A Genome Sequence of Novel SARS-CoV Isolates: the Genotype, GD-*Ins*29, Leads to a Hypothesis of Viral Transmission in South China

E'de Qin[1]*, Xionglei He[2]*, Wei Tian[2]*, Yong Liu[2]*, Wei Li[2]*, Jie Wen[2], Jingqiang Wang[2], Baochang Fan[1],Qingfa Wu[2], Guohui Chang[1], Wuchun Cao[1], Zuyuan Xu[2], Ruifu Yang[1],Jing Wang[2], Man Yu[1], Yan Li[2], Jing Xu[2], Bingyin Si[1], Yongwu Hu[2], Wenming Peng[1], Lin Tang[2], Tao Jiang[1], Jianping Shi[2], Jia Ji[2], Yu Zhang[2], Jia Ye[2], Cui'e Wang[1],Yujun Han[2], Jun Zhou[2], Yajun Deng[2], Xiaoyu Li[1], Jianfei Hu[2], Caiping Wang[2], Chunxia Yan[2], Qingrun Zhang[2], Jingyue Bao[2], Guoqing Li[2], Weijun Chen[2], Lin Fang[2], Changfeng Li[2], Meng Lei[2], Dawei Li[2], Wei Tong[2], Xiangjun Tian[2], Jin Wang[2], Bo Zhang[2], Haiqing Zhang[2], Yilin Zhang[2], Hui Zhao[2], Xiaowei Zhang[2], Shuangli Li[2], Xiaojie Cheng[2], Xiuqing Zhang[2], Bin Liu[2], Changqing Zeng[2], Songgang Li[2], Xuehai Tan[2], Siqi Liu[2], Wei Dong[2], Jun Wang[2], Gane Ka-Shu Wong[2], Jun Yu[2], Jian Wang[2], Qingyu Zhu[1]#, and Huanming Yang[2]#

[1] *Institute of Microbiology and Epidemiology, Chinese Academy of Military Medical Sciences, Beijing 100071, China;* [2] *Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300 & James D. Watson Institute of Genome Sciences, Zhijiang Campus, Zhejiang University, Hangzhou 310008, China.*

We report a complete genomic sequence of rare isolates (minor genotype) of the SARS-CoV from SARS patients in Guangdong, China, where the first few cases emerged. The most striking discovery from the isolate is an extra 29-nucleotide sequence located at the nucleotide positions between 27,863 and 27,864 (referred to the complete sequence of BJ01) within an overlapped region composed of BGI-PUP5 (BGI-postulated uncharacterized protein 5) and BGI-PUP6 upstream of the N (nucleocapsid) protein. The discovery of this minor genotype, GD-*Ins*29, suggests a significant genetic event and differentiates it from the previously reported genotype, the dominant form among all sequenced SARS-CoV isolates. A 17-nt segment of this extra sequence is identical to a segment of the same size in two human mRNA sequences that may interfere with viral genome replication and transcription in the cytosol of the infected cells. It provides a new avenue for the exploration of the virus-host interaction in viral evolution, host pathogenesis, and vaccine development.

Key words: Severe Acute Respiratory Syndrome (SARS), genotype, GD-*Ins*29

Introduction

Severe Acute Respiratory Syndrome (SARS) has infected thousands of people and cost hundreds of deaths globally since it first emerged in Guangdong, China, in November 2002 (ref.*1*; http://www.who.int/csr/sars/en/). Sixteen complete or partial genome sequences of SARS-CoV isolates have been available since March 14 this year (ref. *2-5*; http://www.ncbi.nlm.nih.gov/; Table 1). In the process of surveying all available genomic sequences of the SARS-CoV, we have obtained the complete genomic sequence of an isolate, named Isolate GZ01 originally. The sequence contains an extra 29-nucleotide segment in the vicinity of the viral structural proteins, M (the membrane protein) and N (the nucleocapsid protein). We have renamed the isolate as GD-*Ins*29 to indicate the origin, length, and the type of mutation. We believe that this novel genotype represents one of the early variants of the SARS-CoV. The absence of the 29-nt fragment may help the virus to escape the likely interference from locally homologous RNA molecules of host origin and to become more prevalent in human hosts. This sequence also holds unusually high num-

* **These authors contributed equally to this work.**

Corresponding authors.

E-mail: zhuqy@nic.bmi.ac.cn; yanghm@genomics.org.cn

ber of variations among all the sequenced SARS-CoV isolates, suggesting a more distant relationship from other sequences SARS-CoV genomes published thus far.

Table 1 The Complete Genome Sequences of 17 Isolates of SARS-CoV

Isolate	Genome size (nt)	Accession number	Modification date
SIN2500	29,711	AY283794.1	9-May-03
SIN2677	29,705	AY283795.1	9-May-03
SIN2679	29,711	AY283796.1	9-May-03
SIN2748	29,706	AY283797.1	9-May-03
SIN2774	29,711	AY283798.1	9-May-03
TOR2	29,751	NC_004718.3	22-May-03
Urbani	29,727	AY278741.1	21-Apr-03
CUHK-W1	29,736	AY278554.2	14-May-03
CUHK-Su10	29,736	AY282752.1	7-May-03
HKU-39849	29,742	AY278491.2	18-Apr-03
TW1	29,729	AY291451.1	14-May-03
ZJ01	29,715	AY297028.1	19-May-03
BJ01	29,725	AY278488.2	1-May-03
BJ02	29,740	AY278487	29-May-03
BJ03	29,738	AY278490	29-May-03
BJ04	29,732	AY279354	29-May-03
GD01	29,757	AY278489	29-May-03

Results and Discussion

A 29-nt sequence segment was identified from a viral isolate, SARS-CoV GD01

The genome landscape of this isolate is not greatly different from others reported by our group and other laboratories. Twelve ORFs (open reading frames), including 6 CDSs (coding sequences) and 6 PUPs (postulated uncharacterized proteins) were predicted in the viral genome (Table 2; Figure 1). We annotated two new PUPs, BGI-PUP5 (so named to avoid possible ambiguity in nomenclature) and BGI-PUP6, located between the previously reported PUP4 and the N protein. These PUPs overlap in a 35-nt region (position 27,843 and 27,879, in reference to nucleotide position in the complete sequence of BJ01). Both PUPs share the same consensus leader sequence (5′-agUCUAAAAGAAC-3′) that is immediately followed by the predicted start codon of BGI-PUP5, and approximately 95-nt away from that of BGI-PUP6. The expression, actual existence of protein products, and possible function of the PUPs, such as that of BGI-PUP6, remain to be elucidated experimentally.

An extra sequence segment of 29-nt (GD-*Ins*29) in length was unambiguously identified within the overlapped region of BGI-PUP5 and BGI-PUP6. This sequence was confirmed with 60 high-quality sequencing reads from 35 clones of the site-specific amplicon-library constructed with RT-PCR products from this genomic region. Among them, 25 clones were sequenced from both ends. No other sequence variations were found in the sequences from this particular amplicon-library, even though minor variants are occasionally seen in other amplicon-libraries from different sequences due to minor variations from viral populations and, to a much less extent, RT-PCR generated aberrant products. We have also identified another SARS-CoV isolate that harbors the same sequence segment from a SARS patient in Guangdong Province. These results strongly suggest that we have found a novel yet minor genotype of SARS-CoV, but not encountered a sequence anomaly.

In the GD-*Ins*29 sequence, the 29-nt segment is located in-frame after the second nucleotide of Codon 35 in BGI-PUP5 (position 27,863) and the first nucleotide of Codon 7 of BGI-PUP6. The resulted hypothetical protein is predicted to have 122 amino acids. Without affecting the coding frame, the 29-nt sequence theoretically could be spliced out in two different ways since there is uncertainty whether the uridine is at the begin-

ning or at the end of the extra sequence, 5′-UCCUACUGGUUACCAACCUGAAUGGAAUA-3′ or 5′-CCUACUGGUUACCAACCUGAAUGGAAUAU-3′. No obvious sequence signatures were found within and around the sequence except several short stretches of simple repetitive sequences of 6 to 8 nucleotides in unit length. The actual mechanism, as to how such sequence was deleted, is yet to be revealed.

Table 2 The Predicted ORFs in the GD01 SARS-CoV Genome

ORF	Position[1]	Size (a.a.)	TRS position[2]	TRS sequence
R	246 – 13,379 13,379 - 21,466	7,073	107	A G U A UAAAC - AA UAA UAAA U U U U A
S	21,473 - 25,240	1,255	21,463	CAA CUAAACGAAC
BGI-PUP1	25,249 - 26,073	274	25,237	CACA UAAACGAACUU
BGI-PUP2	25,670 - 26,134	154	25,600	U GCA U C AACG C A - U G UA G AAU UAU
E	26,098 - 26,328	76	26,086	AG U GAGU ACGAACUU
M	26,379 - 27,044	221	26,325	GG UCUAAACGAACU AACU A U U A U U
BGI-PUP3	27,055 - 27,246	63	26,974	A C CG UA UUG GAA AC U AU AAAU UAA
BGI-PUP4	27,254 - 27,622	122	27,184	C CUCUAA - C U AA —G AA G A AU U A
BGI-PUP5[3]	*27,760 - 27,879*	*39*	*27, 750*	AG UCUAAACGAAC
BGI-PUP6[3]	*27,845 - 28099*	*84*	*27, 750*	AG UCUAAACGAAC A U G AAA -C U U C
BGI-PUP	27,760 - 27,863	122	27,750	AG UCUAAACGAAC
(GD-*Ins*29)[4]	-(29bp insertion)- 27,864 - 28,099			
N	28,101 - 29,369	422	28,083	U AAA UAAACGAAC AAAU U AA A
BGI - PUP7	28,111 - 28,407	98	28,083	U AAA UAAACGAAC AAAUU AA AA UG

[1] The nucleotide position is in reference to the complete sequence of BJ01. An ORF includes both start codon and stop codon. An actual position in GD01 is calculated by adding 3 nucleotides.

[2] The position is in reference to the first nucleotide in the consensus leader core sequence (CUAAACGAAC) of the TRS.

[3] The PUPs are equivalent to ORF 10 and ORF 11 in Tor2 (NC_004718) (*3*).

[4] BGI-PUP (GD-*Ins*29) is present only in the minor genotype of GD-*Ins*29.

Since the deletion is very significant in size, we reasoned that it might have been affected by the host factors. Upon searching the sequence against public human sequence databases, we have found two seemingly interesting matches of 17-nt segments in the center of the sequence to three human chromosomal locations. One (5′-GGUUACCAACCUGAAUG-3′) was aligned to protein coding sequences mapped to two chromosomal locations: 15q23 and 9q22; the other (5′-UGGUUACCAACCUGAAU-3′) matches to a sequence segment on chromosome 11, which overlaps with a repetitive sequence of the mammalian interspersed repeat, or MIR, and is most likely non-protein coding. The protein-coding sequences are both members of acidic (leucine-rich) nuclear phosphoprotein 32 family (*6*, *7*). Exhaustive database searches did not yield any other significant matching sequence in the human genomic sequence databases.

Since the virus is propagated within the cytosol, only the host sequences, such as processed transcripts, mRNAs and other operational RNAs, would have possibilities directly interacting with the viral host-dependent cellular processes, such as replication of the viral genome via a negative sense RNA intermediates, transcription of genes encoding viral replicase and structural proteins, and, to a limited extent, translation, if the interfering sequences enabling strand-annealing with viral RNA products. The unique 17-nt sequence that we have discovered from the human genome as mRNA forms is thought to be capable of annealing specifically with the negative sense strand of the viral RNA while the virus is replicating its RNA genomes from the negative sense RNA genome intermediates and transcribing viral transcripts for the replicase and structural proteins, thus interfere with the viral life cycle by reducing the efficiency of both viral replication and transcription. The prerequisite for the interference to happen is that the 17-nt sequence

comes from a protein or RNA coding sequence that exists in infected human cells with reasonable abundance. It is very suggestive that the interaction of the intermediate viral genome products, namely the negative sense RNAs, and the host RNA species might have happened in the propagation processes within a host of SARS-CoV, through intermolecular RNA-RNA recombination (*8*).

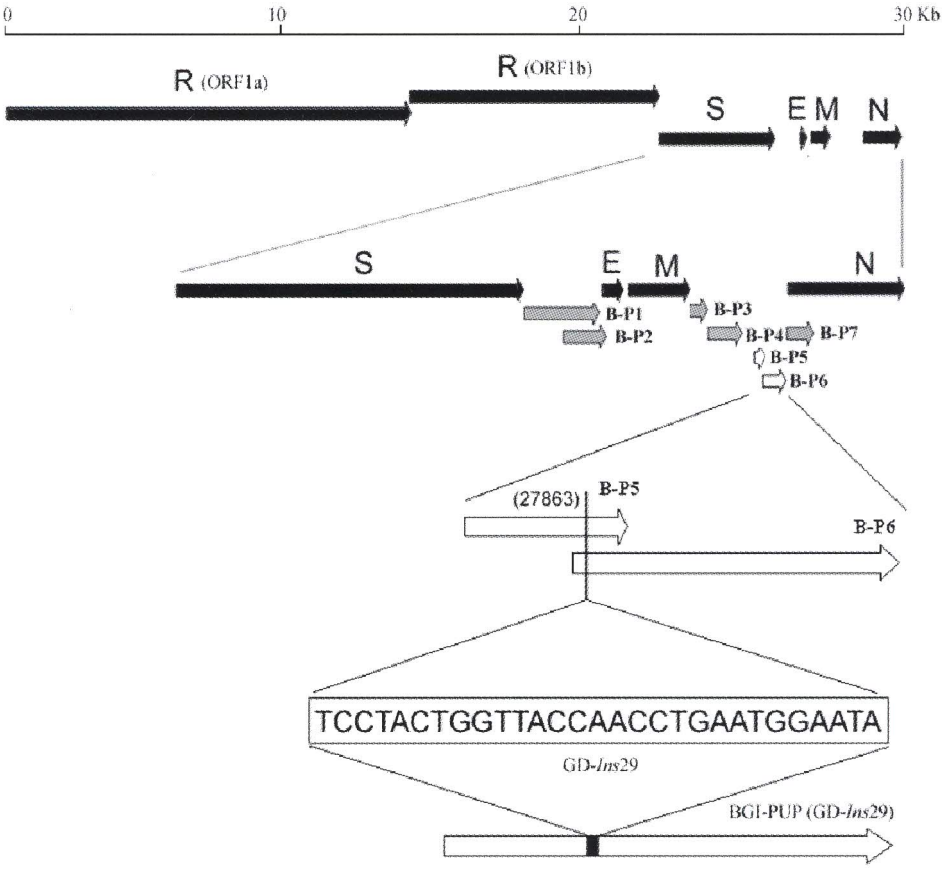

Fig. 1 The GD-*Ins*29 in the genome of Isolate GD01. Solid arrows indicate ORFs and the gray arrows denote the BGI-PUPs. Open arrows highlight the region of interest. The position of the insertion and related annotations are in the parentheses.

A high number of substitutions were also found in Isolate GD01 in comparison with other SARS-CoV genomes

After careful sequence alignment, 137 substitutions were identified in comparison with the other 16 published SARS-CoV genome sequences, yielded an overall-genome mutation rate of approximately 0.46% (Table 3). 45 substitutions were found just between GD01 and BJ01 (Supplementary Table 1). 38 out of 45 are unique to GD01. Special attention was paid to the sequences of the mutation sites and multiple clones from the amplicon-libraries constructed from multiple independent RT-PCR products were sequenced. A substantial portion (70.4%, 100/142) of the substitutions was predicted to be non-synonymous mutations in the ORFs, including 14 in the PUPs. Among the 92 substitutions detected in the R (replicase) protein, 70.7% (65/92) could lead to amino acid changes. Although the S (spike) protein has a low ratio (13/22), the ratio is higher in the other structural proteins, the M (4/4) and N proteins (3/4), as well as in BGI-PUP2 (4/5), BGI-PUP3 (2/2), and BGI-PUP (GD-*Ins*29, 2/2). Even though a large fraction of substitutions in the viral genome have been located in the ORF for the R protein, its substitution rate is actually the lowest among all the defined CDSs, with regard to its large size (21,222 nt). The mutation rate of SARS-CoV is quite high if we take into consideration such a short time period since it was

identified from human hosts. A high mutation rate is consistent with the high error rate of RNA replication (*9*) and the high fraction of non-synonymous substitutions also implies the possibility that the selective pressure from the host may have worked on the virus, albeit there is very little statistical power to support the conjecture with current data set. A very important notion from our data (including tens of thousands of sequencing traces) is the obvious absence of indels (insertions and deletions), suggesting that the viral polymerase is prone to the replication errors in proofreading but remains relatively accurate in frame-moving conveyed by the RdRp (RNA-dependent RNA polymerase) activity.

Table 3 Summarized Substitutions in 17 Isolates of SARS-CoV

ORF	Size (nt)	No. of S[1]	Percentage of substitute (%)	No. of N-Syn[1]	Percentage of N-Syn (%)
R	21,222	92 (26)[2]	0.43	65 (16)	71
S	3,768	22 (7)	0.58	13 (5)	59
BGI-PUP1	825	9 (3)	1.09	6 (2)	67
BGI-PUP2	465	5 (3)	1.08	4 (2)	80
E	231	1 (1)	0.43	1 (1)	100
M	666	4	0.60	4	100
BGI-PUP3	192	2	1.04	2	100
N	1,269	4	0.32	3	75
BGI-PUP (GD-*Ins*29)	369	2 (1)	0.54	2 (1)	100
Non-ORF		1			
Total	29,725	142 (41)[3]	0.46	100	

[1] S: substitution; N-Syn: non-synonymous substitution.

[2] Number in the parenthesis indicates the substitutions contributed solely by Isolate GD01.

[3] A single substitution at the same position in a region overlapped with two ORFs is counted as 2. The total number would be 137 if such a substitution event were calculated as 1, and the total number of substitutions contributed by Isolate GD01 would be 38, accordingly.

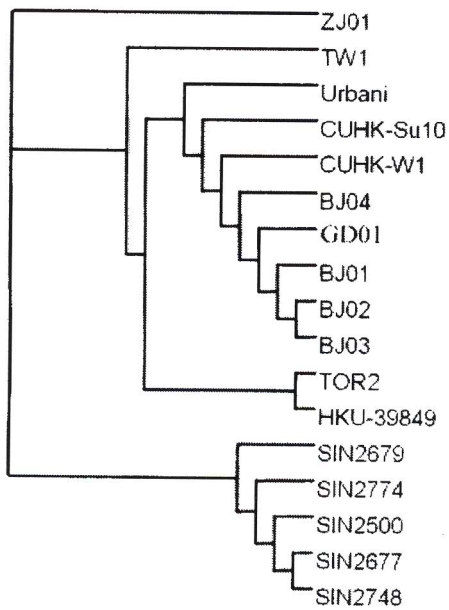

Fig. 2 Phylogenetic analysis of the 17 SARS-CoV isolates based on completed genomes. The proposed rectangular cladogram was generated by Clustalw 1.81 and the bootstrap values were deducted from 1,000 replicates. The sources and abbreviations of the sequences are referred to the text of Table 1.

Based on comparative analyses on the complete genome sequences of the 17 SARS-CoV isolates from patients identified in Canada, USA, Singapore, and China (Beijing, Zhejiang, Guangdong, Hong Kong, and Taiwan), an unrooted phylogenetic tree of the SARS-CoV was constructed (Figure 2). The phylogenetic tree positioned Isolate GD01 in the group composed of BJ01, 02, 03, and 04, two of the Hong Kong isolates (CUHK), and the US isolate (Urbani). This paradigm suggests a possible transmission path among Guangdong, Hong Kong, Beijing and USA. The link between the two isolates (TOR2 and HKU) in another group has been suggested that the Toronto patient, from whom the isolate was obtained, was known to have traveled from Hong Kong (*3*). In addition, the five isolates identified in Singapore seem to form another group. Two relatively distant isolates in Taiwan and Zhejiang, China, reported more recently, are more distant from the rest. Additional molecular epidemiological data, based on genomic sequences of viral isolates identified from different regions, is essential to elucidate the possible routes of global spreading and mutation during its transmission among humans.

The origin of SARS-CoV and possible multiple invasions to humans

The SARS-CoV has nearly identical genome organization, especially its gene order, with other members in *Coronaviridae* found in humans and other animals. The sequences of the SARS-CoV genome from various isolates are also almost identical with only a few dozen nucleotide substitutions per genome compared with each other. Therefore, we are strongly in favor of the hypothesis that the SARS-CoV may have originated from non-human animals, no matter it is virulent or latent to its host, in unknown reservoirs of the wild and recently moved onto humans through a less frequently established contacting route. It is of great importance that such non-human hosts should be promptly identified in order to prevent further evasion into human and other animal populations. The discovery of the GD-*Ins*29 genotype is very useful in differentiating the origin of SARS-CoV. The fact that we have only identified a couple of isolates in Guangdong Province supports the scenario that GD-*Ins*29 is the original form transmitted from non-human host to human ones and later turned into a new genotype after deletion of a DNA segment that may hinders its efficiency of propagation. Therefore, we predict two likely outcomes when surveying the animal reservoirs. First, the non-human reservoir of SARS-CoV may harbor genome sequences close to GD-*Ins*29, not the deleted form or the major genotype found most widespread among its human hosts. Second, both variants of the major and the minor genotypes could be found in the same animal reservoir owing to multiple transmissions of the virus back and forth between its human and non-human hosts.

Alternatively, there might have been at least two genotypes of SARS-CoV spreading in non-human host populations and they have infected human hosts as separate events. It is unlikely but possible that the virus have infected humans multiple times in a not-distant past and the two genotypes co-exist in the non-human hosts but not traceable in current human populations. Although we cannot rule out another possibility that the viruses of the major genotype had acquired a 29-nt sequence and become GD-*Ins*29 genotype during its propagation in human hosts, such event is deemed extremely rare since insertion or deletion is often lethal to the virus that have a rather compact genome.

Genomic and genetic knowledge can provide an extraordinary trove of information about the pathogenesis and virulence of SARS-CoV. More sequences data would significantly expand our knowledge on the etiology and evolution of the virus and are essential for seeking new approaches for diagnostics, vaccine development, and effective therapy of SARS. High quality sequence data and comprehensive analysis, as well as more experimental evidences, are still of urgent need to understand such a potent infectious agent.

Materials and Methods

We isolated the GD01 from a 54-year-old female patient. She was suspected being infected during her hospitalization by indirect contact with one of the "superspreaders", who stayed in the same hospital as she did. She was one of the SARS cases with known transmission path connecting to the "Index Cases" identified in Guangdong Province. The viral isolate was from the autopsied lung tissue of the patient and propagated in Vero-E6 cell culture. The isolates of the same genotype were also discovered from SARS patients in Guangdong Province. Details in their genomic sequences will be reported elsewhere.

Virions were prepared from Vero-E6 cell culture. Aliquots of the RT-PCR products from the viral RNA were sequenced directly and the remaining was cloned

into a plasmid vector (amplicon-library). Multiple clones from each amplicon-library were subsequently sequenced from both directions to confirm the results from those directly acquired from PCR products. High-quality sequences were assembled by using a sequence assembly package, Phred-Phrap-Consed (*10*). The consensus sequences from different amplicon-derived clones are accounted and minor variations among different clones are ignored for the consensus assembly. A total number of 75 amplicons and 2,718 sequencing reads were generated (1,538,993 bp of raw data), equivalent to approximately 52-fold coverage of the viral genome. A complete sequence of 29.757 Kb with an overall error rate of 0.0016% was obtained. The sequence of Isolate GD01 is available in GenBank (Accession No. AY278489).

Acknowledgements

We thank the Ministry of Science and Technology of China, the Chinese Academy of Sciences, National Natural Science Foundation of China, and the Chinese Academy of Military Medical Sciences for financial support. We thank Biopharmaceutical Center of Sun Yat-sen University for technical support. We are indebted to collaborators, clinicians and nurses from Peking Union Medical College Hospital, the National Center of Disease Control of China, Provincial Government of Zhejiang, and the Municipal Governments of Beijing and Hangzhou, as well as the patients and their families.

References

1. Peiris, J.S., *et al.* 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* 361: 1319-1325.
2. Rota, P.A., *et al.* 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*; http://www.sciencemag.org/cgi/rapidpdf/1085952v1.pdf.
3. Marra, M.A., *et al.* 2003. The genome sequence of SARS-associated coronavirus. *Science*; http://www.sciencemag.org/cgi/rapidpdf/1085953v1.pdf.
4. 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.
5. 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.
6. Zhu, L., *et al.* 1997. Cloning and characterization of a new silver-stainable protein SSP29, a member of the LRR family. *Biochem. Mol. Biol. Int.* 42: 927-935.
7. Mencinger, M., *et al.* 1998. Expression analysis and chromosomal mapping of a novel human gene, APRIL, encoding an acidic protein rich in leucines. *Biochim. Biophys. Acta* 1395: 176-180.
8. Domingo, E. and Holland, J.J. 1997. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* 51: 151-178.
9. Uchil, P. D. and Satchidanandam, V. 2003. Characterization of RNA synthesis, replication mechanism, and in vitro RNA-dependent RNA polymerase activity of Japanese encephalitis virus. *Virology* 307: 358-371.
10. Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8: 186-194.