Chinese Science Bulletin 2003 Vol. 48 No.10 941—948

A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01)

QIN E'de, ZHU Qingyu*, YU Man, FAN Baochang, CHANG Guohui, SI Bingyin, YANG Bao'an, PENG Wenming, JIANG Tao, LIU Bohua, DENG Yongqiang, LIU Hong, ZHANG Yu, WANG Cui'e, LI Yuquan, GAN Yonghua, LI Xiaoyu, LÜ Fushuang, TAN Gang, CAO Wuchun, YANG Ruifu

Institute of Microbiology and Epidemiology, Chinese Academy of Military Medical Sciences, Beijing 100071, China

WANG Jian, LI Wei, XU Zuyuan, LI Yan, WU Qingfa, LIN Wei, CHENG Weijun, TANG Lin, DENG Yajun, HAN Yujun, LI Changfeng, LEI Meng, LI Guoqing, LI Wenjie, LÜ Hong, SHI Jianping, TONG Zongzhong, ZHANG Feng, LI Songgang, LIU Bin, LIU Siqi, DONG Wei, WANG Jun, Gane K-S Wong, YU Jun, YANG Huanming*

Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300; National Center for Genome Information, Beijing 101300, China

* Correspondence should be addressed to Zhu Qingyu (e-mail: zhuqy@nic.bmi.ac.cn) and Yang Huanming (e-mail: yanghm@genomics.org.cn).

Abstract **The genome sequence of the Severe Acute Respiratory Syndrome (SARS)-associated virus provides essential information for the identification of pathogen(s), exploration of etiology and evolution, interpretation of transmission and pathogenesis, development of diagnostics, prevention by future vaccination, and treatment by developing new drugs. We report the complete genome sequence and comparative analysis of an isolate (BJ01) of the coronavirus that has been recognized as a pathogen for SARS. The genome is 29725 nt in size and has 11 ORFs (Open Reading Frames). It is composed of a stable region encoding an RNA-dependent RNA polymerase (composed of 2 ORFs) and a variable region representing 4 CDSs (coding sequences) for viral structural genes (the S, E, M, N proteins) and 5 PUPs (putative uncharacterized proteins). Its gene order is identical to that of other known coronaviruses. The sequence alignment with all known RNA viruses places this virus as a member in the family of Coronaviridae. Thirty putative substitutions have been identified by comparative analysis of the 5 SARS-associated virus genome sequences in GenBank. Fifteen of them lead to possible amino acid changes (non-synonymous mutations) in the proteins. Three amino acid changes, with predicted alteration of physical and chemical features, have been detected in the S protein that is postulated to be involved in the immunoreactions between the virus and its host. Two amino acid changes have been detected in the M protein, which could be related to viral envelope formation. Phylogenetic analysis suggests the possibility of non-human origin of the SARS-associated viruses but provides no evidence that they are man-made. Further efforts should focus on identifying the etiology of the SARS- associated virus and ruling out conclusively the existence of other possible SARS-related pathogen(s).**

Keywords: Severe Acute Respiratory Syndrome (SARS), coronavirus, genome, phylogeny.

DOI: 10.1360/03wc0186

Severe Acute Respiratory Syndrome (SARS) is a newly identified infectious disease[1−5]. The global outbreak of SARS has been threatening the health of people worldwide and has killed 353 people and infected more than 5462 in 27 countries, as reported by WHO on April 29, 2003 (http://www.who.int/csr/sarscountry/en). Although it has been recognized that a variant of virus from the family of coronavirus might be the candidate pathogen of SARS[1−5], its identity as the unique pathogen still remains controversial (http://www.nytimes.com/2003/04/24/science/24INFN.html; http://www.sciencenews.org/20030329/fob7.asp).

We have sequenced an isolate of the SARS-associated virus. In this paper, we report the complete sequence of the virus genome and the results of a comparative analysis of all 5 coronavirus genome sequences published to date.

1 Materials and methods

(ⅰ) Source of samples. Between February 10 and March 15, 2003, 10 cases of suspected SARS were identified in both Guangzhou and Beijing. All patients were diagnosed according to WHO guidelines (http://www.who.int/csr/sars/guidelines/en/). Two patients died from respiratory failures and the remaining 8 recovered. Two samples of autopsied lung tissue were obtained from the deceased patients. The samples were cultured with a variety of cell lines, including Vero-E6, MDCK, Hep-2, Hela, BHK-21 and LLC-MK-2[6]. Four isolates were prepared from Vero-E6 cell culture. Viral cDNA were cloned by RT-PCR according to conventional protocols.

(ⅱ) Sequencing, gap-closure, and assembly. Sequencing was performed by using MegaBACE 1000 (Amersham). Base calling was performed by Phred (http://www.phrap.org). Contaminations from human and other resources were removed by CrossMatch and the complete sequence was assembled by using Phrap (http://

Abbreviations: a.a.: amino acid; CDS: coding sequences; dsRNA: double-stranded RNA; HCV: coronavirus; HE: hemagglutinin-esterase; kb: kilobase; kD: kilodalton; MW: molecular weight: nt: nucleotide; pI: isoelectric point; PUP: putative uncharacterized protein; ssRNA: single-stranded RNA; UTR: untranslated region.

www.phrap.org). The gaps, as well as the regions with low quality data identified after preliminary assembly, were filled in or refined by resequencing the PCR products with ABI 377 sequencers (Applied Biosystems).

(iii) Sequence annotation and comparative analysis. The composition of the nucleotide sequence was analyzed by DNA_GC_Calculator (http://www.genome.iastate.edu/ftp/share/DNAgcCal/). The ORFs were identified by using ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Comparative analysis was performed by using Blast against the nr (non-redundant) database (http://www.ncbi.nlm.nih.gov/blast/) for nucleic acid and protein sequences, and the multiple sequence alignment was deployed by using ClustalW1.8 (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW). Putative glycosylation sites in the proteins were examined by using NetNGlyc1.0 (http://www.cbs.dtu.dk/services/NetNGlyc). The hypothesized physical and chemical features of the putative proteins were examined by using Compute PI/MW (http://us.expasy.org/tools/pi_tool.html). All above bioin-formatic analyses were performed on supercomputers DOWNING 2000, DOWNING 3000 (DOWNING Computers Inc.), SUN E10K (SUN Microsystems Inc.), SGI Origin 3800 (Silicon Graphics, Inc.) and IBM p690 (IBM Corp.).

2 Results

There were 207892 nt raw data generated from 549 sequencing reads, equivalent to approximately 7 X coverage of the virus genome according to the estimated size. After gap-closure and improvement of the low quality regions, a complete sequence of 29725 nt with an overall error rate of 0.0093% was obtained. The sequence data have been deposited in GenBank (Accession No. AY278488) and are freely available. All the clones with known sequences are available for collaborators worldwide.

(i) Genome landscape. The genome landscape of the virus that we sequenced (Isolate BJ01) has the representative characteristics of a linear and positive single-stranded RNA molecule, flanked by a 5′-methylated cap and a 3′-poly(A)$^+$ tail. It has a GC content of 40.8% (A : U : C : G = 28.5 : 30.7 : 20.0 : 20.8) with a relatively even distribution over the entire genome (Fig. 1). However, two obvious GC-rich regions were observed at the nucleotide position 200—500 and 28200—28767 near the two termini, corresponding to the ORFs for the RNA-dependent RNA polymerase (orf 1a) and the N (nucleocapsid) protein partially overlapped with PUP5. Several repetitive sequences, which could be related to the secondary structure of a single-stranded RNA molecule, were observed. The 5′terminal sequence was predicted to have a typical secondary structure formed by a long repeat, i.e. the region (nt position 155—211) is highly homologous to the complementary sequence of a 60 nt one at the nucleotide position 861—920 (Fig. 2). At least 4 reversed repeates, with a minimal number of 7 nt in a subunit, were identified (Fig. 3). Approximately 140 segments were postulated to contribute to the possible palindrome structures of the genome in the natural status (See Supplementary Data, http://www. genomics.org.cn/SARS).

(ii) Predicted ORFs and their putative proteins. Eleven ORFs (6 CDSs and 5 PUPs) were predicted with a total length of 28550 bp, accounted for approximately 96% of the whole genome (Table 1 and Fig. 4). In addition to orf 1ab, overlapped ORFs were found between PUP1 and PUP2, as well as between PUP2 and the E protein. The PUP5 (nt position 28111—28407) is embeded in the CDS for the N protein (nt position 28101—29369). Aside from 5′(245 nt) and 3′(355 nt) UTRs, there are only 6 intergenic regions with a total length of 560 nt, with 5 very small ones (6—50 nt) and one large one (478 nt, between PUP4 and the N protein). The available data are insufficient to confirm the reliability and possible function of PUPs identified by computation means. The observed order of CDSs is exactly the same as in other known HE(−) coronaviruses. Codon usage statistics showed that Leu (leusine) had the highest frequency (Table 2).

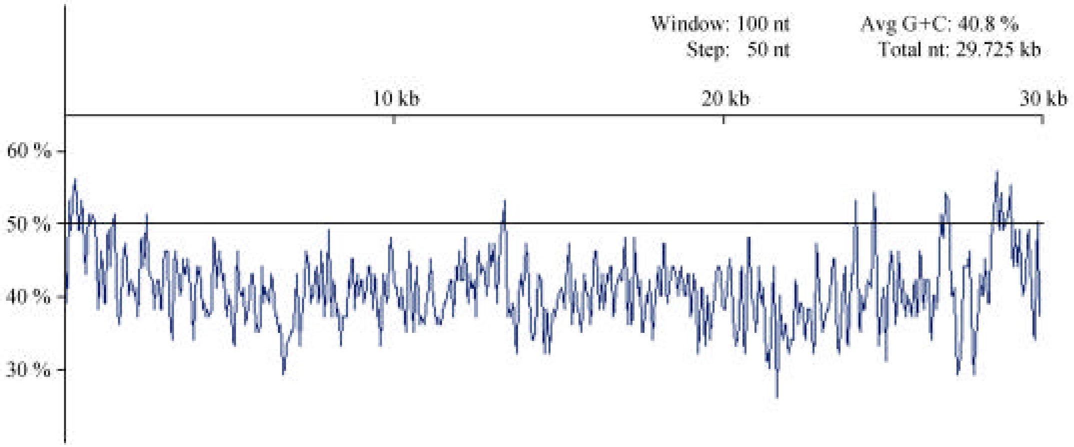

Fig. 1. The GC distribution in the SARS-associated virus genome (Isolate BJ01).

```
155 TCCCTCTTCTGCAGACTGCTTA----CGGTTTCGTCCGTGTTGCAGTCGATCATCAGCATA 211
    : ::::  : ::::: : ::::    : :::: : :: :::::  :: :  : : : :: :
861 TGCCTCAGCAGCAGATTTCTTAGTGACAGTTT-GGCCTTGTTGTTGTTGGCCTTTACCAGA 920
```

Fig. 2. A 5 ′terminal repeat identified in the SARS-associated virus genome (Isolate BJ01).

```
24     ATCTCTTGTAGATCTGTTCTCTA 46
       -------->     <--------

12311  TAACAACATTATCAACAAT 12329
       ------>     <------

14698  TATCAGTGATTATGACTAT 14716
       ------>     <------

26374  TTATCATGGCAGACAACGGTACTATT 26398
       ---------->    <----------
```

Fig. 3. The reversed repeats in the SARS-associated virus genome (Isolate BJ01).

The CDS for the RNA-dependent RNA polymerase (orf 1ab) is accounted for approximately 2/3 of the genome (nt position 246 — 21466). It is composed of 2 ORFs (orf 1a and orf 1b, nt position 246 — 13379 and 13379 — 21466), overlapped by a single nucleotide at the nucleotide position 13379. This result is consistent with the previously proposed "ribosomal frame shifting" mechanism that produces two peptides[7](Table 1).

The S protein (the spike protein, or E2 glycoprotein precursor) was predicted to be a weak acidic glycoprotein (pI 5.5) located immediately after the RNA polymerase (nt position 21473 — 25240). It is the largest protein of (1255 a.a.) among all viral structural proteins with an estimated molecular weight of 139.17 kD, a hydrophobic transmembrane domain, and at least 9 glycosylation sites were identified preliminarily.

The E protein (the small envelope protein) is located at the nucleotide position 26098 —26328. It was predicted to be the smallest protein (8.36 kD, 76 a.a.) with the highest hydrophobicity (47.40%) among all viral proteins,

Table 1 The predicted physical and chemical features of the predicted ORFs in the SARS-associated virus genome (Isolate BJ01)

ORF	Position (nt)	Peptide length (a.a.)	Estimated MW /kD	pI value	Hydrophobicity (%)	Hydrophilicity (%)	Charge (+) (%)	Charge (–) (%)
RNA polymerase	246—13379 13379—21466	7073	790.28	6.3	30.80	44.30	11.80	10.50
S protein	21473—25240	1255	139.17	5.5	30.40	44.80	9.10	9.20
E protein	26098—26328	76	8.36	6.0	47.40	32.90	5.30	5.30
M protein	26379—27044	221	25.06	9.3	40.70	36.20	10.90	5.90
N protein	28101—29369	422	46.03	10.1	17.30	54.00	15.40	8.50
PUP1	25249—26073	274	30.90	5.6	34.70	39.10	8.40	8.00
PUP2	25670—26134	154	17.72	11.0	37.00	51.90	19.50	0.60
PUP3	27055—27246	63	7.54	4.7	47.60	42.90	11.10	15.90
PUP4	27254—27622	122	13.94	8.3	33.60	42.60	13.10	8.20
PUP5	28111—28407	98	10.80	4.9	32.70	46.90	9.20	11.2

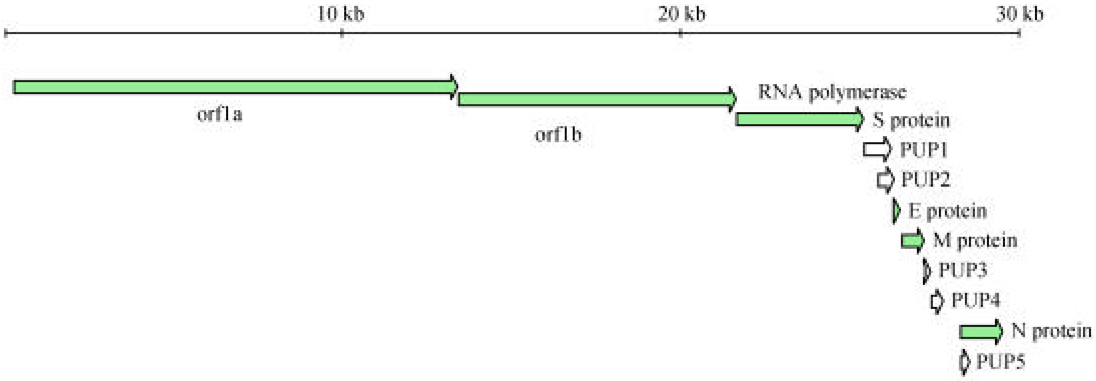

Fig. 4. The genome organization of the SARS-associated virus (Isolate BJ01). Unfilled arrows denote putative uncharacterized proteins (PUPs).

Table 2 The codon usage of the ORFs in the SARS-associated virus genome (Isolate BJ01)

RNA polymerase				S protein				E protein				M protein				N protein			
Leu	(L)	675	9.50%	Thr	(T)	100	8.00%	Val	(V)	14	18.40%	Leu	(L)	31	14.00%	Gly	(G)	45	10.70%
Val	(V)	579	8.20%	Leu	(L)	99	7.90%	Leu	(L)	14	18.40%	Ala	(A)	19	8.60%	Ser	(S)	35	8.30%
Ala	(A)	511	7.20%	Ser	(S)	96	7.60%	Ser	(S)	7	9.20%	Ile	(I)	18	8.10%	Gln	(Q)	34	8.10%
Thr	(T)	495	7.00%	Val	(V)	91	7.30%	Thr	(T)	5	6.60%	Val	(V)	16	7.20%	Ala	(A)	34	8.10%
Ser	(S)	458	6.50%	Ala	(A)	84	6.70%	Asn	(N)	5	6.60%	Gly	(G)	15	6.80%	Thr	(T)	33	7.80%
Lys	(K)	415	5.90%	Phe	(F)	83	6.60%	Tyr	(Y)	4	5.30%	Arg	(R)	15	6.80%	Pro	(P)	31	7.30%
Gly	(G)	419	5.90%	Asn	(N)	81	6.50%	Phe	(F)	4	5.30%	Thr	(T)	13	5.90%	Arg	(R)	31	7.30%
Asp	(D)	395	5.60%	Gly	(G)	78	6.20%	Ala	(A)	4	5.30%	Asn	(N)	13	5.90%	Lys	(K)	29	6.90%
Asn	(N)	366	5.20%	Ile	(I)	77	6.10%	Ile	(I)	3	3.90%	Ser	(S)	12	5.40%	Leu	(L)	26	6.20%
Glu	(E)	348	4.90%	Asp	(D)	74	5.90%	Glu	(E)	3	3.90%	Phe	(F)	11	5.00%	Asn	(N)	25	5.90%
Ile	(I)	343	4.80%	Lys	(K)	60	4.80%	Cys	(C)	3	3.90%	Tyr	(Y)	9	4.10%	Asp	(D)	22	5.20%
Phe	(F)	331	4.70%	Pro	(P)	57	4.50%	Pro	(P)	2	2.60%	Trp	(W)	7	3.20%	Glu	(E)	14	3.30%
Tyr	(Y)	324	4.60%	Gln	(Q)	55	4.40%	Lys	(K)	2	2.60%	Met	(M)	7	3.20%	Phe	(F)	13	3.10%
Pro	(P)	274	3.90%	Tyr	(Y)	54	4.30%	Gly	(G)	2	2.60%	Glu	(E)	7	3.20%	Val	(V)	11	2.60%
Arg	(R)	259	3.70%	Glu	(E)	42	3.30%	Arg	(R)	2	2.60%	Lys	(K)	6	2.70%	Tyr	(Y)	11	2.60%
Gln	(Q)	234	3.30%	Cys	(C)	39	3.10%	Met	(M)	1	1.30%	Asp	(D)	6	2.70%	Ile	(I)	11	2.60%
Cys	(C)	233	3.30%	Arg	(R)	39	3.10%	Asp	(D)	1	1.30%	Pro	(P)	5	2.30%	Met	(M)	7	1.70%
Met	(M)	177	2.50%	Met	(M)	20	1.60%	Trp	(W)	0	0.00%	Gln	(Q)	5	2.30%	Trp	(W)	5	1.20%
His	(H)	160	2.30%	His	(H)	15	1.20%	His	(H)	0	0.00%	His	(H)	3	1.40%	His	(H)	5	1.20%
Trp	(W)	77	1.10%	Trp	(W)	11	0.90%	Gln	(Q)	0	0.00%	Cys	(C)	3	1.40%	Cys	(C)	0	0.00%
Total		7073	100%			1255	100%			76	100%			221	100%			422	100%

and has balanced electric charges (5.30%) both positively and negatively. The consensus UCUAAAC element, located near the beginning of all other CDSs, but a mutant UCUACAC element was identified at –200 nt upstream of the E protein CDS.

The M protein (the membrane, matrix protein, or E1 membrane glycoprotein) was predicted to be a midium-sized protein (25.06 kD, 221 a.a.) with a high predicted pI value (pI 9.3). It is located at the nucleotide position 26379—27044.

The N protein (the nucleocapsid protein), predicted to be the second largest putative viral structural protein (46.03 kD, 422 a.a.) is located at the most 3′end. It is a typical basic protein with a high predicted pI value (pI 10.1), the highest hydrophilicity (54.00%), and the lowest hydrophobicity (17.30%) among all the proteins predicted from the viral sequence.

(iii) Variations and phylogenetics of the coronavirus genomes. It was possible to arbitrarily divide the whole virus genome into 2 regions according to the uneven distribution of substitutions detected in this study. The 5′ stable region of the RNA polymerase has a low substitution rate (0.09%, 19/21,220) with respect to its large size. However, the 3′variable region, representing the 4 CDSs for the viral structural genes (the S, E, M, N proteins) and 5 PUPs, has a significantly higher substitution rate (0.18%, 10/5,447) (Table 3). 30 substitutions were detected by comparative analysis of the 5 published genome sequences of the SARS coronavirus, with an overall substitution rate of approximately 0.10%; a substantial portion (63.3%, 19/30) of them are transitions. The C-T type accounts for 68.4% (13/19), A-G 31.6% (6/19). Seven out of the 30 substitutions were confirmed in 2 independent sequences.

Approximately half (15/30) of the substitutions were predicted to be non-synonymous mutations in the CDSs for proteins, in addition to the 3 in the PUPs (Table 4). Among the 19 substitutions detected in the RNA polymerase, approximately half (10 substitutions) could lead to amino acid changes. The ratio was found even higher in the S proteins (3/4) and in the M proteins (2/2), which were postulated undergoing alterations of their physical/chemical features, irrespective of other possible changes of the higher structure and function of the membrane protein directly related to immunoreactions. No deletion or insertion event was detected.

The RNA-dependent RNA polymerase was found to be highly conservative among the 5 viral genomes, as well as in many other RNA viruses analyzed. The detected substitution rate was only 0.09%, the lowest among the 5 ORFs from which substitutions were identified (Table 3). The highest conserved region of the RNA polymerase resides in the region corresponding to the nucleotide position 14573—16894 in the SARS-associated virus, against

Table 3 Summarized substitutions in the 5 ′SARS-associated virus genomes (Isolate BJ01)

ORF	Size (nt)	Number of substitutions	Substitute rate (%)
RNA polymerase	21220	19	0.09
S protein	3767	4	0.11
M protein	665	2	0.30
PUP1	824	4	0.46
PUP3	191	1	0.52
Non-ORF		1	
Total	29725	30	0.10

Table 4 Substitutions of the nucleotide and amino acid sequences in the 5 ′SARS-associated virus genomes

ORF	nt position in BJ01	a.a. position in the ORF	BJ01[c]		TOR2[c]		US[c]		CUHK[c]		HKU[c]		Relative frequency[a]
RNA polymerase	2582	779	T	(Val)	C	(Val)	T	(Val)	T	(Val)	C	(Val)	C(2)/T(3)
	7727	2494	G	(Pro)	G	(Pro)	G	(Pro)	T	(Pro)	G	(Pro)	T(1)/G(4)
	7900	2552	C	(Ala)	C	(Ala)	T	(Val)	C	(Ala)	C	(Ala)	T(1)/C(4)
	7911	2556	G	(Asp)	G	(Asp)	G	(Asp)	G	(Asp)	A	(Asn)	A(1)/G(4)
	8368	2708	G	(Ser)	G	(Ser)	G	(Ser)	G	(Ser)	C	(Thr)	C(1)/G(4)
	8398	2718	G	(Arg)	G	(Arg)	G	(Arg)	G	(Arg)	C	(Thr)	C(1)/G(4)
	8553	2770	T	(Leu)	G	(Val)	G	(Val)	G	(Val)	G	(Val)	T(1)/G(4)
	9385	3047	C	(Ala)	T	(Val)	T	(Val)	C	(Ala)	T	(Val)	C(2)/T(3)
	9460	3072	T	(Val)	T	(Val)	T	(Val)	C	(Ala)	T	(Val)	C(1)/T(4)
	9835	3197	T	(Val)	C	(Ala)	C	(Ala)	C	(Ala)	C	(Ala)	T(1)/C(4)
	10568	3441	C	(Thr)	A	(Thr)	A	(Thr)	A	(Thr)	A	(Thr)	C(1)/A(4)
	16603	5453	C	(Ala)	C	(Ala)	T	(Ala)	C	(Ala)	C	(Ala)	T(1)/C(4)
	17545	5767	G	(Glu)	T	(Asp)	T	(Asp)	G	(Glu)	T	(Asp)	G(2)/T(3)
	17827	5861	C	(Arg)	C	(Arg)	C	(Arg)	T	(Arg)	C	(Arg)	T(1)/C(4)
	18046	5934	G	(Lys)	G	(Lys)	G	(Lys)	G	(Lys)	A	(Lys)	A(1)/G(4)
	19045	6267	A	(Glu)	A	(Glu)	G	(Glu)	G	(Glu)	A	(Glu)	G(2)/A(3)
	19819	6525	G	(Val)	A	(Val)	A	(Val)	A	(Val)	A	(Val)	G(1)/A(4)
	20577	6778	A	(Gln)	G	(Arg)	A	(Gln)	A	(Gln)	A	(Gln)	G(1)/A(4)
	20891	6883	G	(Asp)	T	(Tyr)	G	(Asp)	G	(Asp)	G	(Asp)	T(1)/G(4)
S protein	21702	77	A	(Asp)	G	(Gly)	G	(Gly)	A	(Asp)	G	(Gly)	A(2)/G(3)
	22203	244	C	(Thr)	T	(Ile)	T	(Ile)	C	(Thr)	T	(Ile)	C(2)/T(3)
	23201	577	T	(Ser)	G	(Ala)	T	(Ser)	T	(Ser)	T	(Ser)	G(1)/T(4)
	24853	1127	T	(Leu)	T	(Leu)	C	(Leu)	T	(Leu)	T	(Leu)	C(1)/T(4)
PUP1	25550	101	T	(Met)	T	(Met)	T	(Met)	T	(Met)	A	(Lys)	A(1)/T(4)
	25654	136	C	(Gln)	A	(Lys)	A	(Lys)	A	(Lys)	A	(Lys)	C(1)/A(4)
	26031	261	C	(Pro)	A	(Pro)	A	(Pro)	A	(Pro)	A	(Pro)	C(1)/A(4)
M protein	26581	68	C	(Ala)	C	(Ala)	C	(Ala)	C	(Ala)	T	(Val)	T(1)/C(4)
	26838	154	T	(Ser)	T	(Ser)	C	(Pro)	T	(Ser)	T	(Ser)	C(1)/T(4)
PUP3	27224	57	T	(Leu)	C	(Pro)	C	(Pro)	C	(Pro)	C	(Pro)	T(1)/C(4)
Non-ORF	27808		C		T		T		C		T		C(2)/T(3)
Number of substitutions related to BJ01[b]			(30)		16	(11)	17	(10)	11	(5)	19	(13)	

a) Number of substitutions detected by comparing the genome sequences at the specified position. b) Number of substitutions in comparison with the nucleotide (nt) and amino acid (a.a.) sequences of Isolate BJ01. c) BJ01/AY278488.2; US/AY278741.1; CUHK/AY278554.1; HKU/AY278491.2;

TOR2/NC_004718.1.

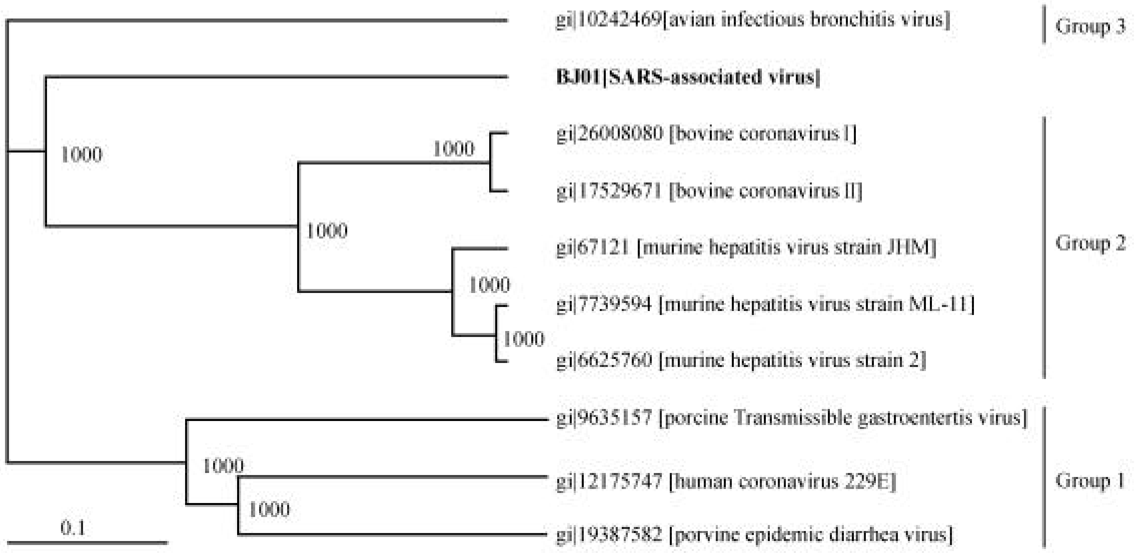
gi|10242469[avian infectious bronchitis virus]
Group 3
BJ01[SARS-associated virus]
gi|26008080 [bovine coronavirus I]
gi|17529671 [bovine coronavirus II]
gi|67121 [murine hepatitis virus strain JHM]
Group 2
gi|7739594 [murine hepatitis virus strain ML-11]
gi|6625760 [murine hepatitis virus strain 2]
gi|9635157 [porcine Transmissible gastroentertis virus]
gi|12175747 [human coronavirus 229E]
Group 1
gi|19387582 [porvine epidemic diarrhea virus]
1000
1000
1000
1000
1000
1000
1000
0.1

Fig. 5. A proposed phylogenetic tree of the SARS-associated viruses based on the complete amino acid sequences of the RNA polymerase. The bar represents the percentage of calculated divergency in evolution, and the bootstrap values deducted from 1000 replicates.

its counterpart in bovine coronavirus (AF391542, AF220295) and murine hepatitis virus (AF208067, AF029248, AF208066, AF207902, AF201929) with an identity of about 70% (640—643/914 a.a.). An exceptionally conserved GC-rich region was identified from the nucleotide position 29571—29602 (32 nt, cgaggccacgcggagtacgatcgagggtacag) at the 3′end of the genome, 3′to the CDS for the N protein among all 17 known coronaviruses.

A phylogenetic tree was proposed on the basis of all previously published nucleotide/amino acid sequences of the RNA-dependent RNA polymerase of all 64 known strains of 17 different coronaviruses in GenBank (Fig. 5). The phylogenetic comparisons covered 4 isolates from birds (avian infectious bronchitis virus, avian infectious laryngotracheitis virus, turkey coronavirus and puffinosis virus), 2 from rodents (murine hepatitis virus and rat coronavirus), 7 from house animals or pets (equine coronavirus, canine coronavirus, feline coronavirus, porcine epidemic diarrhea virus, porcine hemagglutinating encephalomyelitis virus, porcine transmissible gastroenteritis virus and bovine coronavirus), and 4 from human beings (human coronavirus 229E, human coronavirus OC43, human enteric coronavirus 4408, SARS-associated coronavirus). In addition to the 5′stable region of the RNA polymerase, no significant homology in the CDS or PUPs was detected between the SARS-associated virus and all the other genomes sequenced so far.

3 Discussion

(ⅰ) Classification of the SARS-associated virus. A number of hypotheses were postulated regarding to the pathogen(s) of SARS before the SARS-associated coronavirus was recognized. Neither clinical symptoms of phlegm nor increasing leukocytes, nor a curative effect of empirical anti-microbial treatment for SARS, have been the chief forms of evidence cited for excluding bacteria, as well as chlamydia and mycoplasma, as candidate pathogens[3]. The rapid spreading of the disease through respiratory tracts might suggest its possible viral etiology. An RNA virus was considered a likely candidate on the basis of the wide spectrum of atypical symptoms. The atypical clinical features might be related to the replication errors in the RNA viruses, which are significantly higher than that in DNA viruses. Based on this hypothesis, we focused on RNA viruses while still paying attention to other possible pathogen(s).

To date, the complete or partial sequences of 862 RNA viruses available in GenBank include 80 retroid viruses, 219 dsRNA viruses, 100 ssRNA negative-stranded viruses and 463 ssRNA positive-stranded viruses (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html). Our genome-wide alignment with all the known sequences indicates that it is relatively reasonable to place it in the family of Coronaviridae (See Supplement Data, http://www.genomics.org.cn/SARS). This classification is consistent with the morphology under electronic microscope that all members of the Coronaviridae family are characterized by the crown-like appearance (hence the name “crown”)[8]. It should be emphasized that the similarity is mainly attributable to the large RNA-dependent RNA polymerase, which should not be regarded to be directly related to antigenicity, virulence or adaptation of the SARS-associated virus.

The classification and phylogenetic study are essential to the identification of the SARS pathogen(s) and help in elucidating the viral origin and evolution. It would also help in predicting future re-emergence of the deadly infectious diseases. The phylogenetic tree also suggests that SARS virus might have originated from non-human animals, of concern to both professionals and laypersons. If so, the likely candidate could be of bovine- or murine-origin, based on the evidence from the sequence similarity of the RNA polymerases. Other house mammals and pets that are in close contact with human, should not be neglected.

Another critical issue is the possibility that this virus is man-made, released either intentionally by bioterrorists or accidentally from a biolab by carelessness. A man-made origin is very unlikely based on our sequence-based analysis and knowledge of genomics. First, it is absolutely not a simple recombinant product from any known natural genes by means of splicing, gene shuffling or any known measures, which can be concluded from the high sequence variation of all the genes except the RNA polymerase. The only available scientific information that could have been employed to produce this virus artificially is the conserved gene order (not the sequence per se), which is a direct mimic of the coronavirus. Secondly, if it were man-made, it would be either an artificial product of a totally random synthesis without reference to any recent knowledge, or based on an as-of-yet totally unknown technology.

In terms of its evolution, the high rate of base substitution as detected among various isolates together with the typical genome organization, may have made major contributions to its rapid adaptation and deadly virulence. Possible mechanisms, such as RNA recombination or reassortation, which are as-of-yet lacking solid evidence, might also have enhanced significantly the rate of evolution by which the virus obtained growth advantages and/or virulence. Nothing concerning its origin can be concluded with certainty at this moment until further evidence becomes available.

(ii) Phylogenetic analysis of SARS-associated viruses. Based on comparative analysis on the complete genome sequences of the 5′SARS-associated coronaviruses isolated from patients in Canada, USA, and China (Hong Kong, Beijing), a phylogenetic tree of the SARS-associated coronaviruses was proposed (Fig. 6).

The phylogenetic tree suggests that Isolate BJ01 could be closest to one of the isolates (gi|30027610/CUHK) identified in Hong Kong. However, another isolate (gi|30023963/HKU) identified in Hong Kong might be closer to that from Toronto (gi|29826277/TOR2). The Toronto patient, from whom the isolate was obtained, traveled from Hong Kong[1]. The isolate identified in USA appears to be the farthest from all others. One conclusion is that the mutation rate is surprisingly high if we take such a short time period into consideration. A high mutation rate is also consistent with the high error rate of RNA replication[8]. More molecular epidemiological data, based on sequences of multiple isolates identified from different regions together with detailed clinical data, are essential to elucidating the route of global spreading and mutation during transmission.

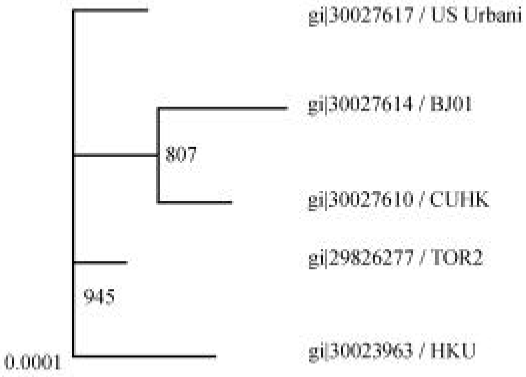

Fig. 6. A proposed phylogenetic tree of the 5′SARS-associated viruses based on the complete genomes. The bar represents the percentage of calculated divergency in evolution, and the bootstrap values deducted from 1000 replicates.

(iii) Sequence variations and medical significance. Even though a large fraction (61.3%) of substitutions in the viral genomes are in the CDS for the RNA polymerase (Table 3), the substitution rate is actually the lowest among the CDSs for the viral proteins with respect to its large size. A major contribution by studying the RNA polymerase is to understand the molecular mechanism of its low fidelity similar to other viral RNA polymerases, even though no direct correlation could be predicted to the viral antigenicity that is the most important to host immune response. More importantly, the RNA polymerase is well known to support the viral transcription independent of its host cells[9]. It could become the target of choice for future drug design or screening to inhibit the replication of the SARS-associated virus by exploiting its unique presence in the RNA viruses, coupled with the fact that the inhibition of RNA-dependent RNA replication should not affect any normal functions in human cells. Treatment of AIDS with “cocktail” is an illustrative example[10].

The substitution rate (0.11%) of the S protein among the known coronaviruses is close to that of the whole genome (0.10%). It is striking that the 3 out of 4 substitutions we have identified are non-synonymous and were found where the change of the hydrophobicity/hydrophilicity of the amino acids could be predicted (Table 4). It has been postulated that coronaviruses might first attach their S protein to a cellular receptor(s), getting the entry into the host cells[11]. Thus it is hypothesized that the primary and the higher order structure of the S protein might be related to the SARS virus' ability to escape from the immune surveillance. Furthermore, the S protein might be the major antigen contributing to both the humoral immune response and that of the cytotoxic T lymphocytes[11]. Therefore, the S protein is potentially a key candidate for developing vaccines and virus-neutralizing antibodies, in addition to its potential use in diagnostics.

It is surprising that the 2 substitutions identified in the M protein are both C-T transitions, and more importantly, non-synonymous. With respect to its relatively small size, the substitution rate (0.30%, 2/665 nt) is the highest among the viral structural proteins, and is also significantly higher than the overall rate of the whole virus genome (Table 3). Further studies are urgently needed to sequence more virus genomes in order to verify the substitutions and to elucidate the possible structural and functional changes and their medical significance. The predicted structure is consistent with its location in that it crosses the lipid bilayer of the envelope and is involved in envelope formation and membrane transport, as well as in immunoreactivities. The M protein appears to be another candidate for diagnostics and therapeutics.

No substitution has yet been found in the two inner proteins, the E protein and the N protein. One reasonable interpretation is that they both are important for the virus itself to survive. The E protein is the smallest among the viral structural proteins and is purported to play a role in envelope assembly as a phosphoprotein located inside the virion. It might be a potential target for drug design if advantage is taken of its conservation in evolution, but more studies are required. The predicted physical and chemical features of the putative proteins indicate that the N protein is a highly basic protein associated with the RNA genome, forming the long helical nucleocapsid. It might also be used as a target for drug design and therapy.

In summary, genomic and genetic knowledge can provide an extraordinary lines of information about the pathogenesis and virulence of the SARS-associated virus. More sequence data of various isolates would significantly expand our vision on the etiology and evolution of the virus and are essential for developing new approaches to diagnostics, prevention, and therapy of SARS. With respect to future diagnostic testing, prevention and treatment based on antibodies and vaccines, as well as exploring the immunoreactions, the S protein, perhaps together with the M protein, appear to be the most important candidates. For developing drugs, promising targets include the RNA polymerase to block the RNA replication, the N protein to interfere with the structure of the nucleocapsid, and the E protein to arrest envelope formation and virus assembly. The M protein and E protein could be the targets for inhibiting signal transduction pathways that are important for the virus to damage the immune system. The expression, or actual existence, and possible function of the PUPs should be explored by a combined means of genomics and proteomics. More comprehensive research and solid data are urgently needed to understand such a witty and powerful organism with all its enigmas and miracles, hidden in such a tiny genome, wrapped with a human coat, successfully surpassing the human defense system, and fatally attacking humans with substances almost entirely hijacked from humans.

Acknowledgements We are indebted to collaborators and clinicians from Chinese Center for Disease Control and Prevention, Peking Union Medical College Hospital, and the Municipal Governments of Beijing and Hangzhou. This work was supported by the Ministry of Science and Technology of China, Chinese Academy of Sciences, and the National Natural Science Foundation of China.

References

1. Poutanen, S. M., Low, D. E., Henry, B. et al., Identification of severe acute respiratory syndrome in Canada, N Engl. J. Med., www.nejm.org, March 31, 2003, 10.1056/NEMoa 030634.
2. Lee, N., Hui, D., Wu, A. et al., A major outbreak of severe acute respiratory syndrome in Hong Kong, N Engl. J. Med., www.nejm.org, April 7, 2003, 10.1056/NEJMoa 030685.
3. Peiris, J., Lai, S., Poon, L. et al., Coronavirus as a possible cause of severe acute respiratory syndrome, Lancet, 2003, 361: 1319—1325.
4. Ksiazek, T. G., Erdman, D., Goldsmith, C. S. et al., A novel coronavirus associated with severe acute respiratory syndrome, N Engl. J. Med., www.nejm.org, April 10, 2003, 10.1056/NEJMoa 020781.
5. Drosten, C., Gunther, S., Preiser, W. et al., Identification of a novel coronavirus in patients with severe acute respiratory syndrome, N Engl. J. Med., www.nejm.org, April 10, 2003, 10.1056/NEJMoa 030747.
6. Zhu, Q.-Y., Qin, E.-D., Wang, C. E., Isolation and identification of a novel coronavirus from patients with SARS, J. Chin. Biotech., 2003, 30 (in press).
7. Alam, S. L., Atkins, J. F., Gesteland, R. F., Programmed ribosomal frameshifting: much ado about knotting! Proc. Natl. Acad. Sci. USA, 1999, 96: 14177—14179.
8. Regenmortel, V., (edi) Virus Taxonomy (7th ed), Academic Press, 2001.
9. Uchil, P. D., Satchidanandam, V., Characterization of RNA synthesis, replication mechanism, and *in vitro* RNA-dependent RNA polymerase activity of Japanese encephalitis virus, Virology, 2003, 307: 358—371.
10. Henkel, J., Attacking AIDS with a " cocktail" therapy? FDA Consum, 1999, 33: 12—17.
11. Popova, R., Zhang, X., The spike but not the hemagglutinin/esterase protein of bovine coronavirus is necessary and sufficient for viral infection, Virology, 2002, 294: 222—236.

(Received April 29, 2003; Published online May 1, 2003)