Genome Organization of the SARS-CoV

Jing Xu[1]*, Jianfei Hu[2,1]*, Jing Wang[2,1]*, Yujun Han[1]*, Yongwu Hu[1,3], Jie Wen[1], Yan Li[1], Jia Ji[1], Jia Ye[1,4], Zizhang Zhang[5], Wei Wei[4], Songgang Li[1,2], Jun Wang[1], Jian Wang[1,4], Jun Yu[1,4#], and Huanming Yang[1,4#]

[1] Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; [2] College of Life Sciences, Peking University, Beijing 100871, China; [3] Wenzhou Medical College, Wenzhou 325003, China; [4] James D. Watson Institute of Genome Sciences, Zhijiang Campus, Zhejiang University and Hangzhou Genomics Institute, Hangzhou 310008, China; [5] College of Materials Science and Chemical Engineering, Yuquan Campus, Zhejiang University, Hangzhou 310027, China.

Annotation of the genome sequence of the SARS-CoV (severe acute respiratory syndrome-associated coronavirus) is indispensable to understand its evolution and pathogenesis. We have performed a full annotation of the SARS-CoV genome sequences by using annotation programs publicly available or developed by ourselves. Totally, 21 open reading frames (ORFs) of genes or putative uncharacterized proteins (PUPs) were predicted. Seven PUPs had not been reported previously, and two of them were predicted to contain transmembrane regions. Eight ORFs partially overlapped with or embedded into those of known genes, revealing that the SARS-CoV genome is a small and compact one with overlapped coding regions. The most striking discovery is that an ORF locates on the minus strand. We have also annotated non-coding regions and identified the transcription regulating sequences (TRS) in the intergenic regions. The analysis of TRS supports the minus strand extending transcription mechanism of coronavirus. The SNP analysis of different isolates reveals that mutations of the sequences do not affect the prediction results of ORFs.

Key words: SARS-CoV, genome annotation, transcription, ORF, PUP, TRS

Introduction

Severe acute respiratory syndrome-associated coronavirus (SARS-CoV), the pathogen of SARS, is a positive single-stranded RNA virus. It is classified as a member of Family *Coronaviridae* taxonomically because its physical profile and genome organization are similar to other known coronaviruses (*1*-*3*).

Five proteins in the SARS-CoV genome, R (replicase), S (spike), E (envelope), M (membrane) and N (nucleocapsid), homologically aligned themselves with those of other well-understood coronaviruses (*4*). The others were previously called PUPs (putative uncharacterized proteins) for their unknown structural or functional features and dissimilarity to those known sequences. However, it has been found that some of the PUPs matched the entries in the NCBI database (*5*).

Coronavirus performs a specific process of transcription known as discontinuous RNA synthesis (*6*, *7*), which is correlated with the primary and secondary structures of its TRS (transcription regulating sequence). Two prevailing but contradictive models, leader-primed transcription and minus-strand extending transcription, have been proposed to interpret this mechanism (*1*, *8*, *9*, *10*). The main discrepancies between them are the temporal process of the transcription and the existence of the subgenomic mRNAs.

In this paper, we report the annotation of the SARS-CoV genome, with the complete sequence of Isolate BJ01 as reference (*11*), and the exploration of its transcription mechanism.

*** These authors contributed equally to this work.**
# Corresponding authors.
E-mail: junyu@genomics.org.cn; yanghm@genomics.org.cn

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Results

Initial annotation of the SARS-CoV genome

The results were generated by a combination of predictions from various gene identification methods. By using FGENSV, 14 ORFs (open reading frames) were predicted and named F1∼F14 (Table 1). Two (F2 and F14) of them were novel to those previously reported, and F14 locates in the minus strand. With parameters trained from the known genes (R, S, E, M, and N), Glimmer (Version 2) predicted nine ORFs that were named G1∼G9. BGFV identified another nine genes that were named B1∼B9. Besides the computational prediction, we manually identified five more ORFs (BGI-PUP-S-1∼S-5) as candidates. Each of these candidates has an upstream region matching the pattern of TRS and its translated sequence is longer than 40 amino acids. All ORFs mentioned above were uniformly listed according to their initial sites along the genome sequence, and the predicted physiochemical properties of these ORFs were presented as well (Figure 1; Table 1).

The major physiochemical properties of different ORFs are various. For example, the GC contents of these ORFs range from 31.2% to 53.5%, while the range of the negative charge varies from 0 to 15.9%.

Table 1 Predicted ORFs and Their Physiochemical Characteristics in the SARS-CoV Genome (Isolate BJ01)

ORF	Position	Length (nt)	GC content (%)	Average MW (kDa)	pI	Hydrophobicity (%)	Hydrophilicity (%)	Charge (+)(%)	Charge (−)(%)
R	246-13,379 13,379-21,466	21,222	40.8	790.28	6.3	30.8	44.3	11.8	10.5
BGI-PUP-R-1	715-1,206	492	46.7	17.74	11.5	33.1	49.1	16.0	1.2
S	21,473-25,240	3,768	38.7	139.17	5.5	30.4	44.8	9.1	9.2
BGI-PUP-S-1	21,936-22,082	147	32.6	5.64	9.7	47.9	37.5	12.5	2.1
BGI-PUP-S-2	22,461-22,595	135	36.2	4.99	9.6	47.7	38.6	11.4	2.3
BGI-PUP-S-3	23,238-23,384	147	40.1	5.75	9.3	50.0	35.4	12.5	2.1
BGI-PUP-S-4	24,798-24,998	201	38.8	7.43	11.0	39.4	53.0	16.7	0.0
BGI-PUP-S-5	25,188-25,310	123	34.9	4.91	9.2	30.0	50.0	15.0	2.5
PUP1	25,249-26,073	825	40.3	30.90	5.6	34.7	39.1	8.4	8.0
PUP2	25,670-26,134	465	40.6	17.72	11.0	37.0	51.9	19.5	0.6
E	26,098-26,328	231	40.3	8.36	6.0	47.4	32.9	5.3	5.3
M	26,379-27,044	666	45.2	25.06	9.3	40.7	36.2	10.9	5.9
PUP3	27,055-27,246	192	31.2	7.54	4.7	47.6	42.9	11.1	15.9
PUP4	27,254-27,622	369	40.1	13.94	8.3	33.6	42.6	13.1	8.2
BGI-PUP4-1	27,619-27,753	135	31.8	5.30	3.9	61.4	27.3	2.3	13.6
PUP-Int-1	27,760-27,879	120	39.1	4.38	9.1	35.9	43.6	17.9	5.1
PUP-Int-2	27,845-28,099	255	40.0	9.56	9.4	31.0	41.7	15.5	3.6
N	28,101-29,369	1,269	48.4	46.03	10.1	17.3	54.0	15.4	8.5
PUP5	28,111-28,407	297	51.8	10.80	4.9	32.7	46.9	9.2	11.2
PUP-N-1	28,564-28,776	213	53.5	7.85	6.3	34.3	35.7	12.9	10.0
BGI-PUP-Neg-1	29,523-29,678	156	44.2	5.90	11.8	52.9	29.4	13.7	0.0

MW: molecular weight; nt: nucleotide; pI: isoelectric point.

Homological and structural analysis of ORFs

The componential and functional features of all the genes or ORFs, including the known nonstructural and structural proteins (R, S, E, M, and N), were explored (*5*). We here focused on the PUPs identified in the viral genome. Three of the PUPs were predicted to have transmembrane domains.

PUP1 is equivalent to ORF3 in Isolate Tor2 (*5*). It got 11 hits in GenBank through BLAST, two of which were putative transmembrane proteins. One was from *Ralstonia solanacearum*, *cytochrome* b-561

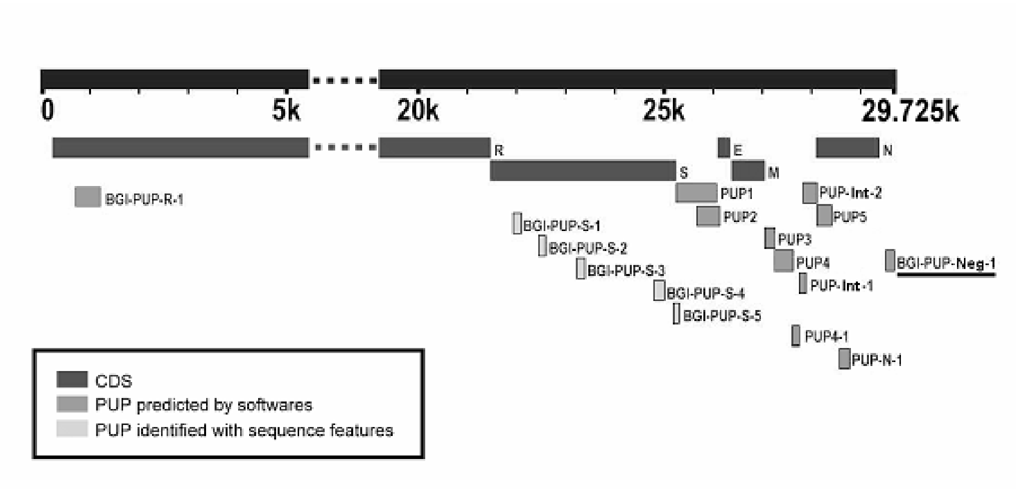


Fig. 1 The genome organization of the SARS-CoV (Isolate BJ01).

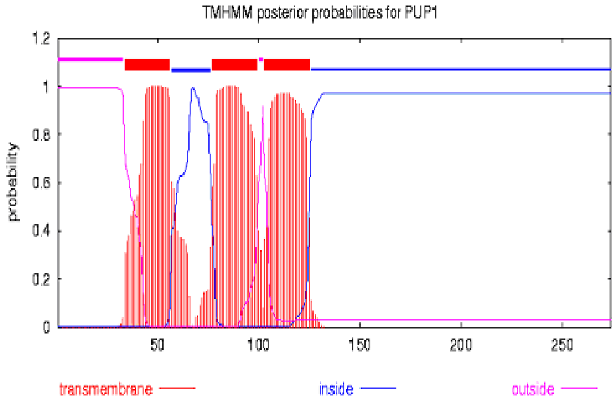


Fig. 2 Predicted transmembrane structure of PUP1 (TMHMM). Red blocks on the top line are predicted transmembrane domains. The abscissa represents the position on sequence, and the ordinate represents the probability of prediction.

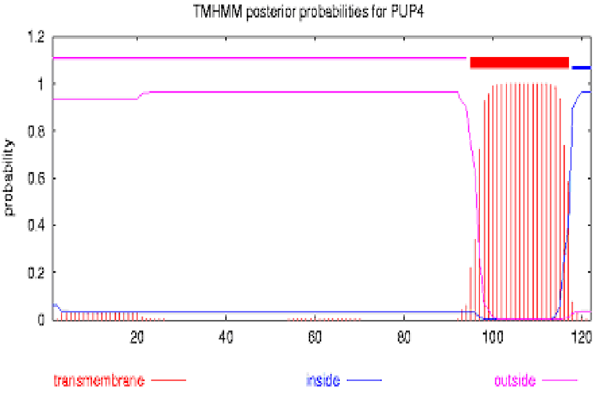


Fig. 3 Predicted transmembrane structure of PUP4 (TMHMM).

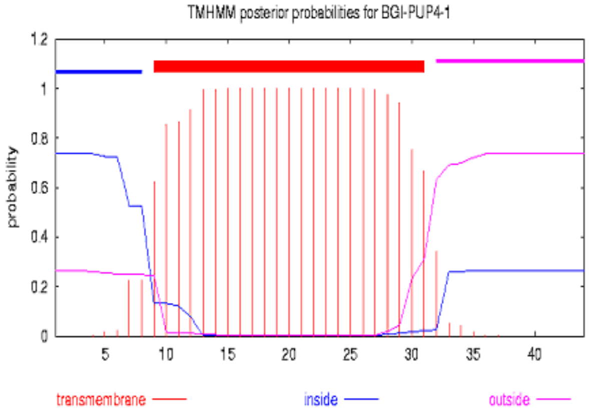


Fig. 4 Predicted transmembrane structure of BGI-PUP4-1 (TMHMM).

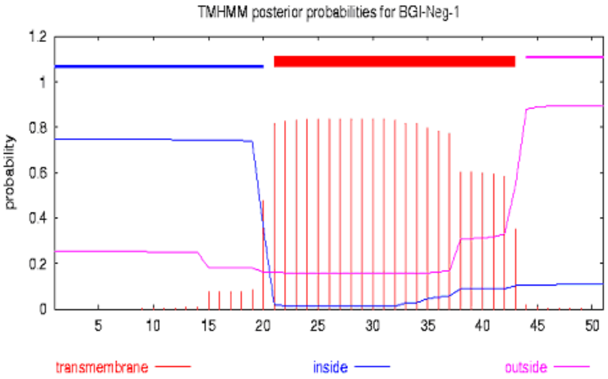


Fig. 5 Predicted transmembrane structure of BGI-Neg-1 (TMHMM).

(195 amino acids), with 97 amino acids of PUP1 aligned. The other was from *Sinorhizobium meliloti*, with 94 amino acids aligned. The identities were 28% and 25%, respectively. TMHMM predicted three transmembrane domains (Figure 2) in PUP1.

PUP4 is an equivalent to ORF8 in Isolate Tor2 (*5*). It aligned a hypothetical protein of *Cytophaga hutchinsonii* with 31% identity over a segment of 51 amino acids. TMHMM predicted a transmembrane region at its C-terminus (Figure 3).

BGI-PUP4-1 overlaps four nt with PUP4 at its N-terminus. In BLAST retrieving, it aligned 41 amino acids to a hypothetical protein, 50 amino acids in size, of *Clostridium perfringens* with an identity of 36%, and 31 amino acids to putative sterol-C5-desaturase of *Arabidopsis thaliana* with an identity of 38%. TMHMM identified one transmembrane domain in BGI-PUP4-1 (Figure 4), and it covers half of the total length (23 bp out of 45 bp). This ORF has a counterpart (ORF9) in Isolate Tor2.

BGI-PUP-Neg-1 is the only ORF detected on the minus strand of the viral genome. It consists of 51 amino acids with a similar TRS on its upstream, and is predicted to have a transmembrane region at 21-43 amino acids. The prediction from TMHMM showed that this ORF had a transmembrane domain (Figure 5).

BGI-PUP-R-1 is entirely embedded in the R protein, and is predicted to encode a protein of 163 amino acids. The BLASTp retrieving result showed its limited similarities to two segments, of 125 amino acids in *Streptococcus cristatus* and of 137 amino acids in *Caenorhabditis elegans*, respectively. Both of the two alignments has identities near 24%.

BGI-PUP-S-1∼S-5 are embedded in (BGI-PUP-S-1∼S-4) or overlapped (BGI-PUP-S-5) with the S protein. They were identified, in addition to criteria for their length (>40 amino acids), by the relatively conserved upstream TRSs. Only two of them, BGI-PUP-S-4 and BGI-PUP-S-5, got hits via BLASTp, retrieving against GenBank. The former matched a 41-amino-acid segment of a putative ethylene receptor in *Pyrus communis*, with an identity of 39%, and a putative nuclear protein family member of *C. elegans*, with 33% identify over an alignment of 69 amino acids; while the latter hit a 69-amino-acid-long segment in the putative nuclear protein of *C. elegans* with an identity of 33%.

PUP2 has a counterpart in Isolate Tor2, the ORF4. It matched 4 segments of different entries in GenBank: 138 amino acids with NADH dehydrogenase subunit2 of *Laudakia stoliczkana*, 137 amino acids with a hypothetical protein of *Methanosarcina barkeri*, 85 amino acids with myosin IXb of *Homo sapiens*, and 85 amino acids with MY9B_HUMAN myosin IXb. All of these alignments have the same identities of 28%.

PUP3 got no hit in GenBank, and no transmembrane or other characteristic domain was predicted with software available. It has a typical ORF with 63 amino acids, and has all other features of a gene, like TRS, start and stop codons. It is equivalent to ORF7 in Isolate Tor2 (*5*).

PUP-Int-1 is thus named since it is a PUP located in the intergenic region between PUP4 and PUP5. It got no hit in GenBank, and no characteristic structure was predicted.

PUP-Int-2 is a protein of 84 amino acids in length, following PUP-Int-1 in the same intergenic region. It matched a putative protein of *C. elegans* (25 amino acids, 48% identity), and a hypothetical protein, MGC28705, of *Mus musculus* (40 amino acids, 42% identity). The two ORFs mentioned above are equivalent to ORF10 and ORF11 in Isolate Tor2.

PUP5 is equivalent to ORF13 in Isolate Tor2. It aligned a segment (69 amino acids) with XP_225244, a hypothetical protein of *Rattus norvegicus*. The identity of their alignment is 26%, similar to the retinoblastoma-associated protein RAP140 of *Homo sapiens*, which has 24% identity over an alignment of 82 amino acids.

PUP-N-1 is entirely embedded in the N protein, which aligned a 64 amino acids segment with DEC-205 of *Mus musculus* (28% identity), and the lymphocyte antigen 75 of *Homo sapiens* as well (25% identity).

Characterization of substitutions

All sequences together with 338 nucleotides (overlapped ORFs may count one nucleotide twice or more times) variations among 42 isolates have been reported (from Jianfei Hu, personal communication). However, after a thorough survey, we have found that the variations do not affect our prediction of ORFs.

Ka and *Ks* are the rates of non-synonymous and synonymous substitutions, and the ratio between them (*Ka/Ks*) indicates the selection pressure of a gene. If *Ka/Ks* is higher than one, the selection pressure the gene takes is heavy; on the contrary, a ratio less than one means a lower pressure. The substitutions and the ratios for those 21 identified ORFs are displayed in Table 2.

Table 2 Substitution Status of ORFs in the Genome of SARS-CoV

ORF	Size (nt)	Substitutions	Non-synonymous Substitutions	Substitute rate (%)	*Ka*	*Ks*	*Ka*/*Ks*
R	21,222	223	171	1.05	0.91	0.98	0.93
BGI-PUP-R-1	492	3	1	0.61	0.57	0.52	1.09
S	3,768	47	38	1.25	1.14	0.94	1.21
BGI-PUP-S-1	147	0	0	0	0.00	0.00	0
BGI-PUP-S-2	135	2	1	1.48	3.72	0.79	4.70
BGI-PUP-S-3	147	0	0	0	0.00	0.00	0
BGI-PUP-S-4	201	3	0	1.49	0.00	1.89	0
BGI-PUP-S-5	123	5	1	4.07	3.18	3.69	0.86
PUP1	825	25	21	3.03	2.88	1.93	1.49
PUP2	465	14	10	3.01	2.45	3.33	0.73
E	231	2	2	0.87	1.02	0.00	0
M	666	8	4	1.2	0.69	2.23	0.31
PUP3	192	8	7	4.17	3.99	2.34	1.71
PUP4	369	3	3	0.81	0.94	0.00	0
PUP4-1	135	0	0	0	0.00	0.00	0
PUP-Int-1	120	5	0	4.17	0.00	4.62	0
PUP-Int-2	255	2	0	0.78	0.00	0.91	0
N	1,269	9	4	0.71	0.36	1.49	0.24
PUP-N-1	213	2	0	0.94	0.00	1.29	0
PUP5	297	2	2	0.67	0.77	0.00	0
BGI-PUP-Neg-1	156	0	0	0	0.00	0.00	0

Regulatory elements in the non-coding regions

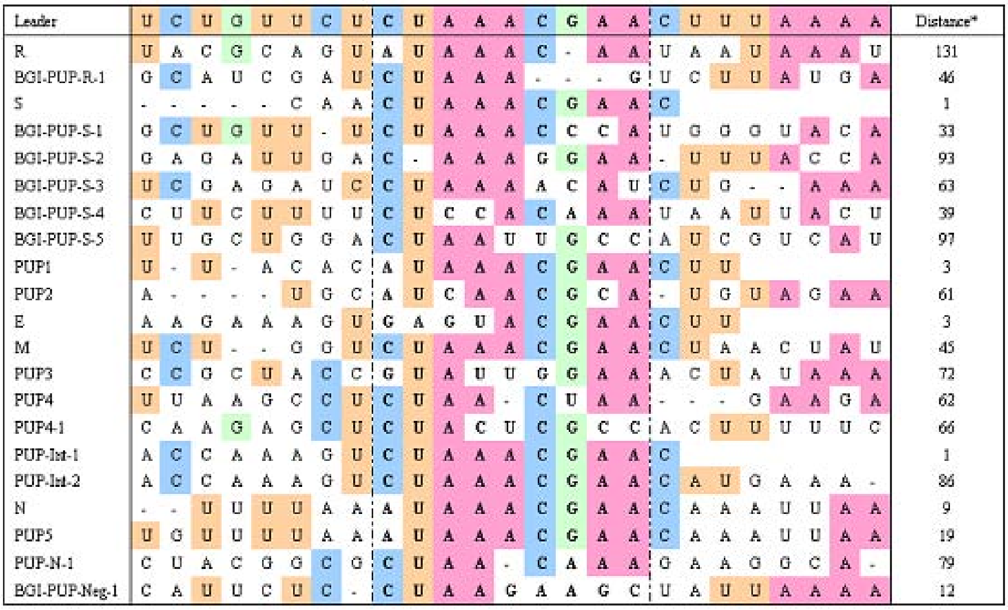
The 5′ UTR of the whole genome contains a special segment with a size variation between 65 and 90 nt for different species of coronaviruses, being notified as leader, which is immediately followed by a segment called leader-mRNA junction (*12*). Both of them are crucial components to the discontinuous transcription model of the coronavirus. For further study, we aligned up the upstream sequence of each ORF. The results showed that these intergenic segments were relatively conserved, and composed a core consensus, CUAAACGAA, which was identical to the junction segment mentioned above. It provided a convincing evidence to support the discontinuous transcription model of the coronavirus. The conserved segments, or TRSs, and their multiple alignments were illustrated in Figure 6, from which we can tell the conserved core consensus apparently. The values of most distances between TRSs and their initial sites of corresponding genes are less than 100 nt (to R, this value is 131 nt). In some cases, two overlapped ORFs refer to the same TRS. Analysis on these segments can help understand the transcriptional mechanism of the coronavirus.

Another remarkable phenomenon is that its 5′ upstream contains a segment that is similar to the 5′ end region of the plus strand (Figure 7), which was first detected in the AIBV (*3*).

The 3′ UTR of the genome is also required for its transcription, in that the truncation of this part can totally inhibit the transcription of subgenomic mRNAs, despite all of the synthesized minus-strand RNAs (*13*).

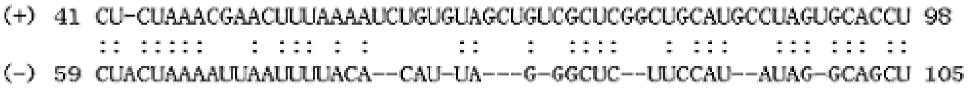
The s2m in the 3′ UTR region is a motif found in Order *Nidovirales*, such as bovine, porcine, and ovine coronaviruses. It is also thought to be a common feature of *Coronaviridae* (*14*). The identification of the motif in those genomes provides supplemental evidence for genetic taxonomy, although the motif may be a gift from RNA recombination rather than a relic of their ancestor. The genome of the SARS-CoV (Isolate BJ01) has homologous sequence to the s2m at the position from 29,567 to 29,607 nt.

Poly(A) is the 3′ end region of the genome, and



Leader	U	C	U	G	U	U	C	U	C	U	A	A	A	C	G	A	A	C	U	U	U	A	A	A	A	Distance*
R	U	A	C	G	C	A	G	U	A	U	A	A	A	C	-	A	A	U	A	A	U	A	A	A	U	131
BGI-PUP-R-1	G	C	A	U	C	G	A	U	C	U	A	A	A	-	-	-	G	U	C	U	U	A	U	G	A	46
S	-	-	-	-	-	C	A	A	C	U	A	A	A	C	G	A	A	C								1
BGI-PUP-S-1	G	C	U	G	U	U	-	U	C	U	A	A	A	C	C	C	A	U	G	G	G	U	A	C	A	33
BGI-PUP-S-2	G	A	G	A	U	U	G	A	C	-	A	A	A	G	G	A	A	-	U	U	U	A	C	C	A	93
BGI-PUP-S-3	U	C	G	A	G	A	U	C	C	U	A	A	A	A	C	A	U	C	U	G	-	-	A	A	A	63
BGI-PUP-S-4	C	U	U	C	U	U	U	U	C	U	C	C	A	C	A	A	A	U	A	A	U	U	A	C	U	39
BGI-PUP-S-5	U	U	G	C	U	G	G	A	C	U	A	A	U	U	G	C	C	A	U	C	G	U	C	A	U	97
PUP1	U	-	U	-	A	C	A	C	A	U	A	A	A	C	G	A	A	C	U	U						3
PUP2	A	-	-	-	-	U	G	C	A	U	C	A	A	C	G	C	A	-	U	G	U	A	G	A	A	61
E	A	A	G	A	A	A	G	U	G	A	G	U	A	C	G	A	A	C	U	U						3
M	U	C	U	-	-	G	G	U	C	U	A	A	A	C	G	A	A	C	U	A	A	C	U	A	U	45
PUP3	C	C	G	C	U	A	C	C	G	U	A	U	U	G	G	A	A	A	C	U	A	U	A	A	A	72
PUP4	U	U	A	A	G	C	C	U	C	U	A	A	-	C	U	A	A	-	-	-	G	A	A	G	A	62
PUP4-1	C	A	A	G	A	G	C	U	C	U	A	C	U	C	G	C	C	A	C	U	U	U	U	U	C	66
PUP-Int-1	A	C	C	A	A	A	G	U	C	U	A	A	A	C	G	A	A	C								1
PUP-Int-2	A	C	C	A	A	A	G	U	C	U	A	A	A	C	G	A	A	C	A	U	G	A	A	A	-	86
N	-	-	U	U	U	U	A	A	A	U	A	A	A	C	G	A	A	C	A	A	A	U	U	A	A	9
PUP5	U	G	U	U	U	U	A	A	A	U	A	A	A	C	G	A	A	C	A	A	A	U	U	A	A	19
PUP-N-1	C	U	A	C	G	G	C	G	C	U	A	A	-	C	A	A	A	G	A	A	G	G	C	A	-	79
BGI-PUP-Neg-1	C	A	U	U	C	U	C	-	C	U	A	A	G	A	A	G	C	U	A	U	U	A	A	A	A	12

Fig. 6 The TRS sequences in the SARS-CoV genome (Isolate BJ01). *This refers to the number of nucleotides between the first nucleotide of the TRSs and the first letter of the start codon of the corresponding ORFs.



```
(+) 41 CU-CUAAACGAACUUUAAAAUCUGUGUAGCUGUCGCUCGGCUGCAUGCCUAGUGCACCU 98
       :: :::::   : ::: : :      ::   :  ::::   : :::   ::: ::: ::
(-) 59 CUACUAAAAUUAAUUUUACA--CAU-UA---G-GGCUC--UUCCAU--AUAG-GCAGCU 105
```

Fig. 7 Homological comparison of the 5′ end and 3′ end of the SARS-CoV genome (Isolate BJ01).

each subgenomic mRNA acquired it as a fused tail during its transcription. Poly(A)-binding proteins (PABPs) from the host cell interact with this terminal region, in order to initiate the transcription and enhance the stability of subgenomic mRNA. Results from experiments indicated that functional and selective pressure forced the shortened Poly(A)s to be repaired or restored their missing part, and the longer the poly(A) is (compared with the wild-type isolate), the higher efficiency the transcription has (*15*).

Discussion

Comparison of the gene prediction software

Genetic information of any life is preserved in its genome, and annotation is the first step to decode the sequence. The length of the SARS-CoV genome is more than 30 Kb, while only 5 structural or non-structural genes seem not to accord with the general characteristics for virus genome and the compactness and concentration of genetic information. In addition to these genes, it may have some non-structural proteins but lack experimental support. The absence probably results from their short existing-time before decomposition. In this study, encouraged by the supposition, we employed four different instruments to predict genes in the SARS-CoV (Isolate BJ01).

Glimmer (Version 2) predicted two ORFs that started with UUG (G5) and GUG (G8), respectively, instead of the usual initiation codon. This fact challenges the prevalent viewpoint that all ORFs start with AUG. The hypothetical minus sense ORF identified by FGENESV (from 48 to 203 nt on the minus strand or 29,523 to 29,678 nt on the plus strand) may be fake, but we should not absolutely deny the probability of the existence of minus ORFs.

Results of four prediction approaches with the genome sequence of the SARS-CoV (Isolate BJ01) were compared. The combined result contradistinguishes with annotations of the isolates from four different areas, one sample per city (Table 3; ref. *5*, *11*, *16*, *17*).

The GC contents of the five well-explored proteins (R, S, E, M and N) range from 38.7% to 48.4% (Table

1). Most of the other ORFs have approximate values, while the GC contents of BGI-PUP-S-1, PUP3, and PUP4-1 are 32.6%, 31.2% and 31.8%, respectively. Further more, BGI-PUP-S-1 and PUP4-1 both have a *Ka/Ks* ratio of zero. These facts may suggest that the probabilities of the two ORFs to be proteins are lower than others.

Table 3 Comparison of Prediction and Annotation of SARS-CoV (Isolate BJ01)

Prediction				Combined result	Annotation		
FGENESV	Glimmer	ZCURVE_CoV	BGFV	BJ01	Tor2	Urbani	SIN2500
				R	ORF1	R	R
F1	G1	orf1a	B1	ORF1a	ORF1a	orf1a	orf1
F2				BGI-PUP-R-1			
F3	G2	orf1b	B2	ORF1b	ORF1b		
F4	G3	S	B3	S	S	S	S
				BGI-PUP-S-1			
				BGI-PUP-S-2			
				BGI-PUP-S-3			
				BGI-PUP-S-4			
				BGI-PUP-S-5			
F5	G4	Sars274	B4	PUP1	ORF3	X1	PUP1
F6				PUP2	ORF4	X2	PUP2
F7		E		E	E	E	E
F8	G5#	M	B5	M	M	M	M
F9		Sars63	B6	PUP3	ORF7	X3	PUP3
F10	G6	Sars122	B7	PUP4	ORF8	X4	PUP4
	G7	Sars44		PUP4-1	ORF9		
F11		Sars39		PUP-Int-1	ORF10		
F12	G8#	Sars84	B8	PUP-Int-2	ORF11	X5	
F13	G9	N	B9	N	N	N	N
				PUP5	ORF13		PUP5
				PUP-N-1	ORF14		
F14*				BGI-PUP-Neg-1			

*The ORF on the minus-strand, predicted by FGENESV.

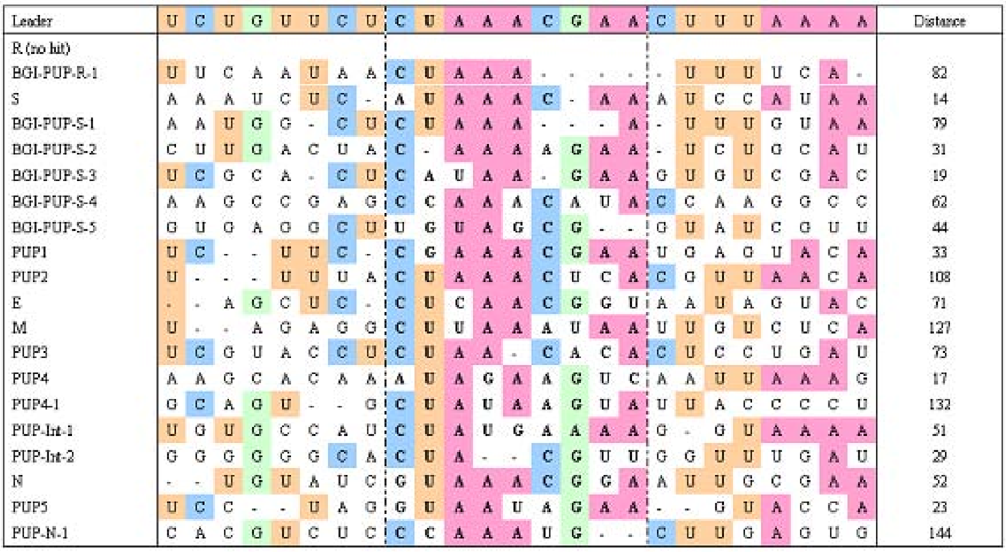
#Glimmer (Version 2) predicted ORFs, not starting with AUG.

Our study was based on two presumptions. Firstly, the identification of the five proteins could not explain the pathogenesis and relevant observations. Secondly, it might be caused by our "one protein, multiple functional domains" deduction, or by overlapped independent genes which have not been explored. Most PUPs (14 out of 16) were embedded in or overlapped with at least one other ORF, indicating the compactness of the viral genome.

Human coronavirus HCoV-OC43 has a hemagglutinin-esterase (HE) protein while another well-explored virus HCoV-229E has not. Being classified as a coronavirus which could infect human, the SARS-CoV seems not to contain such an ORF coding for the HE protein in that there is no space between R protein and S protein, the very position for HE to exist independently. It was found that the relics of HE protein had spread to neighboring regions (from Jianfei Hu, personal communication). This phenomenon suggests that large-scale recombination might have taken place.

Furthermore, we employed FGENESV to explore the sequences of MHV (NC_001846 in NCBI) and AIBV (NC_001451 in NCBI), and compared the results with their previous annotations, respectively. ORFs for their structural proteins were totally predicted, while some hypothetical ORFs were not.

Apart from the leader-mRNA junction segments found in the plus strand, we also detected some other segments similar to the consensus on the minus-strand sequence. They were ahead of the anti-codons of the termini of positive sense ORFs (Figure 8). This discovery should not be simply coincident, but the unique reason to explain the phenomena is unknown and worthy of further exploration.



Leader	U	C	U	G	U	U	C	U	C	U	A	A	A	C	G	A	A	C	U	U	U	A	A	A	A	Distance
R (no hit)																										
BGI-PUP-R-1	U	U	C	A	A	U	A	A	C	U	A	A	A	-	-	-	-	-	U	U	U	U	C	A	-	82
S	A	A	A	U	C	U	C	-	A	U	A	A	A	C	-	A	A	A	U	C	C	A	U	A	A	14
BGI-PUP-S-1	A	A	U	G	G	-	C	U	C	U	A	A	A	-	-	-	A	-	U	U	U	G	U	A	A	79
BGI-PUP-S-2	C	U	U	G	A	C	U	A	C	-	A	A	A	A	G	A	A	-	U	C	U	G	C	A	U	31
BGI-PUP-S-3	U	C	G	C	A	-	C	U	C	A	U	A	A	-	G	A	A	G	U	G	U	C	G	A	C	19
BGI-PUP-S-4	A	A	G	C	C	G	A	G	C	C	A	A	A	C	A	U	A	C	C	A	A	G	G	C	C	62
BGI-PUP-S-5	G	U	G	A	G	G	C	U	U	G	U	A	G	C	G	-	-	G	U	A	U	C	G	U	U	44
PUP1	U	C	-	-	U	U	C	-	C	G	A	A	A	C	G	A	A	U	G	A	G	U	A	C	A	33
PUP2	U	-	-	-	U	U	U	A	C	U	A	A	A	C	U	C	A	C	G	U	U	A	A	C	A	108
E	-	-	A	G	C	U	C	-	C	U	C	A	A	C	G	G	U	A	A	U	A	G	U	A	C	71
M	U	-	-	A	G	A	G	G	C	U	U	A	A	A	U	A	A	U	U	G	U	C	U	C	A	127
PUP3	U	C	G	U	A	C	C	U	C	U	A	A	-	C	A	C	A	C	U	C	C	U	G	A	U	73
PUP4	A	A	G	C	A	C	A	A	A	U	A	G	A	A	G	U	C	A	A	U	U	A	A	A	G	17
PUP4-1	G	C	A	G	U	-	-	G	C	U	A	U	A	A	G	U	A	U	U	A	C	C	C	C	U	132
PUP-Int-1	U	G	U	G	C	C	A	U	C	U	A	U	G	A	A	A	A	G	-	G	U	A	A	A	A	51
PUP-Int-2	G	G	G	G	G	G	C	A	C	U	A	-	-	C	G	U	U	G	G	U	U	U	G	A	U	29
N	-	-	U	G	U	A	U	C	G	U	A	A	A	C	G	G	A	A	U	U	G	C	G	A	A	52
PUP5	U	C	C	-	-	U	A	G	G	U	A	A	U	A	G	A	A	-	-	G	U	A	C	C	A	23
PUP-N-1	C	A	C	G	U	C	U	C	C	C	A	A	A	U	G	-	-	C	U	U	G	A	G	U	G	144

Fig. 8 Comparison of TRS in SARS-CoV Isolate BJ01 (minus sense). Core segments (CUAAACGAA) are mark up in bold style. Distance in part B is the number of nucleotide between the last letters of the TRSs to the terminal codon of their corresponding ORFs.

Transcriptomics – two models of transcription

The conserved TRS is one of the characteristics of discontinuous transcription that *coronaviridae* performs while duplicating. Currently, two different models are applied to interpret the transcription mechanism of coronavirus: leader-primed discontinuous transcription model and minus-strand extending transcription model with subgenomic mRNAs.

Due to the previous failure in detecting the minus-strand subgenomic mRNA, leader-primed transcription is generally accepted. A full-length minus-strand RNA was considered to act as the template for transcription of all subgenomic mRNAs. In this model, duplications of the leader sequence leave the 3′ end of the template, and then move to intergenic regions upstream each mRNA on the minus-strand template. After duplicated leaders fusing again with the reversed TRSs on the Body sequence (to distinguish with the leader portion) through base pairing, the discontinuous transcription procedure was then triggered.

Sawicki *et al.* proposed another model, the minus-strand extending transcription model, in which subgenome-length negative sense segments were detected in infected cells (*18*, *19*). In this postulation, subgenome-length minus strands derives directly from the genome RNA during transcription, gets terminal TRS counterparts from the body sequence, and then fuses on the TRS region to accomplish the minus-strand after getting the counterpart of the leader. The completed minus-strand RNAs serve as templates for subgenomic mRNAs (*9*, *10*).

Generally, it is found that coronavirus mRNAs are synthesized in amounts reversely related to their sizes, and the N protein is richer than any other proteins in infected cells. It suggests that the gradient of subgenomic mRNA amount results in that the large mRNA tends to premature and generates less proteins than the small subgenomic mRNAs (*20*). Site-directed mutations to TRSs along the genome decrease the transcriptional efficiencies or even eliminate the synthesis of the subgenomic RNAs. Mutations introduced in different places demonstrate that the upstream TRSs do not affect the downstream ones, but the latter affect the replication of the former. It provides a possible interpretation why the downstream proteins are richer in infected cells than the upstream ones (*21*). Although mutations in TRSs reduce their opportunities of transcription, some subgenomic mRNAs are synthesized with the mutated sites performed as markers, and the origin of the fused region (here refers TRS) can be traced. These results show that all the TRSs come from the body sequences rather than the leaders, and provide evidence for the minus-strand extending transcription model (*22*).

Even though the proteins of SARS-CoV were predicted by software, and related transcriptional mechanisms were described, further experiments are still required to prove these hypotheses.

Materials and Methods

The annotation and subsequent analysis were mainly performed on the complete genome sequence of Isolate BJ01 (Accession No. AF278488 in GenBank). FGENESV, a program for gene prediction provided by Softberry Inc. (Mount Kisco, USA) through a web-based interface, has been specially modified and trained with parameters for virus (http://www.softberry.com/berry.phtml?topic=gfin dv). Glimmer (Version 2), from TIGR (The Institute for Genomic Research), is a program for gene identification with high performance in handling small genomes like bacteria and archaea (*23*, *24*). ZCURVE-CoV, developed by researchers in Tianjin University, is an approach to recognize ORFs with Z-Curve theory (*25*). BGFV is a program developed by Beijing Genomics Institute, based on the self-organizing theory (http://arxiv.org/abs/physics/0102048). Fundamental principle of BGFV is the compositional discrepancy between coding and non-coding regions, which is relatively distinctive for simpler species. The prediction have been compared with previous annotations of other isolates for cross-checking. The length threshold for ORFs is the same as that applied by Marra *et al.* to Isolate Tor2 (*5*). One ORF is postulated to be a protein-coding region, if its translated sequence is longer than 40 amino acids. A unique segment, the leader-mRNA junction (*26*), should exist upstream to the transcription initiation site, within a distance of 100 nt, except for the R protein (to R, the distance is 131 nt).

For nomenclature, most of the previously reported ORFs were designated by their original names (*11*), while some PUPs got suffixes. "PUP-Int-1" refers to a PUP locating in intergenic region. Especially, those that are first reported in this paper were named with a prefix of "BGI". If an ORF embedded in or overlapped with a known one, the name of the known hosting ORF will be inserted in its name and a sequential number will be attached. BGI-PUP-R-1, for example, stands for the first identified ORF overlapped with the R protein.

Physiochemical features, such as MW and pI, were calculated with a program from Dr. Yan Li, Beijing Genomics Institute (personal communication), and the transmembranous domains of proteins were identified by TMHMM (*27*), while DAS (*28*) provided similar results (figures from DAS are not shown).

Acknowledgements

The authors thank Ministry of Science and Technology of China, Chinese Academy of Science, and National Science Foundation of China for financial support. We are grateful to Jinsong Liu for providing the BGFV program and mathematical calculations.

References

1. Lai, M.M., *et al.* 1984. Studies on the mechanism of RNA synthesis of a murine coronavirus. *Adv. Exp. Med. Biol.* 173: 187-200.
2. Baric, R.S., *et al.* 1987. Studies into the mechanism of MHV transcription. *Adv. Exp. Med. Biol.* 218: 137-149.
3. Boursnell, M.E., *et al.* 1987. Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. *J. Gen. Virol.* 68: 57-77.
4. de Vries, A.A., *et al.* 1997. The genome organization of the Nidovirales: similaritys and differences between Arteri-, Toro-, and coronaviruses. *Seminars in Virology* 8: 33-47.
5. Marra, M.A., *et al.* 2003. The genome sequence of the SARS-associated coronavirus. *Science* 300: 1399-1404.
6. Lin, Y.J., *et al.* 1994. Identification of the cis-acting signal for minus-strand RNA synthesis of a murine coronavirus: implications for the role of minus-strand RNA in RNA replication and transcription. *J. Virol.* 68: 8131-8140.
7. Jeong, Y.S. and Makino, S. 1994. Evidence for coronavirus discontinuous transcription. *J. Virol.* 68: 2615-2623.
8. Baric, R.S., *et al.* 1985. Characterization of leader-related small RNAs in coronavirus-infected cells: further evidence for leader-primed mechanism of transcription. *Virus Res.* 3: 19-33.
9. Sawicki, S.G. and Sawicki, D.L. 1995. Coronaviruses use discontinuous extension for synthesis of subgenome-length negative strands. *Adv. Exp. Med. Biol.* 380: 499-506.

10. Sawicki, S.G. and Sawicki, D.L. 1998. A new model for coronavirus transcription. *Adv. Exp. Med. Biol.* 440: 215-219.
11. Qin, E.D., *et al.* 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chin. Sci. Bull.* 48: 941-948.
12. Hofmann, M.A., *et al.* 1993. Leader-mRNA junction sequences are unique for each subgenomic mRNA species in the bovine coronavirus and remain so throughout persistent infection. *Virology* 196: 163-171.
13. Lin, Y.J., *et al.* 1996. The 3′ untranslated region of coronavirus RNA is required for subgenomic mRNA transcription from a defective interfering RNA. *J. Virol.* 70: 7236-7240.
14. Jonassen, C.M., *et al.* 1998. A common RNA motif in the 3′ end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. *J. Gen. Virol.* 79: 715-718.
15. Spagnolo, J.F. and Hogue, B.G. 2000. Host protein interactions with the 3′ end of bovine coronavirus RNA and the requirement of the poly(A) tail for coronavirus defective genome replication. *J. Virol.* 74: 5053-5065.
16. Rota, P.A., *et al.* 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394-1399.
17. Ruan, Y.J., *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.
18. Sethna, P.B., *et al.* 1989. Coronavirus subgenomic minus-strand RNAs and the potential for mRNA replicons. *Proc. Natl. Acad. Sci. USA.* 86: 5626-5630.
19. Sawicki, S.G. and Sawicki, D.L. 1990. Coronavirus transcription: subgenomic mouse hepatitis virus replicative intermediates function in RNA synthesis. *J. Virol.* 64: 1050-1056.
20. van Marle, G., *et al.* 1995. Regulation of coronavirus mRNA transcription. *J. Virol.* 69: 7851-7856.
21. Pasternak, A.O., *et al.* 2001. Sequence requirements for RNA strand transfer during nidovirus discontinuous subgenomic RNA synthesis. *Embo. J.* 20: 7220-7228.
22. van Marle, G., *et al.* 1999. Arterivirus discontinuous mRNA transcription is guided by base pairing between sense and antisense transcription-regulating sequences. *Proc. Natl. Acad. Sci. USA.* 96: 12056-12061.
23. Delcher, A.L., *et al.* 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27: 4636-4641.
24. Salzberg, S.L., *et al.* 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26: 544-548.
25. Chen, L.L., *et al.* 2003. ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. *Biochem. Biophys. Res. Commun.* 307: 382-388.
26. Makino, S., *et al.* 1986. Leader sequences of murine coronavirus mRNAs can be freely reassorted: evidence for the role of free leader RNA in transcription. *Proc. Natl. Acad. Sci. USA.* 83: 4204-4208.
27. Krogh, A., *et al.* 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305: 567-580.
28. Cserzo, M., *et al.* 1997. Prediction of transmembrane alpha-helices in procariotic membrane proteins: the dense alignment surface method. *Protein Eng.* 10: 673-676.