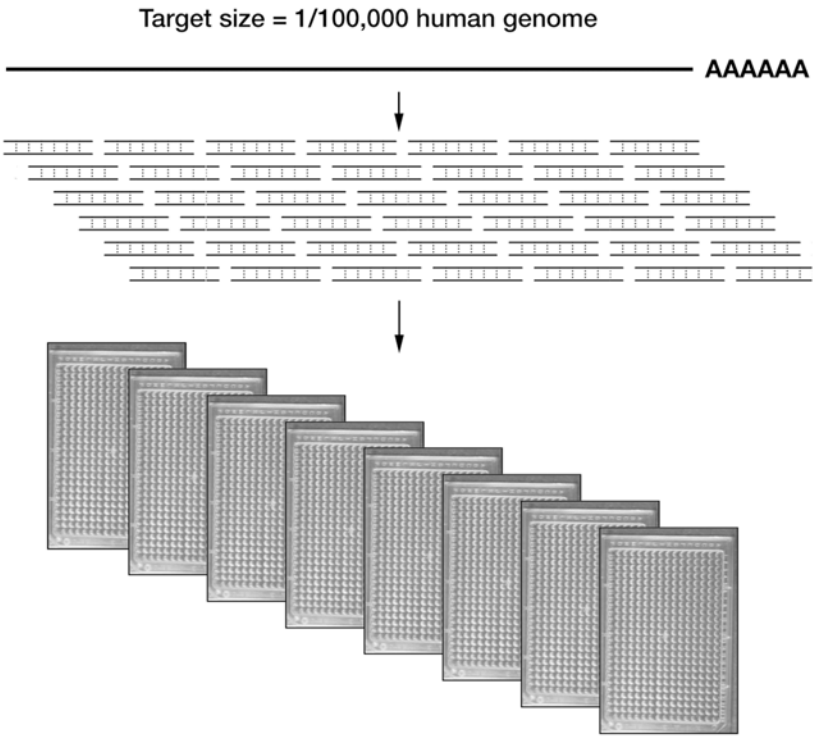


Figure 1. Whole genome shotgun approach used to sequence the SARS coronavirus genome. The SARS genome is approximately 1/100,000 the size of the human genome. A library of fragments (~2,000 bp in length) was generated and arrayed into ten 384-well plates. The clones were sequenced from either end and the data assembled into the full length 29,751 nucleotide sequence [11] (see text for details).

ments were amplified using the EcoR I linker as primers and ligated with the plasmid pCR4-TOPO TA cloning vector (Invitrogen) at the EcoR I site or with pBR194 [5] using the Not I site. Each library was transformed into DH10B T1 cells (Invitrogen) and plated on 22 cm^2 agar plates and transformants obtained using ampicillin selection. After 16 hours, colonies were picked and grown in 2X YT and plasmid DNAs purified using our standard high-throughput alkaline lysis procedure.

DNA sequencing, assembly of reads and analysis of data

The fragments generated by the cloning procedure are illustrated schematically in Figure 1. Each fragment was sequenced from both ends with appropriate primers using BigDye terminator reagent (version 3, Applied Bio-

systems, Foster City, CA, USA). The fragments were resolved by electrophoresis on either an AB 3700 or AB 3730XL automated sequencing instrument and the data was collected automatically and transferred electronically to the Bioinformatics team for analysis. DNA sequence chromatograms were processed and trimmed for sequence quality using PHRED software [6]. Sequence reads were subsequently screened for non-viral contaminating sequences (almost exclusively plasmid vector sequences allowing us to infer that the virus RNA sample provided by the NML had been highly purified. The assembly of the sequence reads was carried out using the PHRAP sequence assembly software [7]. The sequences were also analyzed using BLAST [8] and FASTA [9] to search the viral and non-redundant protein datasets derived from the National Centre for Biotechnology Information (NCBI). This revealed that the sequences were coronavirus related, but quite unlike any previously characterized coronavirus. Sequence data was accumulated until it was apparent that additional reads increased the depth of coverage but not the length of the sequence. Of the first 3,080 sequence reads, 2,634 assembled into one large contig.

Once the full-length genome was assembled (minus the extreme few bases at the 5' end) a manual annotation of the genome was carried out using a series of computational resources. Annotation of the SARS virus sequence was carried out using the ACEDB genome database system [10]. Release 1 of the draft sequence was made public on the BCCA GSC website on April 12, 2003 in addition to being submitted to the NCBI Genbank sequence database. On April 14, the CDC (Atlanta) released their independently derived sequence for the Urbani strain of the SARS virus obtained using an RT-PCR method and direct sequencing of fragments. On comparison of these two sequences 12 nucleotides differed between the two strains. On further analysis four of these differences were resolved, leaving a total of only 8 differences between these strains [11]. Version 2 of our sequence was posted to our website on April 14, 2003 and this update was simultaneously submitted to the Genbank database. Release 2 of the genome sequence had an average PHRED consensus quality score of 89.96 with the lowest quality bases at the 5' and 3'ends of the viral genome. This average score corresponds to an expected error rate in sequence determination of 1 error in $10^{8.96}$ base pairs. However, such an error rate corresponds only to the sequencing technology itself and does not account for other sources of potential error, e.g. reverse transcriptase errors during early phases of library construction. Each base in the sequence was determined on average 60 times (30 times in a forward direction and 30 times in a reverse direction). A subsequent release included the 5' most nucleotides which were determined by the RACE procedure using the RLM-RACE kit (Ambion Inc., Woodward, Austin, TX, USA). Release 3, the complete genome sequence, was deposited in Genbank (accession AY274119). We also immediately made available our clones via our website www.bcgsc.ca.

We subsequently validated all the clones and have also made these validated clones available to qualified researchers.

In addition to the sequencing efforts at the BC GSC [11] and the CDC Atlanta [12] groups, the Washington University Genome Sequence Centre in collaboration with the de Risi group in San Francisco used a method very similar to the one used at the GSC to generate a library of cDNA clones which were sequenced by the shotgun method [4]. These scientists assembled about 25,000 bases of sequence data which was freely sent to the SARS group at the CDC in Atlanta by J. deRisi on April 12, 2003. Shortly after the publication of the Tor2 and Urbani sequences, another group reported determining the sequence of several related isolates by the whole genome shotgun approach [13]; however of the more than 100 sequences now deposited at GenBank, most have been generated by sequencing of RT-PCR fragments using the Tor2 or Urbani sequences to design primers.

Relatedness of the SARS-CoV with other coronaviruses

Coronaviruses are members of the order *Nidovirales*, family *Coronaviridae*, genus *Coronavirus* [14]. Comparison of the Tor2 [11] (Genbank AY274119) and the Urbani sequence [12] (Genbank AY 278741) data showed that the sequences were essentially identical with only 8 bases difference out of ~29,700 bases. The analysis of the sequences by both groups was also very similar. The nomenclature used by each group differed slightly and in this review we maintain the naming system used for the Tor2 sequence [11]. For the convenience of the reader, Table 1 summarizes the nomenclature used by both groups as well as that used by Thiel et al. [15].

Preliminary analysis of the SARS genome sequence using BLAST and FASTA identified a total of 14 putative open reading frames (ORFs) possessing initiating codons (Fig. 2). For completeness, all detected ORFs were annotated in the initial genome analyses. Marra et al. [11] annotated ORFs that did not match database sequences but were identified if they were larger than 40 amino acids and had a strong match to the TRS consensus upstream of the potential initiating methionine residue. In contrast, Rota et al. did not identify potential proteins of less than 50 amino acids [12]. It is important to note that for some of these predicted ORFs no experimental evidence or sequence similarity information currently exists to support the notion that they actually do produce proteins *in vivo*.

All five major ORFs found in all known coronavirus genomes were identified: the replicase proteins (ORFs 1a and 1b), the spike or S protein ORF, the M or membrane glycoprotein ORF, the E or small membrane protein ORF and the N or nucleocapsid ORF. Coronaviruses were originally divided into three serotypes, groups 1 and 2 (predominantly mammalian viruses) and 3 (predominantly avian viruses) based on antigenic cross-reactivity. This grouping of the viruses also agrees fairly well with the

Table 1. Summary of terninology used for SARS-CoV Open Reading Frames

Marra et al.[11]	Rota et al. [12]	Thiel et al. [15]	Protein encoded
1a	1a	1a	Replicase
1b	1b	1b	Replicase
2	S	2	Spike
3	X1	3a	Unknown
4	X2	3b	Unknown
5	E	4	E small membrane
6	M	5	Membrane glycoprotein
7	X3	6	Unknown
8	X4	7a	Unknown
9	N/R	7b	Unknown
10	N/R	8a	Unknown
11	X5	8b	Unknown
12	N/R	9a	Nucleocapsid
13	N/R	9b	Unknown
14	N/R	N/R	Unknown

NB. In animal isolates of the SARS-CoV, ORFs 10 and 11 are replaced by a 122 aa of unknown function. This is the result of an additional 29 nt in the animal genome that are apparently deleted in almost all human isolates. In this paper that protein is referred to as ORF 10'.
N/R, not recognized. Rota et al. did not recognize these ORFs because their minimal length cutoff was greater that 50 amino acids.

phylogenetic similarity of the viral genomes; however, some genomes have diverged sufficiently that the viruses are no longer antigenically cross-reactive. Phylogenetic analysis of four of the SARS ORFs (Rep, S, M, and N) by both the Vancouver and CDC Atlanta groups using unrooted analyses provided evidence (for all four comparisons) that the SARS virus represents a fourth group within the coronavirus family [11, 12]. Figures 3A and B illustrate the comparison for ORF 1a [11] and ORF1b [12], respectively. However, others have now constructed a rooted phylogenetic tree, using the ORF1b gene of equine torovirus. This analysis suggests that the SARS-CoV is more related, albeit distantly, to group 2 coronaviruses than to any other group and shares a common ancestor early in coronavirus evolution (Fig. 3C) [16].

Analysis by both Marra et al. [11] and Rota et al. [12] also suggested that the SARS-CoV is not a *recent* recombinant between two (or more) other known coronaviruses, which may have explained its sudden appearance in humans and its pathogenicity. However, Rest and Mindell [17] have completed a more in-depth analysis of the RNA-dependent RNA polymerase (RdRp) coding sequence and found evidence that during evolution of the SARS-CoV the 3' region of the RdRp gene may have been derived from avian coronaviruses possibly due to a recombination event. Yet another study by Stavrinides and Guttman suggests that the left portion of the SARS genome is derived from mammalian-like viruses while the right portion is derived from avian sources [18]. The position of recombination appears to be within the 3' end of the S protein gene. Stavrinides and

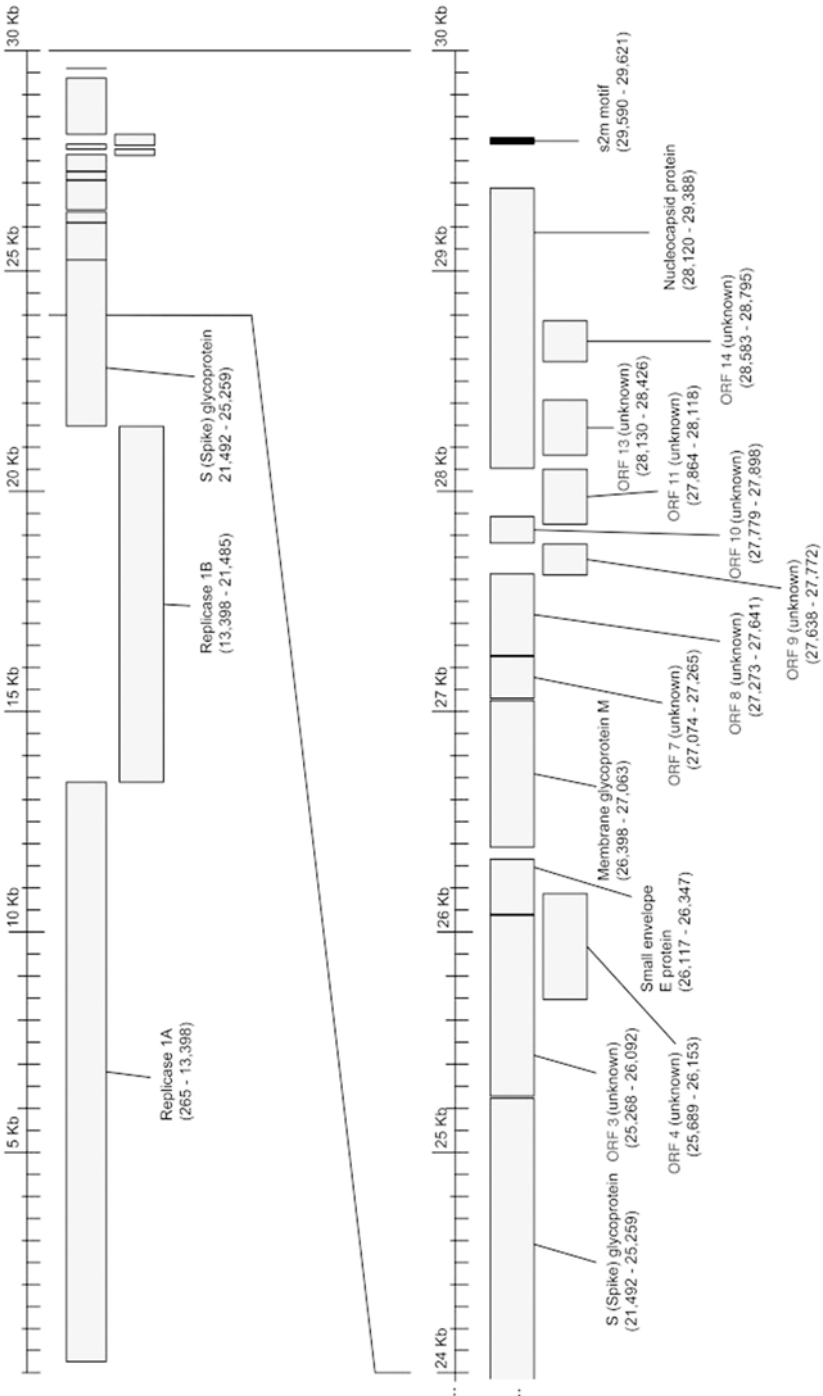


Figure 2. Map of the predicted ORFs and s2m motif in the Tor2 SARS virus genome sequence (reprinted with permission from [11]).

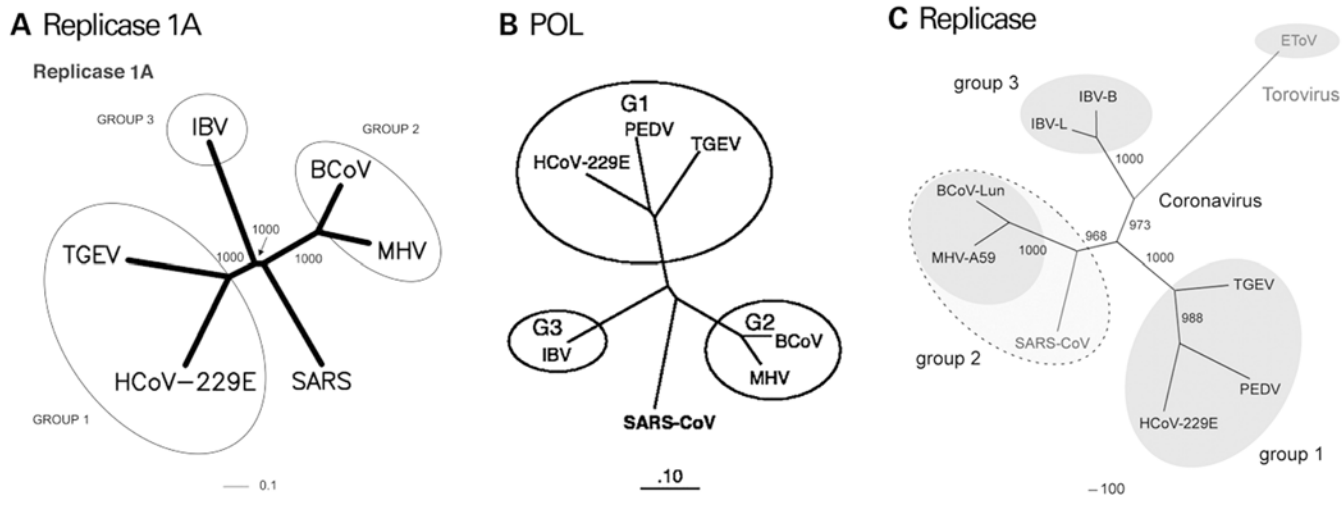


Figure 3. Phylogenetic analysis of SARS coronavirus open reading frames. (A) is ORF1a [11] and (B) is the POL gene from ORF 1b [12] obtained using an unrooted phylogenetic analysis methods, (C) is ORF 1b obtained using a rooted phylogenetic analysis method [16]. (Fig. 3A reprinted with permission from [11], Fig. 3B reprinted from [12], Fig. 3C reprinted from [16], with permission from Elsevier.)

Guttman proposed that this recombination within the S protein gene may have been the event that allowed the SARS virus to spread to human hosts [18]. However, the lack of evidence for a mammalian virus that is closely related to SARS to the left of the recombination point and an avian virus closely related to sequences to the right, leaves room for an alternative interpretation. The SARS-CoV may have arisen by an ancient recombination occurring a long time ago in the natural history of the SARS coronavirus and that subsequently this recombinant virus has been evolving over a long time period would seem like a more rational interpretation of these data. There are currently over 100 SARS-CoV complete or partial sequences in GenBank http://www.ncbi.nlm.nih.gov/query.fcgi?db=nucleotide&cmd=search&term=%22SARS+coronavirus%22, including several isolates from civet cats, one of several possible animal hosts responsible for transmitting the virus to humans [19]. While it is unclear if civet cats are the natural reservoir for this novel coronavirus, there does appear to have been another transfer event (maybe from civet cats or rats) in Guangdong, reported in December 2003. Analysis of this sequence will hopefully shed more light on the origin of the human SARS virus, however at this point, most hypotheses are based on a very limited amount of data. Future characterization of many additional animal SARS virus isolates will help to elucidate the true origin of the human SARS-CoV.

Mechanism of transcript expression by coronaviruses including identification of TRS sequences in the SARS-CoV

The expression of coronavirus genes is thought to occur by discontinuous transcription of the plus sense genomic RNA to generate a series of subgenomic RNAs which are subsequently transcribed into opposite sense mRNA [20, 21] (Fig. 4). Sequences referred to as TRS (transcription regulatory sequences) occur within the genomic RNA upstream of the major ORFs. These TRS sequences which have a common core sequence (5'-UAAACGAAC-3') are believed to facilitate the dissociation of the replicase complex from the genomic RNA and reassociation with similar sequences within the template RNA near the 5' end of the genome referred to as the leader sequence, generating a nested set of minus strand subgenomics RNAs. The mRNAs are then transcribed in a continuous fashion from these subgenomics negative strands, generating a nested set of mRNAs. Hence each of the mRNAs contains a leader sequence of ~72 nt (corresponding with the 5' end of the viral genome) attached to one of the downstream viral ORFs and all sequence 3' to that region.

In the bioinformatic analysis of the SARS genome by Marra et al. [11] candidate TRS sequences upstream of 10 of 13 "internal" ORFs and within the 72 nt leader sequence upstream of ORFs 1a and 1b were identified. Putative TRS sequences containing the central nucleotides of the core

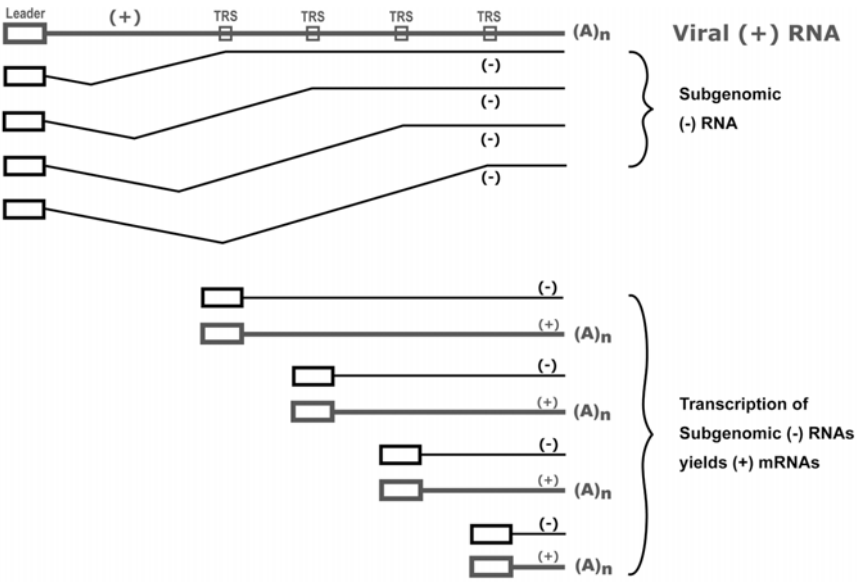


Figure 4. Preferred transcription model for coronaviruses. The mechanism illustrated is believed to be responsible for generating the nested set of mRNAs used to express coronavirus proteins. The coronavirus plus strand genome is transcribed into subgenomic minus strand RNAs. During the transcription process the RdRp encounters one of several TRS (transcription regulatory sequences) located upstream of the viral ORFs. The RdRp then jumps to the 5' leader sequence and finishes synthesis of the transcripts. The family of subgenomics minus strands are then copied into a family of subgenomic plus strands, the mRNAs used for translation of the viral proteins. The mRNA for ORF1a and ORF1b is synthesized from a full length minus strand transcript [22, 23] and during translation can undergo a –1 ribosomal frameshift (see text and Fig. 7 below).

sequence, 5'-ACGAAC-3', were identified upstream of 11 of the 14 ORFs in the SARS-CoV. Some were identified as strong (good matches to the consensus sequence) while others were classed as weak while still other upstream regions appeared not to contain an obvious TRS core sequence. Those lacking obvious TRS core sequences include ORFs 4, 13 and 14. It was speculated that possibly the three putative ORFs without an upstream TRS are translated from a longer transcript (containing an upstream ORF) by a process of internal ribosomal initiation [11]. Functional bicistronic mRNAs have been observed for other coronavirus subgenomic mRNAs [22, 23]. The ORFs 1a and 1b are expected to be translated from full length mRNAs copied from full-length negative sense RNA.

Two groups of investigators have used Northern blotting to identify the subgenomic species of RNA expressed in Vero6 cells infected by the SARS virus. Rota et al. [12] detected 5 subgenomic RNAs while Snijder et al. [16] and Thiel et al. [15] identified a total of 8 subgenomic RNAs (1.8, 2.1, 2.6,

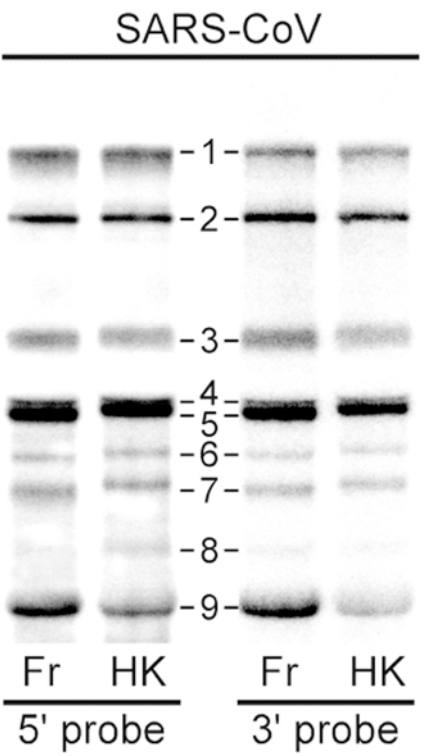


Figure 5. Transcripts generated in cultured cells infected with either the Frankfurt-1 (Fr) or HKU-39849 (HK) isolates of the SARS-CoV [16]. The transcripts were detected using 5' or 3' probe genomic probes. There are 8 subgenomic transcripts ranging from 1.8 to 8.4 kb plus the full-length transcript (reprinted from [16] with permission from Elsevier).

3.0, 3.5, 3.8, 4.6 and 8.4 kb) (Fig. 5) plus the full-length mRNA (29.7 kb) from which ORFs 1a and 1b are expected to be translated. The molar ratio of the fragments indicates that some transcripts are clearly more abundant than others, with the most abundant transcripts the ones that likely are translated into the E and M proteins. If the inability to detect additional subgenomic transcripts is not simply due to their low level of expression, it would appear that the use of bicistronic mRNA is likely the method used to express at least half of the putative proteins encoded by the ORFs in the right one third of the SARS genome. Thiel et al. [15] have also confirmed the leader to body fusion sites on the subgenomic mRNAs using RT-PCR to transcribe the mRNAs expressed in SARS-CoV-infected Vero6 cells. Subsequent PCR amplification using a second "body" primer plus a primer specific for the leader sequence generated fragments that were sequenced to determine the junction of the leader to the body of the mRNA.

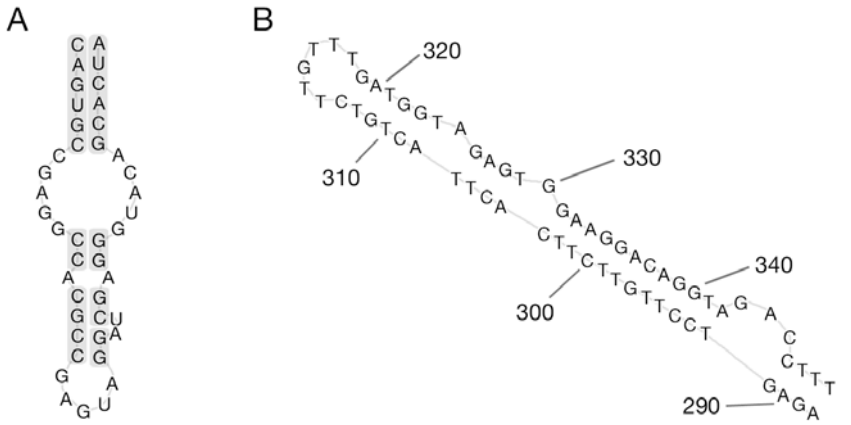


Figure 6. Secondary structural motifs within the SARS-CoV genome. (A) illustrates the conserved s2m motif found near the 3' end of many coronavirus genomes [24], (B) illustrates the putative SARS RNA packaging signal within the 3' end of ORF 1b [26] (reprinted with permission).

Other non-coding features of the SARS-CoV genome

Coronavirus genomes are single-stranded plus sense RNAs which are both capped at the 5' end and polyadenylated at the 3' end. There is a putative 2'-O-methyl transferase and putative mRNA cap 1-methyl transferase within the large replicase gene (see below, section "Coding potential of the SARS CoV, ORFs 1a and 1b-replicase) of the SARS-CoV suggesting that these proteins have a role in synthesizing the cap structure. However, so far there is no putative guanyl transferase gene identified, nor is there a polyA polymerase gene identified. However since the incoming viral genome is polyadenylated and the replication mechanism described above would provide an explanation for the presence of a polyA tract, this is the likely mechanism by which the mRNAs are polyadenylated.

In addition there is a 32 nt region corresponding to the conserved s2m motif [24]. This imperfect hairpin structure (Fig. 6A) is a feature of all astroviruses and has also been found in avian infectious bronchitis virus (avian IBV) and the ERV-2 equine rhinovirus. Due to their high degree of sequence similarity and occurrence in viruses that are evolutionarily distant, Jonassen et al. [24] proposed that this motif is indicative of multiple horizontal transfer events.

A recent bioinformatic analysis of the SARS-CoV genome has identified a putative genome packaging signal. The packaging signal for two laboratory coronavirus models systems, murine hepatitis virus (MHV) and bovine coronavirus (B-CoV) has been identified an ~69 nt region near the

3' end of ORF 1b, the second half of the large ORF encoding the replicase gene. This sequence is capable of driving the packaging of foreign RNA into MHV particles [25]. Qin et al. [26] have characterized the secondary structure of this packaging signals using alignment of the sequences and a RNA secondary structure program, RNA Structure 3.71 [27]. The structural motif for the packaging signals of several coronaviruses appears to be a long stem-loop structure of which the top of loops are similar. Analysis of the SARS genome showed that a similar structure (Fig. 6B) is found in a hypervariable region of ORF1b displaced slightly upstream relative to the packaging signals for MHV and BCoV. This packaging signal needs to be tested experimentally and may prove to be a useful target for development of an antiviral agent.

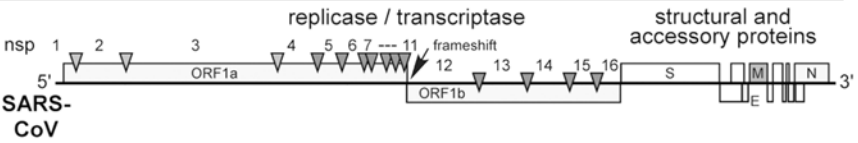
Coding potential of the SARS-CoV

In this section we summarize the coding potential of each of the 14 ORFs identified by Marra et al. [11] and Rota et al. [12] (see Tab. 2). We also include information about the proteins that has become available during the past 9 months since the genome sequences were published. A search on the plus sense strand identified a number of ORFs (Fig. 2). There are a number of smaller ORFs on the minus strand but to date no coronavirus has been shown to encode genes on the minus sense strand.

ORFs 1a and 1b-replicase

The left two thirds (21.2 kb) of the genome was found to contain two large ORFs corresponding to the replicase 1a and 1b regions. ORF 1a is translated into a large polypeptide of ~4,000 aa while a fused ORF1ab can be translated into a polyprotein of ~7,000 aa. As is the case with other coronaviruses, ORFs 1a and 1b can be fused together into one large ORF by a –1 frameshift event which is believed to occur by a process of slippage on the ribosome [28]. In the case of the SARS virus the C-terminal end of the ORF 1a polypeptide is …. L-N-G-F-A-V-STOP (Fig. 7). At least part of the time a –1 frameshift occurs, yielding a read-through polypeptide with the sequence ….L-N-R-V-C-G-V – etc. A pseudoknot, a specialized region of secondary structure within the mRNA, immediately downstream from the position of the frameshift, is believed to cause the ribosome to pause during translation. Upstream of the pseudoknot is a consensus "slippery sequence" NNNAAAN.. or ..NNNTTTN.. where the first three Ns can be any nucleotide but they must be the same nucleotide [29]. In the case of the SARS genome the sequence is …TTTAAACGGG…. When the ribosome encounters the pseudoknot, there is a pause in reading of the mRNA and the mRNA is shifted back one nucleotide, allowing only partial yet stable

Table 2. Predicted SARS-CoV replicase cleavage products and their mode of expression (see text for details). Modified from [16].



Protein order[a] in polyproteins pp1a/pp1ab	Position in polyproteins pp1a/pp1ab (amino acid residues)[b]	Protein size (amino acid residues)	Associated putative functional domain(s)[c]	Predicted mode of expression and release from polyproteins[d]
nsp1-pp1a/pp1ab	1Met-Gly180	180	?	TI + $PL2^{pro}$
nsp2-pp1a/pp1ab	181Ala-Gly818	638	?	$PL2^{pro}$
nsp3*-pp1a//pp1ab	819Ala-Gly2740	1922	Ac, X, $PL2^{pro}$, Y (TM1), ADRP	$PL2^{pro}$
nsp4-pp1a/pp1ab	2741Lys-Gln3240	500	TM2	PL2 + $3CL^{pro}$
nsp5-pp1a/pp1ab	3241Ser-Gln3546	306	$3CL^{pro}$	$3CL^{pro}$
nsp6-pp1a/pp1ab	3547Gly-Gln3836	290	TM3	$3CL^{pro}$
nsp7-pp1a/pp1ab	3837Ser-Gln3919	83	?	$3CL^{pro}$
nsp8-pp1a/pp1ab	3920Ala-Gln4117	198	?	$3CL^{pro}$
nsp9-pp1a/pp1ab	4118Asn-Gln4230	113	?	$3CL^{pro}$
nsp10-pp1a/pp1ab	4231Ala-Gln4369	139	GFL	$3CL^{pro}$
nsp11-pp1a	4370Ser-Val4382	13	?	$3CL^{pro}$ + TT
nsp12-pp1ab	4370Ser-Gln5301	932	RdRp	RFS + $3CL^{pro}$
nsp13-pp1ab	5302Ala-Gln5902	601	ZD, NTPase, HEL1, mRNA CAP 1-mt	RFS + $3CL^{pro}$
nsp14-pp1ab	5903Ala-Gln6429	527	Exonuclease (ExoN homolog)	RFS + $3CL^{pro}$
nsp15-pp1ab	6430Ser-Gln6775	346	NTD, endoRNase (XendoU homolog)	RFS + $3CL^{pro}$
nsp16-pp1ab	6776Ala-Asn7073	298	2′-*O*-MT	RFS + $3CL^{pro}$ + TT

Predictions are based on the SARS-CoV sequences published by Michael Smith Genome Sciences Centre (Vancouver, Canada; Entrez Genomes accession number NC_004718 (AY274119)[a]) and the Centers for Disease Control and Prevention (Atlanta, USA; GenBank accession number AY278741[b]) and an alignment of SARS-CoV with previously characterized coronavirus sequences as summarized in Refs. [32-34].

[a] For convenience, replicase cleavage products were provisionally numbered non-structural protein (nsp) 1–16 according to their position in the polyproteins.

[b] Amino acids of replicase proteins pp1a and pp1ab were numbered assuming that, as in other coronaviruses, a −1 ribosomal frameshift occurs; use of the slippery sequence UUUAAAC[10] is predicted to yield a peptide bond between Asn4378 and Arg4379 in pp1ab.

[c] Abbreviations: $PL2^{pro}$, papain-like proteinase 2; ADRP, adenosine diphosphate-ribose 1″-phosphatase; TM, transmembrane domain; $3CL^{pro}$, 3C-like cysteine proteinase; GFL, growth factor-like domain; RdRp, RNA-dependent RNA polymerase; ZD, putative Zinc-binding domain; HEL1, superfamily 1 helicase; NTD, nidovirus conserved domain; ExoN, 3′-to-5′ exonuclease; 2′-*O*-MT, S-adenosylmethionine-dependent ribose 2′-*O*-methyltransferase. Domains Ac, X, and Y are described in Refs [33, 35]. The mRNA cap-1 methyltransferase (in nsp 13) has been described [37].

[d] Indicated are the SARS-CoV proteinases predicted to be involved in cleavage of the N- and/or C-termini of the cleavage products; TI, translation initiation; TT, translation termination; RFS, ORF1a/ORF1b ribosomal frameshift.

interactions with the peptidyl-tRNA and aminoacyl-tRNA on the ribosome. Figure 7A illustrates the SARS virus mRNA sequence with the codons lined up opposite the anticodons for leucine (L) and asparagine (N) when the normal reading frame is used, while Figure 7B illustrates the position when a –1 frameshift occurs. In Figure 7C the ribosome continues translation in the –1 reading frame, generating a polypeptide ~7,000 nt in length.

Initial annotation of this large ORF identified the RdRp (POL) gene, a putative RNA helicase and two proteases, a chymotrypsin-like protease $3CL^{pro}$ and a papain-like protease PL^{pro} [11, 12]. Gorbalenya and coworkers have been studying the coronavirus replicase ORFs for several years [30–32] and have recently provided an extensive analysis of the SARS-CoV 1a and 1b ORFs [16] (Tab. 2).

These very large “replicase” proteins are processed autocatalytically by two or three viral proteinases encoded within ORF 1a [33]. In all other

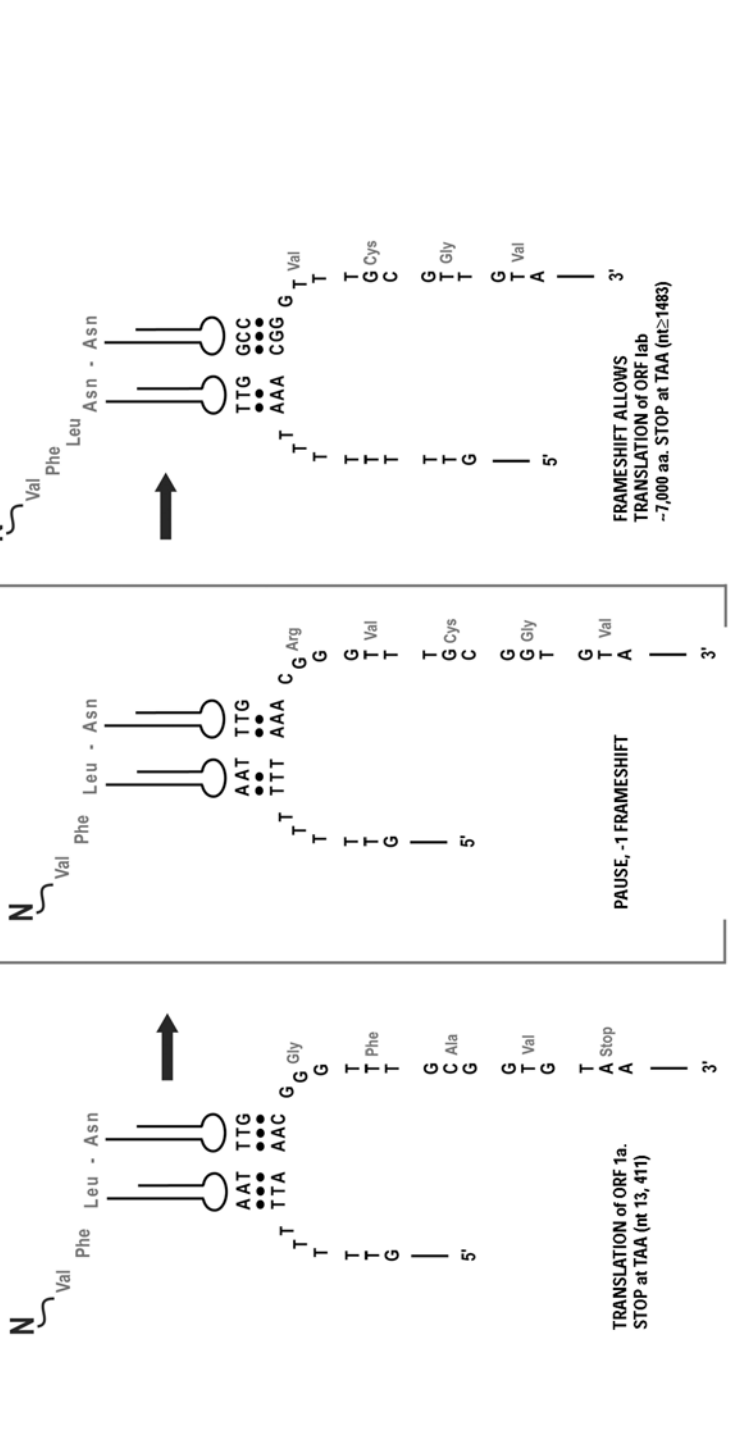


Figure 7. Ribosomal frameshifting within ORF1a and b of the SARS-CoV. Orf1a and ORF1b are large open reading frames encoding as many as 16 non-structural proteins. Translation of the ORF1b proteins is believed to occur due to a –1 frameshift which occurs on the ribosome. There is a pseudoknot within the RNA immediately downstream of the frameshift site. Translation of the RNA is believed to pause near the end of ORF1a and the RNA is shifted back one nucleotide, shifting to a different reading frame (see text for details).

coronaviruses three proteinases, a 3CL^{pro}, PL1^{pro} and PL2 ^{pro} function in cleaving these polyproteins [34, 35]. However, Rota et al. [12] and more recently Snijder et al. [16] proposed that the PL1^{pro} has been lost during evolution such that only the 3CL^{pro} and PL2^{pro} virally encoded proteinases function in processing of the SARS-CoV replicase polyprotein. The putative recognition sites within the 1a and 1b regions are summarized in Table 2. The PL2^{pro} is predicted to autocatalytically cut at the first three arrowheads near the N-terminal end of the polyprotein, while the 3CL^{pro} is predicted to cut at all other sites. If all consensus sites are recognized by these enzymes then the 1a and 1ab polyproteins will generate 16 proteins ranging in size from 18 amino acids up to 1922 amino acids. The sizes and putative function of these proteins is summarized in Table 2 (modified from Snijder et al. [16]). Also, each of these proteins is assigned a number, in order, designated as nsp (nonstructural protein) 1 to 16. Some of these proteins would be translated from both the 1a and the 1ab polypeptides, while others would be translated exclusively from the 1b region on the full-length frame-shifted transcript.

The replicase subunits are thought to exist as a large membrane-associated complex including at least the RdRp, RNA helicase, and several putative RNA processing enzymes such as Poly(U) specific endoribonuclease (XendoU,nsp 15), a 3' to 5' exonuclease (ExoN,nsp 14), S-adenosylmethionine-dependent ribose 2'O-methyltransferase (2'-O-MT,nsp 16,), adenosine diphosphate-ribose 1''-phosphatase (ADRP, nsp3) (Tab. 2). Several other proteins are predicted to be membrane associated and may facilitate formation of the membrane-associate replicase complex. One can imagine a large complex that synthesizes discontinuously the subgenomic minus strands and then transcribes these into the functional mRNAs.

Two functions that seem not to have been identified are a polyA polymerase activity and 5' capping activity. Coronavirus RNAs are known to be both 5' capped and 3' polyadenylated. It is probable that the genomic plus strands that have a polyA tail are transcribed into subgenomic minus strands and when these are converted to mRNAs, the polyA tail is added by copying the template strand. Presumably loss of the polyA tract would destine that genome to be lost from the gene pool. How the 5' cap structure is formed is unclear. There is a consensus sequence for a mRNA cap-1 methyltransferase [36] and a 2'-O-MT [16] which may play some role in 5' cap formation but so far a nucleotidyl transferase activity has not been identified in the viral genome.

Since the large ORF 1 polyproteins must be cleaved to generate the functional enzymes needed for virus replication, there is much interest in the viral protease as targets for the development of antivirals. A crystal structure of human coronavirus 229E 3CL^{pro} [37] was published about the same time the SARS-CoV genome sequence became available. Based on the 229E 3-D structure for the 3CL^{pro}, Anand et al. [37] were able to model the SARS-CoV 3CL^{pro} and a few months later a 3D structure for this protease (both

the free protein and complexed with a peptide inhibitor) appeared [38], providing the basis for drug design. The active protease is believed to be a dimer. Also, Fan et al. [39] have expressed the $3CL^{pro}$ in bacterial cells and provided evidence that the active protease is a dimer. Substrate specificity was analyzed using a series of 11 oligopeptides corresponding to the 11 predicted cleavage sites within the ORF1a and 1b regions (Tab. 2).

Another ORF 1a/1b polypeptide of unknown function, nsp9, has also been expressed, purified and crystallized [40].

ORF 2 – spike

The next largest ORF is ORF2 encoding the spike or S protein. The S protein is predicted to be 1255 amino acids in length and like other coronavirus S proteins it is likely heavily glycosylated. Mutations in this gene have previously correlated with altered pathogenesis and virulence in other coronaviruses [14]. Bioinformatic analysis of the SARS S protein using SignalP [41] showed a likely signal peptide at the N terminus (13–14 amino acids) and TMHMM [42] identified a strong transmembrane domain near the C-terminus, consistent with this protein being a type I membrane protein embedded in the viral envelope with most of the protein exposed on the surface. Proteomic analysis by Krokhin et al. [43] has also confirmed twelve glycosylation sites on this protein and the attached sugars for 4 of these sites have already been identified. Many of the SARS structural proteins appear to be membrane-associated proteins. A model of the predicted membrane-associated nature and orientation of these proteins is shown in Figure 8 (modified from http://athena.bioc.uvic.ca/sars/map/diagram-main.html/). It is believed that three molecules of the coronavirus S proteins form the characteristic peplomers or corona-like structure of this viral family. For some coronaviruses, specific regions of the S protein that bind to cellular receptors have been identified. Also it is known that various coronaviruses use different cellular receptors to mediate entry of the virus into host cells, hence it will be important to establish both the cellular receptor for the SARS-CoV and the viral attachment site on the S protein as possible antiviral targets for this new pathogen.

In several coronaviruses it is known that the S protein is cleaved into an N-terminal S1 subunit and C-terminal S2 subunit either by a viral or host protease [33], however it is not expected that the SARS-CoV S protein is similarly cleaved. Rota et al. [12] noted that the basic amino acid cleavage sites found in group 2 and group 3 coronaviruses (RRFRR, RRSRR, RSRR, RARS, and RARR) are not present in the SARS S protein. The S glycoprotein for the Sars-CoV shows a remarkably low sequence homology with other coronavirus S proteins. This low homology is especially notable in the N-terminal 700 amino acids of the protein. Spiga et al. [44] have presented a molecular modeling analysis of the S1 and S2 regions of

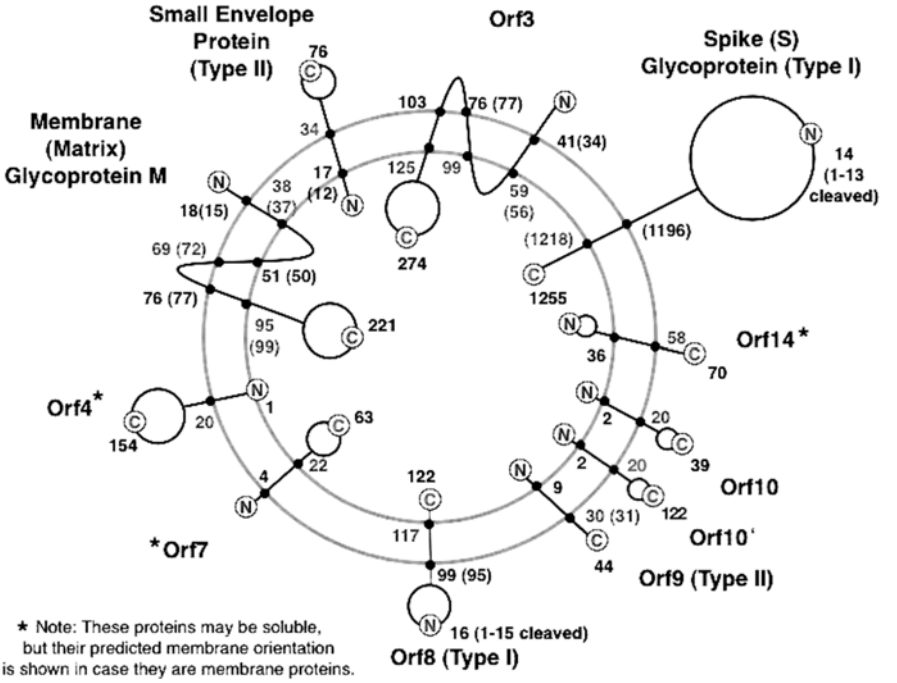


Figure 8. Predicted membrane association and orientation of "structural" proteins encoded by the SARS-CoV. Major coronavirus proteins such the spike (S), membrane glycoprotein (M) and small membrane (E) proteins are known to be part of the viral envelope. The remainder of the proteins may be associated with the viral envelope or may be associated with cellular membranes. For simplicity all the proteins are illustrated as being associated with a hypothetical circular membrane structure. Modified from http://athena.bioc.uvic.ca/sars/ (with permission of R. Roper and C. Upton).

the SARS-CoV S protein using the crystal structure of *Clostridium botulinum* neurotoxin B protein. These molecular models predict the overall shape and surface hydrophobicity of the S1 and S2 subunits and until a crystal structure is available, these will be a good starting point for design of antiviral drugs.

ORF3 – unknown function

ORF3 encodes a 274 aa protein that lacks significant similarities to any known proteins (BLAST, FASTA, or PFAM [45]). SignalP analysis [41] suggests that there may be a signal peptide at the N-terminus and internally, TMPred [46] and TMHMM predict three trans-membrane domains with a

149 aa C-terminal domain inside the viral or cellular membrane (Fig. 8). This region may also contain an ATP-binding domain.

It is important to note that many of the ORFs in the right two thirds of the SARS-CoV genome are predicted to code for membrane-associated proteins. In Figure 8, for convenience, a circular membrane envelope is used to illustrate the predicted orientation of these putative proteins in a membrane. Certainly the E, M and S proteins are embedded in the viral envelope, however there is no evidence that the other proteins are associated with the viral envelope or embedded in host cell membranes.

ORF4 – unknown protein

ORF4 has the potential to encode a 154 aa protein which overlaps completely with ORF 3 and ORF 5, but in a different reading frame. There is no obvious upstream TRS for this ORF, however it may be translated from the mRNA for ORF 3 using internal initiation [11]. This suggestion is supported by the data of Thiel et al. [15] in which the 4.6 kb mRNA transcript for ORF 3 is predicted to be bicistronic (capable of expressing Thiel's 3a and 3b polypeptides). This putative protein appears to encode a single transmembrane helix suggesting that if it is expressed, the product may be a membrane-associated protein (Fig. 8).

ORF 5 – E protein

ORF 5 encodes the small membrane E protein. It is 76 amino acids in length and BLAST and FASTA analyses show significant matches to other coronavirus small membrane E proteins. Both SignalP and TMPred analyses reveal a single transmembrane helix, again suggesting that this is a membrane-associated protein. TMHMM predicts the protein is a type II membrane protein with the majority of the hydrophilic domain and C terminus on the surface of the viral envelope (Fig. 8).

ORF 6 – M protein

The membrane glycoprotein or M protein is encoded by ORF6. This 221 aa protein is related to other coronavirus membrane or matrix glycoproteins. It is believed that during assembly of progeny virions the RNA-nucleocapsid complex associates non-covalently with the M protein embedded in the membranes of the ER, resulting in viral particles budding into the lumen of the ER. The virus then migrates through the Golgi complex and exits the cell, likely by exocytosis [14]. Signal P predicts that the M protein has a signal peptide that is likely not cleaved, while TMHMM and TMpred analyses

suggest that there are three transmembrane domains (aa 15–37, 50–72, and 77–00). The 121 aa hydrophilic domain at the C-teminus of the protein is predicted to be inside the viral particles (Fig. 8) where it is expected to interact with the nucleocapsid protein.

ORF 7 – unknown protein

ORF 7 likely encodes yet another putative membrane protein. Although TMHMM and SignalP do not predict a transmembrane helix, TMPred does. The transmembrane helix is predicted to be between residues 3 and 22, with the N-terminus on the outside of the viral membrane (Fig. 8). The possible function of this 63 aa putative polypeptide is unknown as there are no significant matches using BLAST and FASTA analyses.

ORF 8 – unknown protein

ORF 8 is another ORF that may code for a protein of 122 amino acids. Again BLAST and FASTA searches failed to find significant homology with known proteins and again TMPred and TMHMM analyses predict that there is a single transmembrane domain near the C-terminus. SignalP indicates a signal sequence at the N-terminus that is likely cleaved between residues 15 and 16. However, there is no evidence yet to indicate that this protein is expressed in infected cells.

ORF 9 – unknown protein

The putative protein encoded by ORF 9 is only 44 amino acids in length (Rota et al. [12] used a cutoff of 50 aa, hence they did not include this protein in their analysis) (Tab. 1). Using FASTA there are some weak similarities with a putative sterol-C5 desaturase and a *Clostridium perfringens* protein (SWISS-PROT Q9M883 and CPE2366, respectively). Again TMPred predicts a strong transmembrane helix with no real preference for the orientation of the protein in the membrane. In Figure 8 this protein has been drawn it with the N terminus inside the particle and the C terminus outside.

ORF 10 – unknown protein

ORF 10 is interesting as it is only a 39 aa protein, yet there is a strong TRS sequence upstream of this ORF [11]. From studies on animal SARS virus-

es, ORF 10 and 11 proteins would actually be fused due to an additional 29 nt in these genomes. Hence in these animal SARS virus isolates as well as one human isolate (GZ01) these additional 29 nts result in an ORF that would encode a 122 aa protein (see ORF 10' below) [19]. The small ORF 10 protein is predicted to be another putative transmembrane protein with the N terminus located on the inside of the membrane and the C terminus outside (Fig. 8).

ORF 11 – unknown protein

ORF11 is predicted to encode an 84 aa protein. It contains limited homology to the human E2 glycoprotein precursor. SignalP and TMHMM predict that this protein, if expressed, is likely soluble.

ORF 10' – unknown protein

As discussed above all animal isolates contain an additional 29 nt [19] which predict that the ORF 10 and ORF 11 protein would not be expressed, but rather a fusion protein of 122 amino acid, ORF 10' would be expressed. This protein is also likely a transmembrane protein (see Fig. 8) with the N terminus inside the membrane.

ORF 12 – Nucleocapsid protein

The nucleocapsid gene encodes a 422 aa protein that has significant homology to other coronavirus N proteins. It does however have a short basic region (KTFPPTEPKKDKKKKTDEAQ) that appears to be unique to the SARS-CoV. Further analysis suggested that this region is part of a bipartite nuclear localization signal [11]. This motif may indicate that the SARS-CoV N protein is capable of being transported to the nucleus where it may play some novel role in pathogenesis of the virus. The nucleocapsid protein is known to associate with the viral RNA to form the RNA-protein complex that interacts with the M protein initiating encapsidation of the particle as it passes into the lumen of the ER [14]. Of interest is that a protein (~46 kDa) found in serum from convalescent SARS patients has been characterized by MALDI_TOF MS to be the SARS N protein [43]. Mass spectrometry has also been applied to the S protein allowing confirmation of 12 glycosylation sites in the protein [43].

The final two ORFs (13 and 14) do not have a TRS upstream so if they are expressed, they would likely be translated from the mRNA encoding the N protein.

ORF 13 – unknown protein

This ORF has the potential to encode a protein of 98 aa. There are no transmembrane helices detected and BLAST shows no homology to known proteins.

ORF 14 – unknown protein

The ORF 14 protein also has no known homologies using BLAST and TMPred predicts a single transmembrane helix, hence this protein, if expressed, is likely a membrane associated protein. However like the ORF4 protein the prediction of the membrane location is not strong. If embedded in the membrane it is expected that the N-terminus would be oriented toward the inside of the membrane.

The role of these minor proteins (ORFs 3, 4, 7, 8, 9, 10, 11, 13, and 14) will need further study to evaluate their role, if any, in the SARS-CoV replication cycle. Unpublished preliminary data indicates that at least one of these proteins is expressed.

Future direction for SARS-CoV research

It is quite remarkable that, in the one year since the first clinical reports of a new emerging virus, so much has been learned about the SARS-CoV. In late March 2003 the virus was identified as probably a new coronavirus and within two weeks its entire genome was sequenced [11, 12]. The speed with which this was accomplished is unprecedented. It is also remarkable that more than 1300 papers have been published on SARS in the past year (PubMed). In comparison, it took several years before the human immunodeficiency virus (HIV) was identified in 1983-4 as the agent responsible for a newly described immunodeficiency disease, AIDS [14]. Characterization of HIV has led to the development of a number of drugs that slow the progression of AIDS and multi-drug therapies ("highly active antiretroviral therapy" or HAART) have changed HIV infection from a terminal disease to a chronic one. However, 20 years after the discovery of HIV, no vaccine is available, a reflection of the remarkably intractable task of developing an HIV vaccine.

In contrast, development of improved tests for the SARS-CoV RNA as well as serological test to detect antibodies to viral proteins were well underway very shortly after this viral genome sequence became available [47]. Amazingly, a SARS chip was available by June 2003 [48]. Also, structural information about the main protease $3CL^{pro}$ [38], nsp9 protein [40], E protein [49] and S protein [44] (in the first two cases crystallographic deter-

mination of the structures and in the latter two cases, molecular models) is allowing preliminary, directed development of anti-SARS drugs to begin. In addition, small inhibitory RNA (siRNA) therapy may prove to be effective in treating this viral infection and again preliminary studies have begun to evaluate potential siRNAs [50].

Another area that will be a focus during the next few years is the development of a SARS vaccine. The success of vaccines for avian, bovine, feline, and porcine coronaviruses bodes well for the development of a vaccine against the SARS-CoV [51] and a report of a recent conference in Switzerland indicates preliminary progress from a number of laboratories towards developing a vaccine directed against the S, E and N proteins using a variety of inactivated virus, recombinant vectors, recombinant antigens, as well as DNA-based vaccines (http://www.who.int/vaccine_research/diseases/sars/events/2003/11/en/).

Yet further reason for optimism is the very recent discovery by Li et al. [52] that unlike other coronaviruses, the SARS-CoV uses a receptor on cells – the angiotensin-converting enzyme 2 or ACE2 – and drugs that block this receptor already have been developed for other conditions. The ACE2 inhibitors have been tested in humans, and some appear to be safe, but they have not yet been commercialized [52, 53]. In addition, the other half of the interaction, the receptor binding domain on the S protein, has been localized to residues 303 to 537 of the S protein [53]. Also learning from the experience with HIV virus in which soluble receptor protein (sCD4) proved ineffective in blocking HIV infection while a multivalent CD4-IgG fusion protein does seem to be effective, it has been speculated that soluble ACE2-IgG fusions may prove useful in treating SARS infections [53].

One aspect of the SARS viral genome lends itself to speculation about the origin of the SARS-CoV. While the natural host has not been identified, if appears that the virus spread from civet cats into humans. Since the civet cat virus and only one human isolate of the SARS-CoV contains the additional 29 nt in the ORF 10/11 region, it is tempting to speculate that transfer of the virus to humans was accompanied by a deletion of 29 nt. If this is the case, the characterization and function of ORF 10, ORF11 and ORF 10' may provide considerable insight into why the SARS-CoV is so pathogenic in humans.

Acknowledgements

This work was supported by funds from Genome British Columbia and Genome Canada.

The GSC high throughput large scale sequencing and bioinformatics pipelines are also supported by the BC Cancer Foundation of the BCCA, GBC/GC, NSERC, CIHR, NHGRI, NCI, and Western Diversification.

MAM and SJMJ are Scholars of the Michael Smith Foundation for Health Research.

We wish to acknowledge all those who contributed to the rapid sequencing of the Tor2 SARS genome. At the GSC these include A. Brooks-Wilson, Y. Butterfield, J. Khattra, J. Asano, S. Barber, S. Chan, A. Cloutier, S. Coughlin, D. Freeman, N. Girn, O. Griffiths, S. Leach, M. Mayo, H. McDonald, S. Montgomery, P. Pandoh, A. Petrescu, G. Robertson, J. Schein, A. Siddiqui, D. Smailus, J. Stott, and G. Yang. Our collaborators at the NML were F. Plummer, A. Antonov, H. Artsob, N. Bastien, K. Bernard, T. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Lio, S. Normand, U. Stoher, G. Tipples, S. Tyler, R. Vogrig, D. Ward and B. Watson. Our collaborators at the BC CDC were R. Brunham, M. Krajden, D. Skowronski, and M. Petric and a the U. Victoria they were C. Upton and R. Roper.

This review summarizes research published to Dec 31, 2003. Due to space limitations not all papers could be cited. Two accidental laboratory derived infections (Senior, K.) [55, 56] and only one new confirmed case of SARS have occurred since June 2003. The new community acquired case occurred in Guangdong province in December 2003. It appears that surveillance and quarantine procedures in place helped to identify and isolate this case preventing further transmission to the patient's contacts.

References

1 Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, Nicholls J, Yee WK, Yan WW, Cheung MT (2003) Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* 361: 1319–1325

2 Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim XX et al (2003) A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* 348: 1848–1951

3 Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA et al (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N Engl J Med* 348(20): 1967–1976

4 Wang D, Urisman A, Lui Y-T, Springer M, Ksiazek TG, Erdman DD, Mardis ER, Hickenbotham M et al (2003) Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* 1: e2 (2003)

5 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, (2001) The sequence of the human genome. *Science* 291: 1304–1351

6 Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194

7 Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8: 195–202
8 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410
9 Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. *Science* 227: 1435–1441
10 Durbin R,Thierry J Mieg (1991) A *C. elegans* database documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov
11 Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, Khattra J, Asano JK, Barber SA, Chan SY et al (2003) The genome sequence of the SARS-associated coronavirus. *Science* 300: 1399–1404 (published online 1 May, 2003)
12 Rota PA, Oberste MS, Monroe SS, Nix WA, Campagnoli R, Icenogle JP, Penaranda S, Bankamp B, Maher K, Chen MH et al (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394–1399 (published online, 1 May, 2003)
13 Ruan YJ, Wei CL, Ee AL, Vega VB, Thoreau H, Su ST, Chia JM, Ng P, Chiu KP, Lim L et al (2003) Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* 361: 1779–1785
14 Fields BN, Knipe DM, Howley PM, Griffin DE (eds) (2001) *Fields Virology.* Lippincott Williams and Wilkins, Philadelphia
15 Thiel V, Ivanov KA, Putics A, Hertzig T, Schelle B, Bayer S, Weissbrich B, Snijder EJ, Rabenau H, Doerr HW et al (2003) Mechanisms and enzymes involved in SARS coronavirus genome expression. *J Gen Virol* 84: 2305–2315
16 Snijder, EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE (2003) Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J Mol Biol* 331: 991–1004
17 Rest JS, Mindell DP (2003) SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. Infect. Genet. and Evol. 3: 219–225
18 Stavrinides J, Guttman DS (2004) Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J Virol* 78: 76–82
19 Guan Y, Zheng BJ, He YQ, Liu ZX, Zhuang CL, Cheung CL, Luo SW, Li PH, Zhang LJ, Guan YJ et al (2003) Isolation and characterization of viruses related to the SARS coronavirus from Southern China. *Science* 5643: 276–278
20 Lai MMC, Cavanagh D (1997) The molecular biology of coronaviruses. *Adv Virus Res* 48: 1–100
21 Sawicki SG, Sawicki DL (1998) A new model for coronavirus transcription. *Adv Exp Med Biol* 440: 215–220
22 Lai MMC, Holmes KV (2001) *Coronaviridae*: the viruses and their replication. In: Fields BN, Knipe DM, Howley PM, Griffin DE (eds) (2001) *Fields Virology.* Lippincott Williams and Wilkins, Philadelphia, 1163–1185

23 Siddell SG (1995) The *Coronaviridae*. Plenum Press, New York
24 Jonassen CM, Jonassen TO, Grinde BJ (1998) A common RNA motif in the 3' end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. *Gen Virol* 79: 715–718
25 Woo K, Joo M, Narayanan K, Kim KH, Makino S (1997) Identification and characterization of a coronavirus packaging signal. *J Virol* 71: 824–827
26 Qin L, Xiong B, Luo C, Guo Z-M, Hao P, Su J, Nan P, Feng Y, Shi Y-X, Yu X-J et al (2003) Identification of probable genomic packaging signal sequence from SARS-CoV genome by bioinformatics analysis. *Acta Pharmacologica Sinica* 24: 489–496
27 Matthews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288: 911–940
28 Brierley I, Digard P, Inglis SC (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. *Cell* 57: 537–547 (1989)
29 Geidroc DP, Theimer CA, Nixon PL (2000) Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol* 298: 167–185
30 Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1989) Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. *Nucl Acids Res* 17: 4847–4861
31 Gorbalenya AE (2001) Big nidovirus genome. When count and order of domains matter. *Adv Exp Biol Med* 494: 1–17
32 Ziebuhr J, Thiel V, Gorbalenya AE (2001) The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. *J Biol Chem* 276: 33220–33232
33 Zeibuhr J, Snijder EJ, Gorbalenya AE (2000) Virus-encoded proteinases and proteolytic processing in the *Nidovirales*. *J Gen Virol* 82: 853–879
34 Gorbalenya AE, Koonin EV, Lai MM (1991) Putative papain-related thiol proteases of positive-strand RNA viruses. Identification of rubi- and aphthovirus proteases and delineation of a novel conserved domain associated with proteases of rubi-, alpha- and coronaviruses. *FEBS Letts* 288: 201–205
35 Baker SC, Yokomori K, Dong S, Carlisle R, Gorbalenya AE, Koonin EV, Lai MM (1993) Identification of the catalytic sites of a papain-like custeine proteinase of murione coronaviruses. *J Virol* 67: 6056–6063
36 von Grotthuss M, Wyrwicz LS, Rychlewski L (2003) mRNA cap-1 methytransferase in the SARS genome. *Cell* 113: 701–702
37 Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R (2003) Coronovirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 300: 1763–7
38 Yang H, Yang M, Ding Y, Liu Y, Lou Z, Zhou Z, Sun L, Mo L, Ye S, Pang H (2003) The crystal structure of severe acute respiratory syndrome virus main

protease and its complex with an inhibitor. *Proc Natl Acad Sci USA* 100: 13190–13195

39 Fan K, Wei, P, Feng, Q, Chen S, Huang C, Ma L, Lai B, Pei J, Liu Y, Chen J et al (2004) Biosynthesis, purification and substrate specificity of SARS coronavirus 3C-like proteinase. *J Biol Chem* 279: 1637–1642

40 Campanacci V, Egloff MP, Longhi S, Ferron F, Rancurel C, Salomoni A, Durousseau C, Tocque F, Bremond N, Dobbe JC et al (2003) Structural genomics of the SARS coronavirus: cloning, expression, crystallization and preliminary crystallographic study of the Nsp9 protein. *Acta Crystallogr D Biol Crystallogr* 59: 1628–1631

41 Nielson H, Engelbrecht S, Brunak, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10: 1–6

42 Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6: 175–182

43 Krokhin O, Li Y, Andonov AQ, Feldman H, Flick R, Jones S, Stroeher U, Bastien U, Dasuri KV, Cheng K et al (2003) Mass spectrometric characterization of proteins from the SARS Virus: A preliminary report. *Mol Cell Proteomics* 2: 346–356

44 Spiga O, Bernini A, Ciutti A, Chiellini S, Menciassi N, Finetti F, Caurarono V, Anselmi F, Prisch F, Niccolai N (2003) Molecular modelling of S1 and S2 subunits of SARS coronavirus spike glycoprotein. *Biochem Biophys Res Commun* 310: 78–83

45 Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. *Nucl Acids Res* 30: 276–280

46 Hoffman K, Stoffel W (1993) *Biol Chem Hoppe-Seyler* 374: 166

47 Basu P (2003) Biotech firms jump on SARS bandwagon. *Nat Biotech* 21: 720

48 Frankish H (2003) SARS genome chip available to scientists: the NIAID hopes that widespread access to the SARS genome chip will catalyse research into effective treatments for the virus. *Lancet* 361: 2212

49 Shen X, Xue JH, Yu CY, Luo HB, Qin L, Yu XJ, Chen J, Chen LL, Xiong B, Yue LD et al (2003) Small envelope protein E of SARS: cloning, expression, purification, CD determination, and bioinformatics analysis. *Acta Pharmacol Sin* 24: 505–511

50 Zhang R, Guo Z, Lu J, Meng J, Zhou C, Zhan X, Huang B, Yu X, Huang M, Pan X et al (2003) Inhibiting severe acute respiratory syndrome-associated coronavirus by small interfering RNAs (siRNAs). *Chin Med J* 116: 1262–1265

51 Cavanagh D (2003) Several acute respiratory syndrome vaccine development: experiences of vaccination against avian infectious bronchitis coronavirus. Avian Pathol. 32: 567–582

52 Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, Somasundaran M, Sullivan JL, Luzuriaga K, Greenough TC et al (2003) Angiotensin-converting

enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 426: 450–454

53 Dimitrov DS (2003) The secret life of ACE2 as a receptor for the SARS virus. *Cell* 652–653

54 Xiao X, Chakraboti S, Dimitrov AS, Gramatikoff K, Dimitrov DS (2003) *Biochem Biophys Res Commun* 312: 1394–1399

55 Senior K (2003) Recent Singapore SARS case a laboratory accident. *Lancet Infec Dis* 3: 679

56 Lim PL, Kurup A, Gopalakrishna G, Chang KP, Wong CW, Ng LC, Se-Thoe SY, Oon L, Bai X, Stanton LW et al (2004) Laboratory-acquired severe acute respiratory syndrome. *New Engl J Med* 350: 1740–1745