Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

遗传学报 *Acta Genetica Sinica*, April 2006, 33 (4): 354–364 ISSN 0379-4172

Polymorphism of SARS-CoV Genomes

SHANG Lei[1,3*], QI Yan[2*], BAO Qi-Yu[2,①], TIAN Wei[1], XU Jian-Cheng[1], FENG Ming-Guang[3], YANG Huan-Ming[1]

1. *James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou* 310008, *China*;
2. *Institute of Biomedical Informatics, Wenzhou Medical College, Wenzhou* 325000, *China*;
3. *Institute of Microbiology, Zhejiang University, Hangzhou* 310029, *China*

Abstract: In this work, severe acute respiratory syndrome associated *coronavirus* (SARS-CoV) genome BJ202 (AY864806) was completely sequenced. The genome was directly accessed from the stool sample of a patient in Beijing. Comparative genomics methods were used to analyze the sequence variations of 116 SARS-CoV genomes (including BJ202) available in the NCBI GenBank. With the genome sequence of GZ02 as the reference, there were 41 polymorphic sites identified in BJ202 and a total of 278 polymorphic sites present in at least two of the 116 genomes. The distribution of the polymorphic sites was biased over the whole genome. Nearly half of the variations (50.4%, 140/278) clustered in the one third of the whole genome at the 3′ end (19.0 kb–29.7 kb). Regions encoding Orf10-11, Orf3/4, E, M and S protein had the highest mutation rates. A total of 15 PCR products (about 6.0 kb of the genome) including 11 fragments containing 12 known polymorphic sites and 4 fragments without identified polymorphic sites were cloned and sequenced. Results showed that 3 unique polymorphic sites of BJ202 (positions 13 804, 15 031 and 20 792) along with 3 other polymorphic sites (26 428, 26 477 and 27 243) all contained 2 kinds of nucleotides. It is interesting to find that position 18379 which has not been identified to be polymorphic in any of the other 115 published SARS-CoV genomes is actually a polymorphic site. The nucleotide composition of this site is A (8) to G (6). Among 116 SARS-CoV genomes, 18 types of deletions and 2 insertions were identified. Most of them were related to a 300 bp region (27 700–28 000) which encodes parts of the putative ORF9 and ORF10-11. A phylogenetic tree illustrating the divergence of whole BJ202 genome from 115 other completely sequenced SARS-CoVs was also constructed. BJ202 was phylogeneticly closer to BJ01 and LLJ-2004.

Key words: severe acute respiratory syndrome associated *coronavirus* (SARS-CoV); genome; polymorphism

Severe acute respiratory syndrome (SARS) is a new infectious disease that first emerged in Guangdong province, China, in November, 2002 and then quickly spread worldwide before being successfully controlled in 2003 by classical public health measures[1]. It had the characteristic of high mortality and morbidity. Within a short period of six months since its first outbreak, it affected 8 096 people and led to 774 deaths[2].

Since the publication of the first complete genomic sequence of SARS-CoV[3], 115 SARS-CoV genomic sequences have been completed and hundreds of additional partial sequences are available in the NCBI GenBank, all of which provide a strong foundation for a better understanding of the transmission and molecular evolution of SARS-CoV. However, it is still a great challenge to establish the relationship between the observed genomic variations and the biology of SARS-CoV.

SARS-CoV, an enveloped, positive-stranded RNA virus, was determined to be a new member of the *Coronaviridae* family and the supposed transmission was from the wild animal to human being[4]. It has the largest genome size among RNA viruses and a broad host range[5,6].

Received: 2005-04-16; Accepted: 2005-09-01

This work was supported by the Science Foundation of Wenzhou City(No. Y2003A005) and Zhejiang University (No. 181130-544301).

* Contributed equally to this study

① Corresponding author. E-mail: baoqy@genomics.org.cn; Tel: +86-577-8889 2799

The entire 29 700-base genome of SARS-CoV contains 12 putative open reading frames including four major structural proteins: the spike protein (S protein, 1 255 aa) which mediates attachment to cellular receptors and entry by fusion with cell membranes; the small envelope protein (E protein, 76 aa) which acts as a scaffold protein to trigger assembly; the membrane protein (M protein, 221 aa) which is an integral membrane protein involved in budding and interaction with the nucleocapsid and S proteins and the nucleocapsid protein (N protein, 422 aa)[3,7-10]. The functions of some non-structural proteins, such as polyproteins of the replicase complex encoded by Orf1a and Orf1b, have also been identified.

A characteristic of RNA viruses is their genetic instability, which enables the viruses to escape attack by the host immune system and to change their host range and tissue tropism more frequently. On the other hand, over a long epidemic period, the rate of synonymous mutations of the coding sequences of SARS-CoV was constant, while the rate of non-synonymous mutation (amino acid substitution) decreased[4,5]. Pairwise analysis of the Ka/Ks for the genotypes in each epidemic phase showed that the average Ka/Ks for the early phase was significantly higher than that of the middle phase, and even higher than that in the later phase[4, 5]. In this work, the genome of one SARS-CoV isolated directly from the stool sample of a SARS patient was completely sequenced. Comparative genomics analysis was performed to reveal the biological characteristic of the SARS-CoV genome.

1 Materials and Methods

1.1 SARS patient and sample collection

A male patient was hospitalized in Beijing Youan Hospital on April 29, 2003, a week after the onset of the respiratory disease. Based on the clinical characteristics that satisfied the WHO case definitions (http://www.who.int/csr/sars/casedefinition/en/), he was diagnosed as a SARS patient. One stool sample for direct SARS-CoV genome sequencing was collected on May 23, 2003 (31 d after onset of the disease), from which SARS-CoV BJ202 genome sequence was completed.

1.2 Viral RNA extraction and RT-PCR method

About 0.2 gram of stool sample was taken into a 1.5 mL tube, and 1 mL of 0.8% NaCl was added. The tube was vortexed to suspend the stool sample thoroughly and then was centrifuged for 3 min at 5 000 r/min. About 140 μL of the supernatant was collected to extract viral RNA with an RNA extraction kit (Qiagen). The RNA was dissolved in 100 μL diethyl pyrocarbonate (DEPC)-treated water containing 1 U DNase I (Promega). cDNA was synthesized by reverse transcription from 10 μL RNA at 45℃ for 50 min in a 20 μL solution containing 50 mmo/L Tris/HCl (pH 8.3), 75 mmol/L KCl, 3 mmol/L $MgCl_2$, 10 mmol/L DTT, 100 ng random hexamer primers, 200 U Moloney murine leukemia virus reverse transcriptase (MMLV, Promega), 25 U RNasin (Promega) and 0.5 mmol/L dNTPs. According to GZ02 genome sequence, 75 pairs of PCR primers covering the whole genome were designed. The primary PCR was carried out in a 25 μL mixture containing 2 μL cDNA, 10 mmol/L Tris/HCl (pH 8.4), 50 mmol/L KCl, 2.5 mmol/L $MgCl_2$, 100 μmol/L dNTPs, 1 U *Taq* DNA polymerase (Promega), 0.25 μmol/L forward and reverse primers, respectively.

1.3 Cloning of PCR products

PCR products containing polymorphic sites were ligated into the T-vector (Promega). About 20 recombinant clones from each ligation were isolated and sequenced. Only the high quality nucleotides (Phred/Phrap/Consed, >Q40) at the polymorphic sites were calculated.

1.4 Sequencing and sequence assembly

The PCR products were used for direct sequencing analysis on ABI 377 sequencers (Applied Biosystems) and MegaBACE 1 000 (Amersham). We used Phred/Phrap/Consed package version 13.0 for processing all of the raw sequence data. 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 using Phrap (http://www.phrap.org). The gaps, as well as the regions with low quality

data, identified after the preliminary assembly, were filled in or refined by re-sequencing the PCR products.

1.5 Annotation and comparative genome analysis

115 complete SARS-CoV genome sequences for comparative analysis were retrieved from NCBI (Table 1). All the open reading frames were identified with ORF Finder (http://www.ncbi.nlm.nih.gov/gOrf/Gorf.html). Comparative analysis was performed using BLAST against the nr (non-redundant) database

Table 1 115 SARS-CoV genomes for comparative genomics analysis

Viral strain	Accession No.	Viral strain	Accession No.	Viral strain	Accession No.
TJF	AY654624	TW8	AY502931	CUHK-AG03	AY345988
WH20	AY772062	TW7	AY502930	CUHK-AG02	AY345987
CFB/SZ/94/03	AY545919	TW6	AY502929	CUHK-AG01	AY345986
HC/GZ/32/03	AY545918	TW5	AY502928	CUHK-Su10	AY282752
HC/GZ/81/03	AY545917	TW4	AY502927	PUMC03	AY357076
HC/SZ/266/03	AY545916	TW3	AY502926	PUMC02	AY357075
HC/SZ/DM1/03	AY545915	TW2	AY502925	PUMC01	AY350750
HC/SZ/79/03	AY545914	TW11	AY502924	GZ50	AY304495
PC4-227	AY613950	TW10	AY502923	SZ16	AY304488
PC4-136	AY613949	TW1	AY291451	SZ3	AY304486
PC4-13	AY613948	GZ02	AY390556	AS	AY427439
GZ0402	AY613947	ZS-C	AY395003	Sin2774	AY283798
GZ0401	AY568539	LC5	AY395002	HKU-39849	AY278491
GD69	AY313906	LC4	AY395001	GD01	AY278489
HC/SZ/61/03	AY515512	LC3	AY395000	TWC3	AY362699
CDC#200301157	AY714217	LC2	AY394999	TWC2	AY362698
Sin3408L	AY559097	LC1	AY394998	Sin2748	AY283797
Sin850	AY559096	ZS-A	AY394997	Sin2679	AY283796
Sin847	AY559095	ZS-B	AY394996	Sin2677	AY283795
Sin846	AY559094	HSZ-Cc	AY394995	Sin2500	AY283794
Sin845	AY559093	HSZ-Bc	AY394994	Urbani	AY278741
SinP5	AY559092	HGZ8L2	AY394993	ZMY 1	AY351680
SinP4	AY559091	HZS2-C	AY394992	TWY	AP006561
SinP3	AY559090	HZS2-Fc	AY394991	TWS	AP006560
SinP2	AY559089	HZS2-E	AY394990	TWK	AP006559
SinP1	AY559088	HZS2-D	AY394989	TWJ	AP006558
Sin3725V	AY559087	HZS2-Fb	AY394987	TWH	AP006557
Sin849	AY559086	HSZ-Cb	AY394986	CUHK-W1	AY278554
Sin848	AY559085	HSZ-Bb	AY394985	Taiwan TC3	AY348314
Sin3765V	AY559084	HSZ2-A	AY394983	Taiwan TC2	AY338175
Sin3408	AY559083	GZ-C	AY394979	Taiwan TC1	AY338174
Sin852	AY559082	GZ-B	AY394978	TWC	AY321118
Sin842	AY559081	NS-1	AY508724	BJ04	AY279354
LLJ-2004	AY595412	ShanghaiQXC1	AY463059	BJ03	AY278490
WHU	AY394850	ShanghaiQXC2	AY463060	BJ02	AY278487
TOR2	AY274119	FRA	AY310120	ZJ01	AY297028
HSR 1	AY323977	SoD	AY461660	BJ01	AY278488
Frankfurt 1	AY291315	Sino3-11	AY485278		
TW9	AY502932	Sino1-11	AY485277		

(http://www.ncbi.nlm.nih.gov/blast/) for nucleic acid and protein sequences, and the multiple sequence alignment was deployed using ClustalW1.8 (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW). TreeView (Win32) version 1.66 and MEGA 2.0 were used to draw the phylogenetic tree.

2 Results

2.1 Characteristics of the BJ202 genome

The complete genome sequence of BJ202 was 29 751 bp in length. Compared with the genome sequence of GZ02, it had a deletion of 29 bp corresponding to the region of 27 884–27 912 bp (Orf10-11) in the GZ02 genome. Except for the polyA region of the 3′ end, no other deletions or insertions were identified in the BJ202 genome.

Comparative analysis between BJ202 and GZ02 genomes revealed 41 polymorphic sites between them. A total of 38 variations could be found in other published SARS-CoV genomes, of which one polymorphic site (at position 26 428, G-A) was only found in TJF and Sin2500. The remaining 37 sites could be found in three or more SARS-CoV genomes. There were three unique variations in the BJ202 genome (at positions 13 804, T-C, 15 031, C-T and 20 792, T-A, Table 2). It is interesting to find a unique polymorphic site 15 031 in BJ202 in a region with the lowest mutation frequency (Fig.1).

2.2 The distribution of polymorphic sites in 116 SARS-CoV genomes

We aligned 116 complete genome sequences of SARS-CoV (including BJ202) to analyze their single nucleotide polymorphism (SNPs). There were

Table 2 Sequence comparisons between BJ202 and GZ02

Position in genome	GZ02	BJ202	AA change	Position in ORF	ORF	Position in genome	GZ02	BJ202	AA change	Position in ORF	ORF
508	T	G	Cys-Gly	{82}	Orf1a	20840	A	G	S[a]	{2481}	Orf1b
1206	C	T	S[a]	{314}	Orf1a	20992	A	G	Lys-Arg	{2532}	Orf1b
3326	C	T	Ala-Val	{1021}	Orf1a	21479	T	C	S[a]	{2694}	Orf1b
3626	C	T	Thr-Ile	{1121}	Orf1a	22145	C	T	S[a]	{218}	OrfS
4220	G	A	Arg-Lys	{1319}	Orf1a	22207	T	C	Leu-Ser	{239}	OrfS
5251	A	C	Ile-Leu	{1663}	Orf1a	22422	A	G	Arg-Gly	{311}	OrfS
6612	T	G	Phe-Leu	{2116}	Orf1a	22517	G	A	S[a]	{342}	OrfS
6929	A	G	Tyr-Cys	{2222}	Orf1a	22522	G	A	Arg-Lys	{344}	OrfS
8502	G	T	Trp-Cys	{2746}	Orf1a	23823	G	T	Asp-Tyr	{778}	OrfS
8559	C	T	S[a]	{2765}	Orf1a	24566	C	T	S[a]	{1025}	OrfS
8815	T	C	S[a]	{2851}	Orf1a	24978	G	A	Glu-Lys	{1163}	OrfS
8946	A	T	S[a]	{2894}	Orf1a	25779	C	A	Ala-Glu,	{171},	Orf3,
9095	T	C	Ile-Thr	{2944}	Orf1a				Gln-Lys	{31}	Orf4
9176	C	T	Ala-Val	{2971}	Orf1a	25844	T	A	Trp-Arg,	{193},	Orf3,
9479	C	T	Ala-Val	{3072}	Orf1a				S[a]	{52}	Orf4
9854	C	T	Ala-Val	{3197}	Orf1a	26032	A	T	S[a],	{255},	Orf3,
10029	A	G	S[a]	{3255}	Orf1a				Gln-Leu	{115}	Orf4
13804[b]	T	C	Ile-Th	{136	Orf1b	26428	G	A	Glu-Lys	{11}	OrfM
15031[b]	C	T	Ala-Val	{545}	Orf1b	26586	C	T	S[a]	{63}	OrfM
17131	C	T	Ser-Leu	{1245}	Orf1b	27243	C	T	Pro-Leu	{57}	Orf7
19838	A	G	S[a]	{2147}	Orf1b	28118	T	C	Phe-Leu	{114}	Orf10-11
20792[b]	T	A	S[a]	{2465}	Orf1b	29276	A	G	S[a]	{376}	OrfN

[a] Synonymous mutations;

[b] Unique variations in BJ202 genomes.

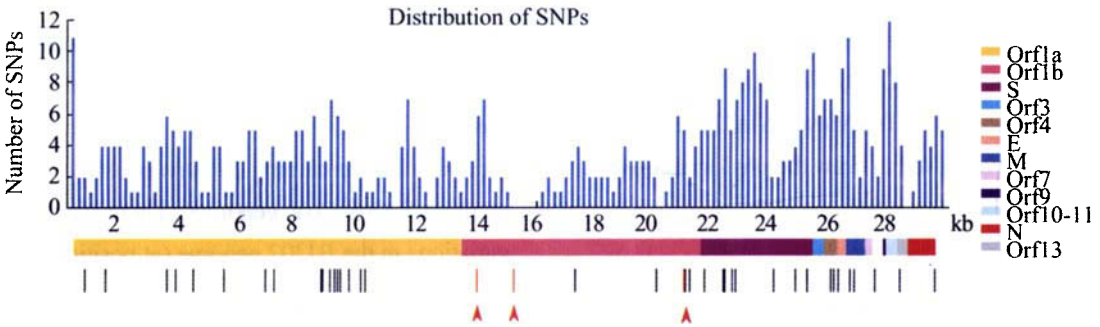

Fig. 1 Distribution of polymorphic sites over the SARS-CoV genome

278 recurrent polymorphic sites are illustrated in columns in the upper part of the diagram (The diagram was drawn with 400 bp windows and 200 bp intervals). Red triangles show the unique polymorphic sites and 41 short lines above the red triangles represent 41 polymorphic sites in the BJ202 genome.

278 polymorphic sites present in at least two genome sequences, and 170 sites present in at least three genome sequences. Nearly half of the variations (50.4%, 140/278) were located in the last one third of the genome (19.0–29.7 kb). The genomic region of 27.7–28.3 kb, which encodes Orf10-11, had the highest mutation frequency, where about 2.0% (0.6/29.7 kb) of the genome clustered 6.1% (17/278) of the polymorphic sites. The average mutation rate in this region was up to 28.3 per kb (17/0.6 kb), more than 3 times of that of the whole genome (278/29.7 kb, or 9.4 polymorphic sites per kb). The 25.2–26.8 kb region which encodes Orf3/4, E protein and M protein had the second highest mutation frequency in that about 5.4% (1.6/29.7 kb) of the sequence contained 12.2% (34/278) of the polymorphic sites. The average mutation rate in this region was up to 21.3 per kb (34/1.6 kb), more than 2 times that of the whole genome. The 21.9–23.9 kb region, which falls into OrfS, had the third highest mutation frequency, in which 39 polymorphic sites were found in the nearly 2 kb stretch of genomic sequence (19.5 SNPs in 1 kb). The other two thirds of the SARS-CoV genomic sequence had a lower density of polymorphic sites except for the 6.9–9.9 kb region within Orf1a, which contained 32 polymorphic sites in the 3 kb sequence (or 10.6 polymorphic sites per kb). The region of 14.4–17.3 kb had the lowest mutation frequency of the whole genome. Only 6 polymorphic sites were scattered over nearly 3 kb of this part of the genome (Fig. 1).

2.3 Nucleotide composition of the polymorphic sites in BJ202 genome

We cloned 15 PCR products (about 6.0 kb of the genome) including 11 fragments containing 12 known polymorphic sites and 4 fragments without identified polymorphic sites. Results showed that all 3 unique polymorphic sites of BJ202 (positions 13 804, 15 031 and 20 792) contained 2 kinds of nucleotides. Polymorphic sites 26 428, 26 477 and 27 243 also contained mixed nucleotides. It is interesting to find that position 18379 which has not been identified to be polymorphic in any of the other 115 published SARS-CoV genomes is a polymorphic site. The nucleotide composition of this site is A (8) to G (6) (Table 3).

2.4 Deletions and insertions

Among 116 SARS-CoV genomes, 18 types of deletions and 2 insertions were identified. Most of them were related to a 300 bp region (27 700–28 000 bp of the genome) which encoded part of the putative ORF9 and ORF10-11. Eighty-six genomes had the 29 bp deletion. 10 genomes had 2 deletions and one genome (GZ-C) had deletions of 3 fragments (D12, D14 and D18). Seventeen genomes (including GZ0401, GZ0402, HC/SZ/79/03, HSZ-Bc, SZ16, HC/SZ/266/ 03, HC/SZ/61/03, CFB/SZ/94/03, HC/SZ/DM1/03, HSZ-Cb, HSZ-Cc, GD01, HC/GZ/32/03, GZ02, SZ3, PC4-136 and PC4-13) were free of any deletion. They mainly constituted genomes of SARS-CoV isolated from early phase clinical samples or from animals (Table 4)

Table 3 Nucleotide composition of some polymorphic sites in BJ202 genome

Position	Nucleotide in GZ02	Polymorphic nucleotide	Dominant nucleotide in BJ202	Polymorphism of genomes [a]	Polymorphism of clones [b]
9404	C	T	C	33-83	16-0
9854	C	T	T	105-11	0-16
13804 [c]	T	C	C	115-1	7-15
15031 [c]	C	T	T	115-1	6-7
17564	G	T	G	42-74	13-0
18379 [d]	A	G	A	116-0	8-6
19838	A	G	G	105-11	0-12
20792 [c]	T	A	A	115-1	2-10
25298	G	A	G	113-3	12-0
26428	G	A	A	113-3	16-20
26477	T	G	T	88-28	9-7
27243	C	T	T	106-10	1-20
27827	C	T	C	45-71	13-0

[a] Number of genomes with the same nucleotides as GZ02 to number of genomes with the polymorphic nucleotides;

[b] Number of clones with the same nucleotide as GZ02 to the number of clones with the polymorphic nucleotide;

[c] Unique polymorphic sites in BJ202 genome;

[d] Not identified as polymorphic in any of the other 115 SARS-CoV genomes, but polymorphic in the genome of BJ202. The nucleotide ratio of A to G is 8 to 6.

Table 4 Deletions and insertions in 116 SARS-CoV genomes

Types [a]	Length (bp)	Start position	End position	ORF	Genome
D01	578	5914	6491	ORF1a	SHANGHAIQXC2
D02	2	27067	27068		TWJ [b]
D03	2	27167	27168	ORF7	TWJ [b]
D04	277	27411	27687	ORF9	WH20
D05	137	27677	27813	ORF9,ORF10-11	SIN846
D06	49	27761	27809	ORF9,ORF10-11	SIN849
D07	6	27782	27787	ORF10-11	SIN2677
D08	57	27798	27854	ORF10-11	SIN852
D09	5	27810	27814	ORF10-11	SIN2748
D10	2	27808	27809	ORF10-11	TWC, WHU
D11	415	27719	28133	ORF9,ORF10-11	LC2 [b], LC3 [b], LC4 [b], LC5 [b]
D12	39	27771	27809	ORF10-11	GZ-C [c], GZ-B
D13	82	27858	27939	ORF10-11	ZS-A [b], ZS-B [b], ZS-C [b]
D14	29	27884	27912	ORF10-11	86 viral strains
D15	31	27883	27913	ORF10-11	Sino3-11 [b], PUMC02 [b]
D16	32	27882	27913	ORF10-11	GZ-C [c]
D17	2	27912	27913	ORF10-11	HC/GZ/81/03, PC4-227 [b]
D18	12	28162	28173	ORFN	GZ-C [c]
I01 [d]	10	20373	20374	ORF1b	GD69
I02 [d]	6	27658	27659	ORF9	LLJ2004

[a] D1-D18 are deletions and I01-I02 are insertions;

[b] SARS-CoV genomes with 2 kinds of deletions. TWJ has not the 29 bp (D14) deletion and the rest have the 29 bp deletion (D14);

[c] GZ-C genome with 3 deletions;

[d] I01 is inserted between 20 373 and 20 374 in GD69 genome and I02 is inserted between 27 658 and 27 659 in LLJ2004 genome.

2.5 Phylogeny

A phylogenetic tree based on the divergence of whole genome from BJ202 and the other 115 completed SARS-CoVs places BJ202 closest to BJ01 (Fig.2).

3 Discussion

To date, complete genomic sequences of 115 SARS-CoVs are available in NCBI GenBank, which provides a foundation for a better understanding of the polymorphism and molecular evolution of SARS-CoV.

As a member of RNA viruses, the genome of SARS-CoV has a higher mutation rate than DNA viruses[11]. Using GZ02 genome as the reference, nearly 685 (N/M/W/R/Y and SNPs located in 100 bp region of the 5′ and 3′ end not included) polymorphic sites have been identified in all 116 completed SARS-CoV genomes, among which 278 polymorphic sites are present in at least two genomes. The density of the polymorphic sites is up to 9.4 per kb (278/29.7 kb) of the genome.

The skewed distribution of the polymorphic sites over the genome is another characteristic of the SARS-CoV genome. More than half of the variations (50.4%, 140/278) have been identified in the last one third of the genome (19–29 kb). The region encoding Orf10-11, Orf3/4, E, M and S protein are more variable than other parts of the genome. About 32.4% (90/278) of all the polymorphic sites are clustered in about 14.1% (4.2/29.7 kb) of the genome.

Forty-one polymorphic sites have been identified in BJ202 genome. Similar to the polymorphic sites identified in other 115 SARS-CoV genomes, they are not evenly distributed over the whole genome. Four regions (8.5–10.1, 19.8–21.0, 22.1–22.6 and 25.7–26.6 kb) which cover only 14.1% (4.2/29.7 kb) of the whole genome take up 56.1% of all the polymorphic sites (23/41; 9, 4, 5 and 5 polymorphic sites in each respective region). These regions encode parts of Orf1a (265–13 398 bp), Orf1b (13 398–21 485 bp), S protein (21 492–25 259 bp), Orf3/Orf4 (25 268–26 153) and M protein (26 398–27 063), respectively.

In contrast, other regions of the SARS-CoV genome are highly conserved [12]. The 14.4–17.3 kb region has the lowest mutation frequency. Only 6 polymorphic sites scatter over this part of the genome which is nearly 3 kb long. This region, within Orf1b, encodes part of the RdRp (NSP9, 13379–16147) and HEL (NSP10, 16148–17950). NSP9 is a non-structural protein and has RNA-dependent RNA polymerase activity [13].

To verify the polymorphic sites in BJ202, we sequenced some cloned PCR products. It is interesting to find that 3 unique polymorphic sites (positions 13 804, 15 031 and 20 792) in the BJ202 genome were all composed of 2 different nucleotides. Positions 26428, 26477 and 27243, which were not unique polymorphic sites to BJ202 genome, were also composed of mixed nucleotides. Position 18379 was actually polymorphic, although it had not been discovered as such in any of the other SARS-CoV genomes. Among 14 clones sequenced, 8 were nucleotide A and 6 were nucleotide G at this position. The results verifies the polymorphic nature of the SARS-CoV genome as previously reported[14]. It also warns us that when we directly sequence the PCR products, we may fail to identify many polymorphic sites.

Mapping the deletions and insertions in SARS-CoV genome has been used to analyze genotype groups and the molecular evolution of SARS-CoV [15]. To date, there have been about 18 kinds of deletions and 2 insertions ranging from 2 to 578 bp in length identified in the 116 SARS-CoV genomes including BJ202. The deletions are located in Orf9, Orf1a, Orf7, OrfN and mostly in the region encoding Orf10-11. As discussed above, the region encoding Orf10-11 also has the highest frequency of sequence variations. Of the two insertions, one (10 bp) is in Orf1b of GD69 genome and the other (6 bp) in Orf9 of the LLJ2004 genome. Only early phase clinical SARS-CoVs (such as GD01 and GZ02 etc.) and animal origin SARS-CoVs (SZ3 and SZ16 etc.) are free of deletion [15]. It might mean that Orf10-11 is related to host range or tissue tropism of SARS-CoV against animals and is not a contributing factor in the infection against human beings.

Most SARS-CoV genomic sequences in the NCBI GenBank were from viral isolates of Vero E6 cell

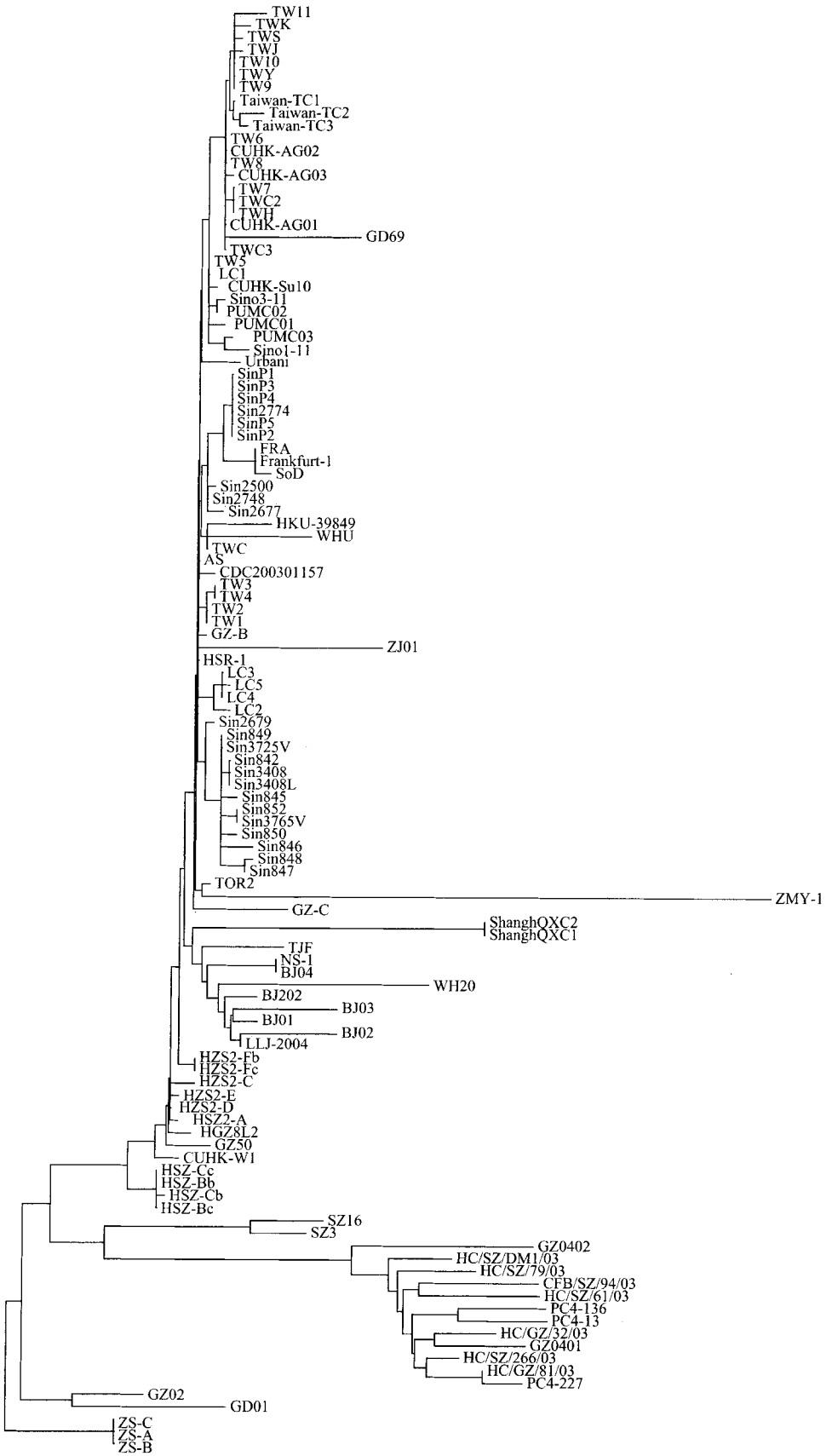

Fig. 2 Phylogenetic tree of 116 complete SARS-CoV genomes

The tree was constructed using the nucleotide number of differences. Bootstrap=1 000.

culture [3, 7, 16]. Only a few of them were directly from clinical samples[5, 17]. Although it was reported that the *in vitro* mutation rate of the SARS-CoV in Vero cell passage was negligible[18], there might be difference between the genomic sequences obtained directly from clinical samples and from isolates of the cell culture. As we know from this work, when SARS-CoVs of different genotypes were mixed in the same sample, only the dominant genotypes would be identified by the strategy of sequencing the RT-PCR products. Moreover, when the clinical sample is subjected to cell culture, viruses with different genotypes duplicate at different rates. Those with dominant genotypes in the sample might not be the prominently duplicated ones and would be diluted by their counterparts during the cell culture. Consequently, it is difficult to effectively identify the polymorphic sites induced *in vivo* by sequencing the RT-PCR products of the cell culture isolates.

As a member of the RNA viruses, SARS-CoV has a higher mutation rate than DNA viruses. However, most genomic sequences of SARS-CoV available in the NCBI GenBank were derived from viral isolates of cell cultures, which might cause some problems when we analyze the sequence variations of SARS-CoV, because sequence from the cell culture isolates might be different from the direct clinical sample. In this work, we completed the sequencing of SARS-CoV genome directly from the stool sample and analyzed the polymorphism of the SARS-CoV genome. All these would provide valuable experimental data for further research in the mechanism of SARS-CoV mutation.

Acknowledgements: We are indebted to collaborators and clinicians from Hangzhou Genomics Institute, Beijing Genomics Institute and Beijing Youan Hospital.

References:

[1] World Health Organization. Severe acute respiratory syndrome (SARS). *Wkly Epidemiol Rec*, 2003, 78 : 81-83.

[2] Groneberg D A, Zhang L, Welte T, Zabel P, Chung K F. Severe acute respiratory syndrome: global initiatives for disease diagnosis. *QJM*, 2003, 96 (11) : 845-852.

[3] Marra M A, Jones S J, Astell C R, Holt R A, Brooks-Wilson A, Butterfield Y S, Khattra J, Asano J K,Barber S A, Chan S Y, Cloutier A, Coughlin S M, Freeman D, Girn N, Griffith O L, Leach S R, Mayo M,McDonald H, Montgomery S B, Pandoh P K, Petrescu A S, Robertson A G, Schein J E, Siddiqui A, Smailus D E, Stott J M, Yang G S, Plummer F, Andonov A, Artsob H, Bastien N, Bernard K, Booth T F, Bowness D,Czub M, Drebot M, Fernando L, Flick R, Garbutt M, Gray M, Grolla A, Jones S, Feldmann H, Meyers A,Kabani A, Li Y, Normand S, Stroher U, Tipples G A, Tyler S, Vogrig R, Ward D, Watson B, Brunham R C,Krajden M, Petric M, Skowronski D M, Upton C, Roper R L. The genome sequence of the SARS-associated *coronavirus*. *Science*, 2003, 300 (5624) : 1399-1404.

[4] Guan Y, Zheng B J, He Y Q, Liu X L, Zhuang Z X, Cheung C L, Luo S W, Li P H, Zhang L J, Guan Y J, Butt K M, Wong K L, Chan K W, Lim W, Shortridge K F, Yuen K Y, Peiris J S M, Poon L L M. Isolation and characterization of viruses related to the SARS *coronavirus* from animals in Southern China. *Science*, 2003, 302 (5643) : 276-278.

[5] Chinese SARS Molecular Epidemiology Consortium. Molecular evolution of the SARS *coronavirus* during the course of the SARS epidemic in China. *Science*, 2004, 303 (5664) : 1666-1669.

[6] Rota P A, Oberste M S, Monroe S S, Nix W A, Campagnoli R, Icenogle J P, Penaranda S, Bankamp B, Maher K, Chen M H, Tong S, Tamin A, Lowe L, Frace M, DeRisi J L, Chen Q, Wang D, Erdman D D, Peret T C, Burns C, Ksiazek T G, Rollin P E, Sanchez A, Liffick S, Holloway B, Limor J, McCaustland K, Olsen-Rasmussen M, Fouchier R, Gunther S, Osterhaus A D, Drosten C, Pallansch M A, Anderson L J, Bellini W J. Characterization of a novel *coronavirus* associated with severe acute respiratory syndrome. *Science*, 2003, 300 (5624) : 1394-1399.

[7] Qin E'de, Zhu Q Y, Yu M, Fan B C, Chang G H, Si B Y, Yang B A, Peng W M, Jiang T, Liu B H, Deng Y Q, Liu H, Zhang Y, Wang Cui'e, Li Y Q, Gan Y H, Li X Y, Lu F H, Tan G, Cao W H, Yang R F. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chinese Science Bulletin*, 2003, 48 (10) : 941-948.

[8] Yeh S H, Wang H Y, Tsai C Y, Kao C L, Yang J Y, Liu H W, Su I J, Tsai S F, Chen D S, Chen P J, National Taiwan University SARS Research Team. Characterization of severe acute respiratory syndrome *coronavirus* genomes in Taiwan: Molecular epidemiology and genome evolution. *PNAS*, 2004, 101 (8) : 2542-2547.

[9] Simmons G, Reeves J D, Rennekamp A J, Amberg S M, Piefer A J, Bates P. Characterization of severe acute res-

piratory syndrome-associated *coronavirus* (SARS-CoV) spike glycoprotein-mediated viral entry. *PNAS*, 2004, 101 (12) : 4240-4245.

[10] CHEN Yin-Jia, GAO Ge, BAO Yi-Ming, Lotez R, WU Jian-Min, CAI Tao, YE Zhi-Qing, GU Xiao-Cheng, LUO Jing-Chu. Initial analysis of complete genome sequences of SARS *coronavirus*. *Acta Genetica Sinica*, 2003, 30 (6) : 493-500 (in Chinese with an English abstract).

[11] Ruan Y J, Wei C L, Ee A L, Vega V B, Thoreau H, Su S T, Chia J M, Ng P, Chiu K P, Lim L, Zhang T, Peng C K, Lin E O, Lee N M, Yee S L, Ng L F, Chee R E, Stanton L W, Long P M, Liu E T. Comparative full-length genome sequence analysis of 14 SARS *coronavirus* isolates and common mutations associated with putative origins of infection. *Lancet* , 2003, 361 (9371) : 1779-1785.

[12] Xu Z, Zhang H, Tian X, Ji J, Li W, Li Y, Tian W, Han Y, Wang L, Zhang Z, Xu J, Wei W, Zhu J, Sun H, Zhang X, Zhou J, Li S, Wang J, Wang J, Bi S, Yang H. The R protein of SARS-CoV: analyses of structure and function based on four complete genome sequences of isolates BJ01-BJ04. *Genomics Proteomics Bioinformatics*, 2003, 1 (2) : 155-165.

[13] XU Xiang, LIU Yun-Qing, Weiss S, Arnold E, Sarafianos S G, DING Jian-Ping. Molecular model of SARS *coronavirus* polymerase: implications for biochemical functions and drug design. *Nucleic Acids Res*, 2003, 31 (24) : 7117-7130.

[14] XU Dong-Ping, WANG Fu-Sheng, ZHANG Ling-Xia. Genetic variation analysis of SARS *coronavirus*. *Acta Genetica Sinica,* 2004, 31 (6) : 634-40 (in Chinese with an English abstract).

[15] Pavlovic-Lazetic G M, Mitic N S, Beljanski M V. Bioinformatics analysis of SARS *coronavirus* genome polymorphism. *BMC Bioinformatics*, 2004, 5 (1) : 65.

[16] Stephen T K W, Stephen C S C, Dennis L Y M. *Coronavirus* genomic-sequence variations and the epidemiology of the severea cute respiratory syndrome. *N Engl J Med*, 2003, 349 (2) : 187-188.

[17] Chim S S C, Tsui S K W, Chan K C A, Au T C C, Hung E C W, Tong Y K, Chiu R W K, Ng E K O, Chan P K S, Chu C M, Sung J J Y, Tam J S, Fung K P, Waye M M Y, Lee C Y, Yuen K Y, Lo Y M D, members of the CUHK Molecular SARS Research Group. Genomic characterization of the severe acute respiratory syndrome *coronavirus* of Amoy Gardens outbreak in Hong Kong. *Lancet*, 2003, 362 (9398) : 1807-1808.

[18] Vega V B, Ruan Y, Liu J, Lee W H, Wei C L, Se-Thoe S Y, Tang K F, Zhang T, Kolatkar P R, Ooi E E, Ling A E, Stanton L W, Long P M, Liu E T. Mutational dynamics of the SARS *coronavirus* in cell culture and human populations isolated in 2003. *BMC Infect Dis,* 2004, 4 (1) : 32-40.

SARS 病毒基因组的多态性

商 磊[1,3*]，齐 艳[2*]，包其郁[2]，田 薇[1]，徐建成[1]，冯明光[3]，杨焕明[1]

1．浙江大学沃森基因组科学研究院，杭州 310008；

2．温州医学院生物医学信息研究所，温州 325000；

3．浙江大学微生物研究所，杭州 310029

摘 要：对 SARS 病人粪便样本直接测序，得到 SRAS-CoV BJ202 全基因组序列(AY864806)。应用比较基因组研究方法对 GenBank 中公布的 115 株 SARS-CoV 基因组序列以及 BJ202 进行分析。以 GZ02 序列为参照，发现 2 个以上基因组中同时存在单核苷酸多态(SNP)位点共 278 个。多态位点在 SARS-CoV 基因组中呈偏态分布，大约一半突变位点(50.4%, 140/278)发生在基因组 3′末端 1/3 区域。编码 Orf10-11、Orf3/4、E 蛋白、M 蛋白和 S 蛋白区域突变率较高。克隆并测序含有 BJ202 基因组 12 个多态位点的 11 个 cDNA 以及 4 个不含已知多态位点的 cDNA 片段（15 个片段总长度为 6.0 kb），结果显示：BJ202 特有的 3 个多态位点（13 804、15 031 和 20 792）以及另外 3 个多态位点（26 428、26 477 和 27 243）均检出两种不同核苷酸；位点 18 379 虽在已公布的 115 株 SARS-CoV 基因组中未发现突变，实际上也是多态位点。14 个克隆中有 8 个克隆该位点为 A，6 个克隆为 G。全部 116 个 SARS-CoV 基因组中共有 18 种缺失类型和 2 种插入类型。大部分缺失发生在编码 ORF9 和 ORF10-11 区域(基因组序列 27 700-28 000 bp 处)。以邻位连接法（Neighbor-Joining）构建了 116 株 SARS-CoV 系统发育树，BJ202 与 BJ01 和 LLJ-2004 等 SARS-CoV 的亲缘关系较接近。

关键词：SARS 相关冠状病毒；基因组；多态性

作者简介：商磊（1979-），男，浙江杭州人，硕士研究生，专业方向：生物信息学

齐艳（1978-），女，山西长治人，硕士，专业方向：临床检验与诊断学