Complete Genome Sequences of the SARS-CoV: the BJ Group (Isolates BJ01-BJ04)

Shengli Bi[1]*, E'de Qin[2]*, Zuyuan Xu[3,4]*, Wei Li[3]*, Jing Wang[5,3]*, Yongwu Hu[6,3]*, Yong Liu[7]*, Shumin Duan[1], Jianfei Hu[3,5], Yujun Han[3], Jing Xu[3], Yan Li[3], Yao Yi[1], Yongdong Zhou[1], Wei Lin[1], Jie Wen[3], Hong Xu[1], Ruan Li[1], Zizhang Zhang[8], Haiyan Sun[3], Jingui Zhu[3], Man Yu[2], Baochang Fan[2], Qingfa Wu[3], Wei Lin[3], Lin Tang[3], Bao'an Yang[2], Guoqing Li[3], Wenming Peng[2], Wenjie Li[3], Tao Jiang[2], Yajun Deng[3], Bohua Liu[2], Jianping Shi[3], Yongqiang Deng[2], Wei Wei[4], Hong Liu[2], Zongzhong Tong[3], Feng Zhang[3], Yu Zhang[2], Cui'e Wang[2], Yuquan Li[2], Jia Ye[3,4], Yonghua Gan[2], Jia Ji[3], Xiaoyu Li[2], Xiangjun Tian[3,4], Fushuang Lu[2], Gang Tan[2], Ruifu Yang[2], Bin Liu[3], Siqi Liu[3], Songgang Li[3,5], Jun Wang[3],Jian Wang[3,4], Wuchun Cao[2], Jun Yu[3,4#], Xiaoping Dong[1#], and Huanming Yang[3,4#]

[1] Center of Disease Control and Prevention, Beijing 100050, China; [2] Institute of Microbiology and Epidemiology, Chinese Academy of Military Medical Sciences, Beijing 100071, China; [3] Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; [4] James D. Watson Institute of Genome Sciences, Zhijiang Campus, Zhejiang University and Hangzhou Genomics Institute, Hangzhou 310008, China; [5] College of Life Sciences, Peking University, Beijing 100871, China; [6] Wenzhou Medical College, Wenzhou 325003, China; [7] Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing 100730, China; [8] College of Materials Science and Chemical Engineering, Yuquan Campus, Zhejiang University, Hangzhou 310027, China.

Beijing has been one of the epicenters attacked most severely by the SARS-CoV (severe acute respiratory syndrome-associated coronavirus) since the first patient was diagnosed in one of the city's hospitals. We now report complete genome sequences of the BJ Group, including four isolates (Isolates BJ01, BJ02, BJ03, and BJ04) of the SARS-CoV. It is remarkable that all members of the BJ Group share a common haplotype, consisting of seven loci that differentiate the group from other isolates published to date. Among 42 substitutions uniquely identified from the BJ group, 32 are non-synonymous changes at the amino acid level. Rooted phylogenetic trees, proposed on the basis of haplotypes and other sequence variations of SARS-CoV isolates from Canada, USA, Singapore, and China, gave rise to different paradigms but positioned the BJ Group, together with the newly discovered GD01 (GD-*Ins*29) in the same clade, followed by the H-U Group (from Hong Kong to USA) and the H-T Group (from Hong Kong to Toronto), leaving the SP Group (Singapore) more distant. This result appears to suggest a possible transmission path from Guangdong to Beijing/Hong Kong, then to other countries and regions.

Key words: SARS, SARS-CoV, haplotype, substitution, phylogeny

China is the prime victim of the sudden outbreak of SARS (severe acute respiratory syndrome) due to a newly identified variant of coronavirus, the SARS-CoV (*1*). Beijing metropolitan area is one of the epicenters that have been severely attacked by the virus since its first patient was identified on March 1, 2003. Both the numbers of cases and of deaths of SARS in Beijing alone were accounted for 29.92% and 23.57%, respectively, of those in the entire world up to June 10, 2003 (Figure 1). We have been working on sequencing SARS-CoV genomes from clinical isolates since early April of this year and have released our data as soon as we acquired them. We now report our detailed comparative analyses on a group of four complete genome sequences (Isolates BJ01, BJ02, BJ03, and BJ04), named the BJ Group, by referencing the thirteen SARS-CoV genome sequences in the public databases (*2-4*).

*** These authors contributed equally to this work.**
[#] Corresponding authors.
E-mail: dongxp@public.fhnet.cn.net;
yanghm@genomics.org.cn

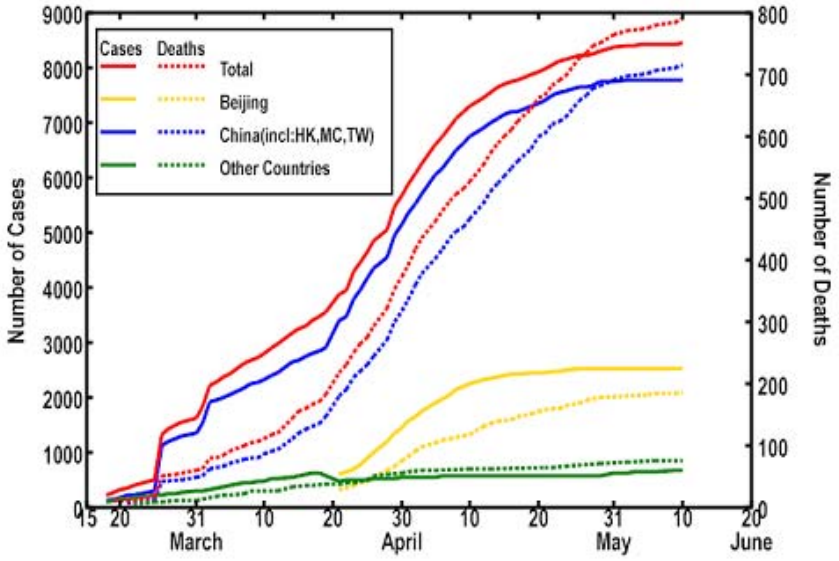

Fig. 1 Accumulated number of probable cases and deaths of SARS in Beijing, China, and the world. Data sources: http://www.moh.gov.cn/zhgl/yqfb/index.htm; http://www.who.int/csr/sars/country/en/.

SARS Patients, SARS-CoV Isolates and Genome Sequencing

SARS patients were clinically diagnosed in March 2003 according to World Health Organization (WHO) guidelines (http://www.who.int/csr/sars/guidelines /en/). SARS-CoV isolates were maintained in Vero-6 cell cultures that were inoculated from autopsies and biopsies of the deceased or recovered SARS patients (Table 1). Viral RNA was purified from virions prepared from the cultures and subjected for cDNA syntheses. A set of primers that cover the entire viral genome were used for the reverse transcription and PCR amplification, in a product size range of 400-800 bp. PCR-amplified fragments were then cloned into amplicon-libraries and two dozens or more clones were sequenced for each PCR-amplified fragments to ensure sequence quality and to avoid errors in the procedures of the RT-PCR and cloning. Only the consensus sequences with absolute majority votes were used for the genome assembly and gene annotation, although every read was assembled and some sequence variations were clearly visible. For comparative genomic analyses, 13 other full-length SARS-CoV genome sequences were downloaded from GenBank (Table 2; http://www.ncbi.nlm.nih.gov). The nucleotide positions of Isolate BJ01 were used as the reference for analyses.

Table 1 Samples and Clinical Data for the BJ Group (Isolates BJ01-BJ04)

Isolate	GenBank accession number	Tissue	Sample	Clinical outcome
BJ01	AY278488	Lung	Autopsy	Deceased
BJ02	AY278487	Nose & throat	Swabs (mixed patients infected by BJ01)	Recovered
BJ03	AY278490	Liver & lymph nodes	Autopsy (same as BJ01)	Deceased
BJ04	AY279354	Lung	Autopsy	Deceased

Table 2 Information Summary of the Complete Genome Sequences of the BJ Group and Other 13 Isolates of SARS-CoV*

Isolate	Genome size (nt)	Accession number	Modification date
BJ01	**29,725**	**AY278488.2**	**1-May-03**
BJ02	**29,745**	**AY278487**	**5-June-03**
BJ03	**29,740**	**AY278490**	**5-June-03**
BJ04	**29,732**	**AY279354**	**29-May-03**
GD01	29,757	AY278489	29-May-03
ZJ01	29,715	AY297028.1	19-May-03
TW1	29,729	AY291451.1	14-May-03
CUHK-W1	29,736	AY278554.2	14-May-03
CUHK-Su10	29,736	AY282752.1	7-May-03
Urbani	29,727	AY278741.1	21-Apr-03
HKU-39849	29,742	AY278491.2	18-Apr-03
TOR2	29,751	NC_ 004718.3	22-May-03
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

*Data were retrieved from GenBank: http://www.ncbi.nlm.nih.gov.

Isolates BJ01 and BJ03 were derived from the autopsied lung tissue (BJ01) and liver/lymph nodes (BJ03) of the same patient who was the father of the first patient (the Index Case of Beijing) diagnosed on March 1, 2003 in Beijing. His own daughter infected him in the last week of February in Shanxi Province after she traveled to Guangdong Province during the period of February 18th to 23rd. He was 53 years old and had no detectable symptoms of hepatitis, AIDS, or heart diseases. He, as the second SARS patient hospitalized (March 5th) in Beijing, died two days after the diagnosis. His daughter, however, had recovered completely from the horrifying ordeal after medical treatments. The sequence differences between Isolates BJ01 and BJ03 are expected to reflect the sequence variations within the same patient after viral infection. Isolate BJ02 was inoculated from pooled samples of nose/throat swabs from seven patients who were all evidenced to be infected by the first patient when she (the Index Case of Beijing) was hospitalized in the first week of March at the same hospital. All of these patients were recovered after effective treatments. The sequence differences between BJ01 and BJ02 should be regarded as the variation from the first circle or round of the infection. Isolate BJ04 was inoculated from the autopsied lung tissue of a single deceased patient who had no direct contact with the Index Case of Beijing. The sequence differences between BJ04 and other BJ cases should yield variations during viral infections outside the first circle. As a whole, the group of isolates, or the BJ Group, represents the early rounds of disease transmissions in Beijing. Another relevant case concerns Isolate GD01 (previously named GZ01). It was isolated from the autopsied lung tissue of a single deceased patient who was a female of 54 years old. 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. Its genome sequence serves as a root or an anchor point for the BJ Group. In addition, the genome sequence of GD01 harbors a 29-nt insertion, a rare genotype so far only found in the samples from Guangdong Province (*5*).

The sequences of the four complete genomes have a size range of 29,725-29,745 nucleotides (nt). The difference is primarily due to the extension of the 5′-end sequence, which has no significant changes in the overall genome structure and gene size. In the genomes of these isolates, the predicted number of ORFs (open reading frames) remains as 13, in contrast to the genome of Isolate GD01, where a 29-

nt insertion was discovered. Sequence evidence from another SARS-CoV isolate collected in Guangdong has confirmed this significant finding (BGI unpublished data). Referring to the BJ01 genome annotation, we have previously reported that, in the GD01 genome, two additional BGI-PUPs (BGI-Postulated Uncharacterized Proteins), BGI-PUP5 (nt positions 27,763-27,882) and BGI-PUP6 (nt positions 27,848-28,102), were predicted in the largest (478 nt) intergenic region between the ORFs for BGI-PUP4 and the N protein (nucleoprotein); PUP5 was previously so named in BJ01 and now renamed as BGI-PUP7. Putative leader sequences were determined in these genomes, based on alignments of the leader sequence at the 5′-end of the genome (5′-UCUCUAAACGAACUUUAAAAUCUG) to the sequences upstream of each ORF. The organization and general features of viral proteins, including the R (replicase) protein and the structural proteins (the spike or S, envelope or E, membrane or M, and N proteins), among all the isolates are nearly identical except the 29-nt insertion in the GD isolates and the limited number of nucleotide substitutions.

A Novel Haplotype Found Unique to the BJ Group

A salient discovery in our study is that all members of the BJ Group share a common haplotype of 7 loci, *C/t-T-G-C-A-C-G-C-T-C* (letters in uppercases indicate the major allele at the locus and letters in lowercases specify the minor allele; Table 3); a new mutation occurred in BJ04 at Locus 9,385 is responsible for the minor allele *t*. Isolate BJ04, which was harbored by a patient who did not have any direct contacts with the other BJ cases, is clustered with most members of the other groups. If GD01 is included in the group, the haplotype remains as *C/t-T/c-G-C-A-C-G-C-T-C*; it introduces another minor allele *c* at Locus 9,835. Even if we exclude either BJ01 or BJ03, the two isolates from a single patient, the haplotype still stands out; it has differentiated the group from all other isolates (*T/c-C-T/g-C/t-A/g-C/t-A-T/c-C-T/c*) identified so far. This BJ Group-specific haplotype represents the first population of SARS-CoV responsible for the acute outbreak in the metropolitan area of Beijing.

Table 3 The Group-Specific Haplotypes of SARS-CoV

ORF	Locus[a]	BJGroup					H-U Group			H-T Group				SP Group				
		BJ01	BJ02	BJ03	BJ04	GD01	CUHK-W1	CUHK-Su10	Urbani	HKU-39849		TOR2	ZJ01	TW1	SIN2679	SIN2748	SIN2774	SIN2500	SIN2677
	9,385	C[b]	C	C	t	C	c	T	T	T		T	T	T	T	T	T	T	T
	9,835	T	T	T	T	c	C	C	C	C		C	C	C	C	C	C	C	C
	17,545	G	G	G	G	G	g	T	T	T		T	T	T	T	T	T	T	T
R	17,827	C	C	C	C	C	T	T	c	C		C	C	C	C	C	C	C	C
	19,045	A	A	A	A	A	G	a	G	A		A	A	A	A	A	A	A	A
	19,065	C	C	C	C	C	C	C	C	C		C	C	C	c	T	T	T	T
	19,819	G	G	G	G	G	A	A	A	A		A	A	A	A	A	A	A	A
S	22,203	C	C	C	C	C	c	T	T	T		T	T	T	T	T	T	T	T
BGI-PUP3	27,224	T	T	T	T	T	C	C	C	C		C	C	C	C	C	C	C	C
BGI-PUP5	27,808	C	C	C	C	C	c	T	T	T		T	T	T	T	T	T	T	T

[a] A locus is defined by the nucleotide position with reference to the complete genome sequence of BJ01.
[b] Capital letters indicate the major allele with the highest frequency at the locus within the corresponding group, and low-case letters denote the minor allele. Shaded areas highlight the possible haplotypes.

The H-U Group (from Hong Kong to USA) has a different haplotype, *T/c-C-T/g-T/c-G/a-C-A-T/c-C-T/c*. The major allele T within the group at Locus 17,827 that is shared by CUHK-W1 and CUHK-Su10, together with the major allele *G* at Locus 19,045 shared by CUHK-W1 and Urbani (named after Dr. Carlo Urbani who was the first WHO officer to identify the outbreak of this new disease in an American businessman who had been admitted to a hospital in Hanoi), establishes the internal association within

the H-U Group. The fact that the H-U Group overlaps with the BJ Group by Isolate CUHK-W1 (*c-C-g-T-G-C-A-c-C-c*) suggests a possible link or path of transmission between the BJ and H-U Groups. The H-U Group appears to be a mosaic genotype of the BJ and other groups. The result suggests the existence of an intermediate state in the transmission path connecting China to the outside. In addition to the haplotype, *T-C-T-C-A-T-A-T-C-T*, identical to the H-T Group (from Hong Kong to Toronto) as well as Isolates ZJ01 (Zhejiang) and TW1 (Taiwan), the SP Group (Singapore) possesses a monolocus marker *T*, as opposed to *C* for all others at Locus 19,065, thus establishing the internal phylogenetic relationship within the group and differentiating it from all others. GD01 shares an almost identical haplotype with the BJ Group at the defined loci, except for the minor allele *c* at Locus 9,835, which distances itself from the BJ Group but links it to all the other groups.

Sequence Variations in the BJ Group in Comparison with other Isolates

137 sequence variations (142 if counted independently by each ORF) were defined by comparing the sequences of the BJ Group with other isolates, including 3 (GD01, ZJ01, and TW1) identified in China, and 10 others elsewhere (Figure 2, Table 4). Out of the total, 42 were contributed by the BJ Group alone, amounting to nearly 30% of the grand total. These variations were confirmed among over a dozen high-quality sequence reads from re-sequencing of the RT-PCR products directly and clones from the corresponding amplicon-derived libraries. Although possible sequence variations acquired during the limited generations in the host body and viral culture as well as sequencing errors arisen from the RT-PCR and amplification/cloning could not be easily excluded, the likelihood that all BJ isolates mutate in a similar way is rather slim since the sequencing was done in a systematic way with overlapping segments and high quality. The quality of complete genome sequences from other contributors was assessed by manually checking the submitted sequence traces when publicly available. In the substitutions identified from the BJ group, 76% are categorized as non-synonymous, slightly higher than the average (70%, 100/142) calculated from all 17 isolates. As a benchmark, we have summarized a few commonly used parameters for viral evolution studies together with those of two other RNA viruses, the HIV-1 and influenza (Table 5). The SARS-CoV in many ways is quite different from these two viruses, except that their genomes are all RNA in nature. In particular, the HIV lives with the host for a long time so the constant escape from the host immune system is of essence for its survival until it overwhelms the system (*6*). The influenza virus that has an 85-year recorded history of infecting humans causes recurrent annual epidemics until a novel virus rises to stir up major worldwide pandemics (*7*). Our results did not lead to any strong conclusions due to insufficient data points compared to both HIV and influenza but it is quite necessary to compare the SARS-CoV data with those of the two prevalent viruses from time to time. Aside from mutation rate calculations, a popular test for selection on a particular protein is the ratio of Ka/Ks (*8*). Generally, most non-synonymous SNPs (single nucleotide polymorphisms) are believed to be deleterious and rapidly removed from the given population of viruses by selection, leaving Ka/Ks less than one. Conversely, Ka/Ks greater than one is a strong indicator of positive selection. Although we have seen a high ratio in the case of PUP2, the indication of such a number from an uncharacterized putative transcript remains to be elucidated. At the present time, limited by the amount of experimental data, the ratio of Ka/Ks in the SARS-CoV data is higher than that of most genes in HIV and influenza viruses, in which selection and adaptation are both playing significant roles in changing the rate of non-synonymous substitutions, especially for the structural proteins that are the targets of the host immune systems.

Instead, we have inspected these substitutions individually to the nucleotide positions in a context of a codon. It is not surprising to find that the three-nucleotide positions within a codon have a similar frequency to be mutated when insufficient evolutionary processes are yet to be experienced by the newly emerged virus. Among all the 139 substitutions (including 2 counted in 2 ORFs), 45 are at the first nucleotide position, 47 are at the second one, and 47 are at the third one. It implies that the selective pressure has yet to work on these mutations that were most likely generated in the initial viral population due to replication errors of the viral machinery. Many interesting non-synonymous substitutions were found among the group. A common A/C transversion (Locus 26,031, C in both BJ01 and BJ03, A in all other isolates, including two of the BJ Group, BJ02 and

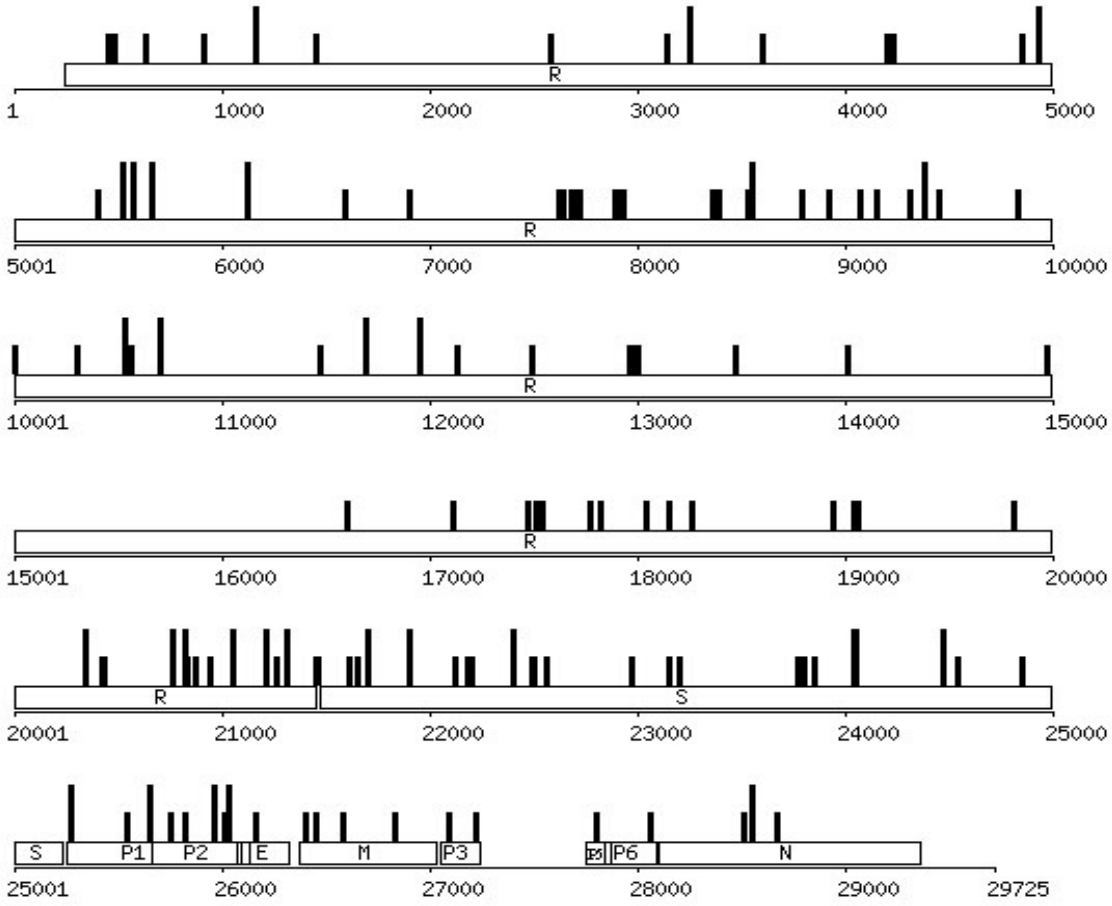

Fig. 2 Distribution of the non-synonymous substitutions in the SARS-CoV genome from the known cases. The tall vertical bars represent the substitutions detected from the BJ Group and the low bars denote those from non-BJ Groups. The scale marks the nucleotide positions in reference to BJ01.

Table 4 Summarized Substitutions Identified in the BJ Group and other 13 Isolates of SARS-CoV

ORF	Size(nt)	No.of S*		Percentage of substitutions (%)		No. of N-Syn*		Percentage of N-Syn (%)	
		BJ Group	All	BJ Group	All	BJ Group	All	BJ Group	All
R	21,222	25	92	0.12	0.43	21	65	84	71
S	3,768	9	22	0.24	0.58	6	13	67	59
BGI-PUP1	825	4	9	0.48	1.09	2	6	50	67
BGI-PUP2	465	2	5	0.43	1.08	2	4	100	80
E	231	0	1	0	0.43	0	1		100
M	666	0	4	0	0.60	0	4		100
BGI-PUP3	192	0	2	0	1.04	0	2		100
BGI-PUP5	120	0	1	0	0.83	0	1		100
BGI-PUP6	255	0	1	0	0.39	0	1		100
N	1,269	1	4	0.08	0.32	1	3	100	75
Non-ORF		1	1						
Total	29,725	42†	142†	0.13	0.46	32	100	76	70

*S and N-Sys stand for synonymous and non-synonymous substitutions, respectively. †A single substitution at the same position in a region overlapping with two ORFs was counted as 2. The total number is 137 when such a substitution event was counted only once so the total number of substitutions contributed by the BJ Group is reduced to 40.

Table 5 Comparison of the Mutation Rates in SARS-CoV, Influenza Virus, and HIV*

Virus	ORF	Size (nt)	No. of substitutions	Substitute rate (%)	No. of non-synonymous substitution	Non-synonymous substitute rate (%)	dN/dS	Ka	Ks	Ka/Ks
	R	21,222	92	0.43	65	0.31	2.38	0.075	0.111	0.67
	S	3,768	22	0.58	13	0.35	2.00	0.108	0.188	0.57
	PUP1	825	9	1.09	6	0.73	1.77	0.171	0.340	0.50
	PUP2	465	5	1.08	4	0.86	4.88	0.217	0.152	1.43
	E	231	1	0.43	1	0.43		0.093	0.000	
SARS-CoV	M	666	4	0.60	4	0.60		0.126	0.000	
	PUP3	192	2	1.04	2	1.04		0.492	0.000	
	PUP4	369	0	0.00	0	0.00		0.000	0.000	
	N	1,269	4	0.32	3	0.24	3.00	0.049	0.054	0.91
	PUP5	297	0	0.00	0	0.00		0.000	0.000	
	Total	29,725	137(142)	0.46	98	0.33	2.52	0.085	0.119	0.72
	HA	1,701	698	41.03	323	18.99	0.62	64.5	382.8	0.17
	M2	294	98	33.33	60	20.41	0.83	178.0	815.9	0.22
	M1	759	243	32.02	79	10.41	0.24	79.9	1,112	0.07
	NA	1,413	418	29.58	181	12.81	0.67	12.3	69.4	0.18
Influenza	NP	1,497	573	38.28	186	12.42	0.31	107.0	1,231	0.09
Virus A	NS	366	128	34.97	56	15.30	0.34	90.6	1,087	0.08
	PA	2,151	689	32.03	181	8.41	0.20	31.4	596.4	0.05
	PB1	2,274	851	37.42	204	8.97	0.13	27.5	805.7	0.03
	PB2	2,280	805	35.31	203	8.90	0.18	36.2	719.3	0.05
	Total	13,638	4,784(4,803)	35.08	1,473	10.80	0.27	48.1	659.1	0.07
	Gag	1,476	979	66.33	676	45.80	0.88	5,917	25,290	0.23
	Pol	3,000	1,894	63.13	1,232	41.07	0.61	3,974	24,510	0.16
	Vif	579	392	67.70	300	51.81	1.42	7,613	21,014	0.36
	Vpr	291	202	69.42	140	48.11	0.76	5,924	28,054	0.21
HIV-1	Tat	306	216	70.59	172	56.21	2.64	11,101	16,108	0.69
	Rev	303	237	78.22	189	62.38	1.70	10,107	19,080	0.53
	Vpu	249	215	86.35	182	73.09	2.23	14,947	25,620	0.58
	Env	2,574	1,992	77.39	1,531	59.48	1.70	10,765	23,342	0.46
	Nef	624	473	75.80	370	59.29	1.12	8,632	29,346	0.29
	Total	9,680	6,536(7,141)	67.52	4,792	49.50	1.15	7,445	24,094	0.31

*SARS-CoV data are from 17 isolates and referred to the notes of Table 2. Influenza Virus data are from 50 strains of HA segment, 100 strains of M segment, 24 strains of NA segment, 88 strains of NP segment, 89 strains of NS segment, 53 strains of PA segment, 58 strains of PB1 segment, and 58 strains of PB2 segment, downloaded from http://www.flu.lanl.gov. HIV data are from 405 strains, downloaded from http://hiv-web.lanl.gov/. Only those minor alleles that are present in at least two sequences were considered as real substitutions.

BJ04), leading to a Gln (C<u>A</u>A)-to-Pro (C<u>C</u>A) change, was noticed in the BJ Group specific variations. Since it resides in the PUP2 coding sequence, the functional significance of such a mutation is yet to be revealed experimentally in the future. A G/A transition (Locus 25,280, A in BJ03 and BJ02, G in all others) leading to Gly/Pro (from polar to non-polar) was also identified. These changes are not as drastic in terms of biochemical characteristics in the amino acid composition as the one found within the BJ01/BJ03 patient, indicating there might be possible selection on the latter cases. BJ04 has all the alleles at the loci mentioned above, the same as all the other isolates, assuring its relationship with isolates outside the BJ

Group. GD01 shares the same alleles at three loci with either BJ01 or BJ02, which are different from all the others, suggesting that these mutations are early replication errors before the viral invasion into Beijing, providing a link between the BJ Group and the GD Group in Guangdong where the first major epidemics of SARS occurred.

The sequence variation between BJ01 and BJ03 was expected to reflect the replication error rates, emerging from both replication cycles and in different tissues inside a single host. The sequence variation between BJ01/03 and BJ02 would reflect the replication errors between the first and the second round of the infection among the hosts, as well as the selective pressure from the host or possible advantages taken by the viruses. Somewhat to our surprises, we have noticed a "square-root rule" of the non-synonymous substitutions among the member isolates in the BJ Group (Figure 3). There are approximately 15 to 16

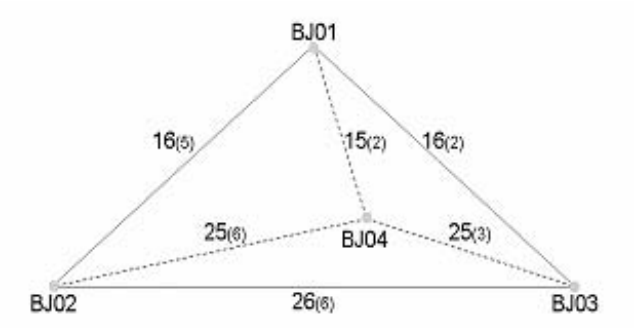

Fig. 3 Two-scale substitution rates of the SARS-CoV in the first and second round of the transmission of the BJ Group. The numbers above the lines that connect each isolate are nonsynonymous substitution counts believed as results of the first round and second round of the transmission. The numbers in the parentheses are synonymous substitution counts between the connected isolates.

non-synonymous substitutions in the first round of the transmission (15 between BJ01 and BJ03, 16 between BJ01 and BJ02, 16 between BJ01 and BJ04); it approximately equals to 4^2 or 2^4. In the second round of the infection, there are 25 to 26 non-synonymous substitutions (25 between BJ02 and BJ04, 25 between BJ02 and BJ03, 25 between BJ03 and BJ04); it is close to 5^2, or to 2^5 when both non-synonymous and synonymous substitutions are accounted. Such a "square-root rule" implies that the mutations occur freely without any constraints from patients to patients, perhaps due to lack of section pressure and adaptation during early transmissions or there is not enough time for them to become obvious, even though there might be a slight reduce in numbers in the second round of the infection.

We summarized the non-synonymous substitutions according to their subregions defined by a combination of structural and/or functional properties in the corresponding ORFs (Table 6). Data were classified according to the computationally predictable changes of physiochemical features and/or the secondary structure they would make in the corresponding subregions. For example, the predicted alterations by the substitutions of the M protein would lead to an increased pI (isoelectric point) of the N-terminal exterior region, decreased or increased hydrophobicity in the TM (transmembrane) domains, and decreased hydrophilicity in the C-terminal interior region, respectively. These predictable changes should suggest that the virus might benefit from these substitutions with remarkable changes that may be advantageous for the virus to defend the host immune system or drift to a new status ready for coming back. No single base insertion or deletion has been found so far in all the sequences published to this date; it states clearly the fidelity of the viral replication machinery.

The BJ Group as a Subset of SARS-CoV Isolates Represents a Discrete Viral Transmission Path

Rooted phylogenetic trees (Figures 4 and 5), proposed on the basis of the haplotypes from each group, synonymous or non-synonymous substitutions, and the sum of all substitutions of the 17 genome sequences of SARS-CoV isolates from patients identified in Canada, USA, Singapore, and China (Beijing, Zhejiang, Guangdong, Hong Kong, and Taiwan), gave different paradigms but positioned the BJ Group, together with the newly discovered GD01 (GD-*Ins29*) in the same clade, followed by the H-U Group, then the H-T Group, leaving the SP Group (Singapore) more distant. This paradigm suggests a possible transmission path from Guangdong to Beijing/Hong Kong, then to other countries and regions. It appears consistent with the epidemiological data presently available, and would suggest a possible transmission path among Guangdong, Hong Kong, Beijing, and USA.

Table 6 Predicted Subregional Changes by the Non-Synonymous Substitutions in the 17 SARS-CoV Genomes

ORF	Subregion (Position in ORF)	Position (nt)	No. of Substitution	Position nt (a.a.)	Type of Substitution	Size (a.a.)	pI value	Hydrophobicity (%)	Hydrophilicity (%)	Charge (+)(%)	Charge (−)(%)
R LP	(1-179)	246-783	4	454(70)	Cys(1)/Phe(16)	179	5.4-5.5	29.1-29.6	46.4-47.5	14.0-14.5	(14.5)*
				483(80)	Asn(16)/Tyr(1)						
				489(82)	Cys(1)/Gly(16)						
				633(130)	Arg(1)/Gly(16)						
p65L	(180-818)	784-2,700	2	918(225)	Gln(1)/Lys(16)	639	6.0-6.1	(30.2)	(45.4)	12.7-12.8	(11.9)
				1,161(306)	Phe(1)/Val(16)						
	interBHID_PLP (999-1,631)	3,241-5,139	6	3,255(1,004)	Asn(16)/His(1)	633	5.9-6.0	29.1-29.2	47.2-47.4	13.1-13.3	(12.5)
				3,607(1,121)	Ile(16)/Thr(1)						
				4,201(1,319)	Arg(1)/Lys(16)						
NSP1				4,231(1,329)	Phe(16)/Ser(1)						
				4,957(1,538)	Ser(16)/Thr(1)						
				4,933(1,563)	Lys(1)/Met(16)						
	PLP (1,632-1,845)	5,140-5,781	4	5,529(1,762)	Ile(16)/Leu(1)	214	(7.5)	26.6-27.1	43.5-43.9	(11.7)	(8.4)
				5,572(1,776)	Gln(16)/Pro(1)						
				5,575(1,777)	Gln(16)/Pro(1)						
				5,662(1,806)	Gly(16)/Val(1)						
	BHOD (1,846-2,138)	5,782-6,660	2	6,129(1,962)	Ile(1)/Leu(16)	293	(5.7)	(27.0)	(50.2)	(11.9)	(11.9)
				6,593(2,116)	Leu(16)/Phe(1)						
	interBHOD_HOD1 (2,139-2,978)	6,661-9,180	11	6,910(2,222)	Cys(16)/Tyr(1)	840	8.1-8.3	33.0-33.2	42.7-43.0	10.8-11.1	8.1-8.2
				7,624(2,460)	Pro(16)/His(1)						
				7,683(2,480)	Ile(1)/Leu(16)						
				7,900(2,552)	Ala(16)/Val(1)						
				7,911(2,556)	Asn(1)/Asp(16)						
				7,935(2,564)	Pro(1)/Ser(16)						
				8,368(2,708)	Ser(16)/Thr(1)						
				8,398(2,718)	Arg(16)/Thr(1)						
				8,553(2,770)	Leu(2)/Val(15)						
				9,076(2,944)	Ile(1)/Thr(16)						
				9,157(2,971)	Ala(1)/Val(16)						
	HOD1 (2,979-3,240)	9,181-9,966	5	9,312(3,023)	Pro(1)/Ser(16)						
				9,313(3,023)	Leu(1)/Ser(16)						
				9,385(3,047)	Ala(5)/Val(12)						
				9,460(3,072)	Ala(2)/Val(15)						
				9,835(3,197)	Ala(13)/Val(4)						
3CLP	(NSP2) (3,241-3,546)	9,967-10,885	2	10,531(3,429)	Gln(16)/Pro(1)	306	(6.30)	(31.0)	42.8-43.1	(10.1)	(8.5)
				10,709(3,488)	Asp(16)/Glu(1)						
HOD2	(NSP3) (3,547-3,836)	10,886-11,754	2	11,472(3,743)	Asn(1)/Tyr(16)	290	(8.90)	(43.8)	30.7-31.0	(9.0)	(3.8)
				11,698(3,818)	Asn(16)/Thr(1)						
NSP4	(3,837-3,919)	11,755-12,003	2	11,952(3,903)	Asn(1)/Asp(16)	83	5.2-5.7	(38.6)	(50.6)	(12.0)	10.8-12.0
				11,955(3,904)	Ile(16)/Phe(1)						
NSP5	(3,920-4,117)	12,004-12,597	2	12,498(4,085)	Ile(16)/Leu(1)	198	(6.30)	27.8-28.3	48.0-48.5	(11.1)	(11.1)
				12,499(4,085)	Asn(1)/Ile(16)						
NSP7	(4,231-4,369)	12,939-13,353	4	12,963(4,240)	Asn(16)/His(1)	139	5.9-6.4	(22.3)	(40.3)	8.6-10.1	(7.9)
				12,975(4,244)	Ile(1)/Leu(16)						
				13,006(4,254)	Ala(16)/Gly(1)						
				13,008(4,255)	Gln(1)/Lys(16)						
RdRp	(NSP9) (4,370-5,301)	13,354-16,149	4	13,475(4,411)	Ile(1)/Val(16)	932	6.0-6.1	30.3-30.5	44.4-44.6	(12.9)	11.5-11.6
				13,476(4,411)	Gly(1)/Val(16)						
				14,019(4,592)	Glu(1)/Val(16)						
				14,973(4,910)	Gln(16)/Leu(1)						
HEL	(NSP10) (5,302-5,902)	16,150-17,952	3	17,112(5,623)	Leu(16)/Ser(1)	601	8.5-8.6	27.6-27.8	44.1-44.4	12.5-12.6	(8.7)
				17,474(5,744)	Arg(16)/Cys(1)						
				17,545(5,767)	Asp(11)/Glu(6)						

Table 6 ***(Continued)***

ORF	Subregion (Position in ORF)	Position (nt)	No. of Substitution	Position nt (a.a.)	Type of Substitution	Size (a.a.)	pI value	Hydrophobicity (%)	Hydrophilicity (%)	Charge (+)(%)	Charge (−)(%)
	NSP11 (5,903-6,429)	17,953-19,533	2	18,263(6,007) 19,065(6,274)	Ile(1)/Leu(16) Ile(4)/Thr(13)	527	(7.5)	30.4-30.6	43.6-43.8	(13.9)	(9.7)
	NSP12 (6,430-6,775)	19,534-20,571	2	20,344(6,700) 20,429(6,729)	Ile(1)/Met(16) Asn(1)/Asp(16)	346	(5.1)	(32.9)	(46.2)	(11.8)	(13.3)
	NSP13 (6,776-7,073)	20,572-21,465	7	20,762(6,840) 20,829(6,862) 20,875(6,877) 21,053(6,937) 21,200(6,992) 21,268(7,008) 21,314(7,024)	Leu(1)/Met(16) Gln(16)/Pro(1) Asp(16)/Glu(1) Ala(16)/Pro(1) Asp(1)/Glu(16) Asn(16)/Lys(1) Gln(1)/Lys(16)	298	7.3-7.5	(32.9)	44.4-44.6	11.1-11.4	(9.1)
S	Region 3 (60-115)	21,650-21,815	1	21,702(77)	Asp(5)/Gly(12)	56	7.9-9.0	(33.9)	44.6-46.4	(10.7)	3.6-5.4
	Region 4 (115-145)	21,815-21,905	1	21,902(144)	Leu(1)/Met(16)	31	(5.8)	(38.7)	(38.7)	(6.5)	(6.5)
	Region 7 (220-255)	22,130-22,235	2	22,188(239) 22,203(244)	Leu(1)/Ser(16) Ile(11)/Thu(6)	36	(8.0)	33.3-36.1	30.6-33.3	(5.6)	(2.8)
	Region 8 (255-320)	22,235-22,430	1	22,403(311)	Arg(2)/Gly(15)	66	5.1-6.0	(28.8)	48.5-50.0	10.6-12.1	(12.1)
	Region 14 (560-580)	23,150-23,210	1	23,201(577)	Ala(1)/Ser(16)	21	(4.3)	(23.8)	47.6-52.4	(9.5)	(19.0)
	Region 20 (830-880)	23,960-24,110	2	24,050(860) 24,053(861)	Leu(1)/Val(16) Arg(1)/Ser(16)	51	4.1-4.6	(33.3)	(27.5)	2.0-3.9	(5.9)
	Region 24 (975-1,010)	24,395-24,500	1	24,474(1,001)	Arg(16)/Met(1)	36	9.7-10.8	27.8-30.6	47.2-50.0	11.1-13.9	(5.6)
BGI-PUP1	(1-274)	25,249-26,073	6	25,279(11) 25,280(11) 25,550(101) 25,654(136) 25,760(171) 25,825(193)	Arg(1)/Gly(16) Glu(2)/Gly(15) Lys(1)/Met(16) Gln(2)/Lys(15) Ala(1)/Glu(16) Arg(16)/Trp(1)	274	5.4-5.8	34.7-35.0	39.1-39.4	8.0-8.8	7.7-8.4
BGI-PUP2	(1-154)	25,670-26,134	4	25,760(31) 25,965(99) 26,013(115) 26,031(121)	Gln(1)/Lys(16) Ile(1)/Thr(16) Gln(1)/Leu(16) Gln(15)/Pro(2)	154	9.9-10.9	37.0-40.1	51.9-52.3	18.8	0.60-0.62
E	(1-76)	26,098-26,328	1	26,167(24)	Met(1)/Val(16)	76	(6.0)	(47.4)	(32.9)	(5.3)	(5.3)
	Exterior (1-18)	26,379-26,433	1	26,409(11)	Glu(16)/Lys(1)	18	4.2-4.8	(33.3)	(55.6)	5.6-11.1	16.7-22.2
M	TM I (19-37)	26,434-26,487	1	26,458(27)	Cys(1)/Phe(16)	18	(5.6)	68.4-73.7	(10.5)	(0.0)	(0.0)
	TM II(50-72)	26,529-26,585	1	26,581(68)	Ala(16)/Val(1)	22	(7.9)	63.6-68.2	(9.1)	(4.5)	(0.0)
	Interior (99-221)	26,676-27,044	1	26,838(154)	Pro(1)/Ser(16)	122	(9.8)	(30.9)	44.7-45.5	(15.4)	(7.3)
BGI-PUP3	(1-63)	27,055-27,246	2	27,092(13) 27,224(57)	Glu(16)/Gly(1) Leu(5)/Pro(12)	63	4.7-4.8	46.0-47.6	41.3-42.9	(11.1)	14.3-15.9
N	Region 7 (137-178)	28,509-28,634	2	28,519(140) 28,560(154)	Leu(16)/Trp(1) Asn(16)/Tyr(1)	42	7.9-8.0	14.3-16.7	47.6-52.4	(16.7)	(5.6)
	Region 8 (179-223)	28,635-28,769	1	28,677(193)	Cys(1)/Gly(16)	45	12.0-12.1	(8.9)	(57.8)	(13.3)	(2.2)

*Numbers in the parentheses for the last five columns indicate the percentage of the parameters that are not affected by the amino acid changes.

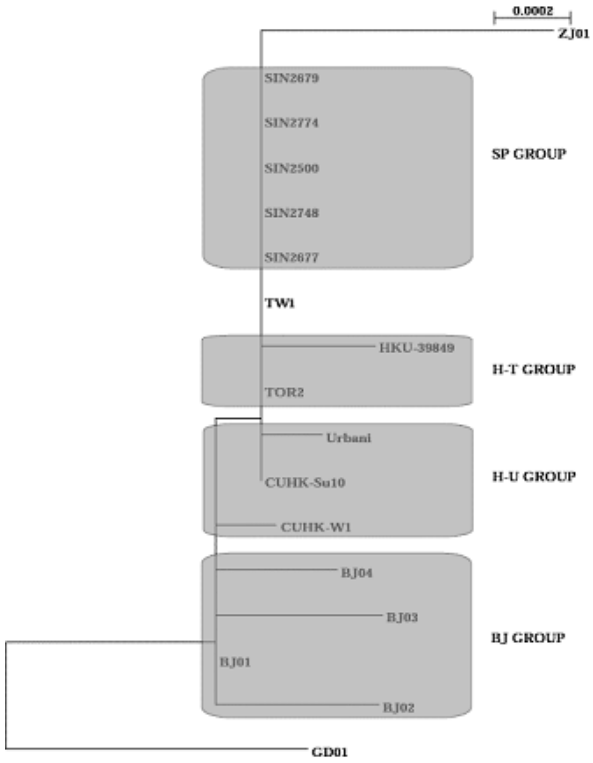

Fig. 4 A rooted phylogenetic tree (GD01 as the postulated root) indicates the defined haplotypes and possible transmission path of the SARS-CoV based on the complete genome sequences of 17 SARS-CoV isolates. The neighbor-joining trees are generated by using the program Clustalw 1.81. The sources and abbreviations of the sequences are referred to Table 2.

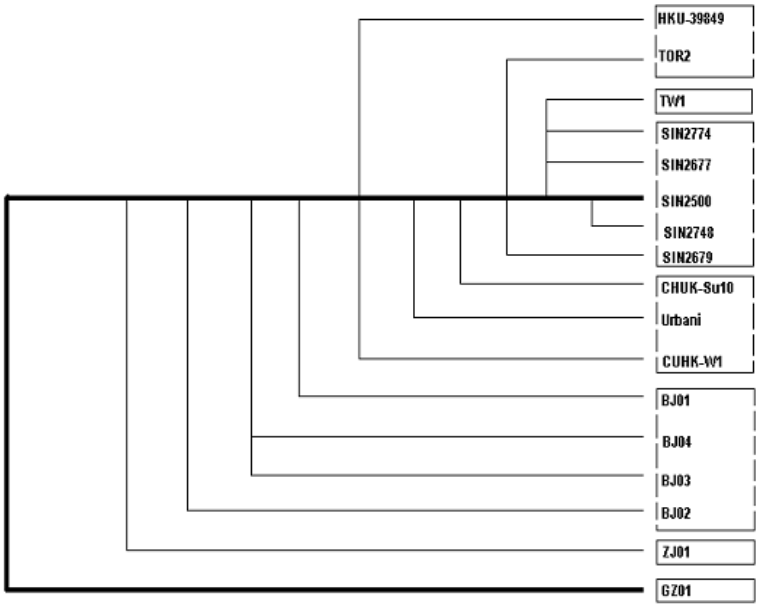

Fig. 5A (all substitutions)

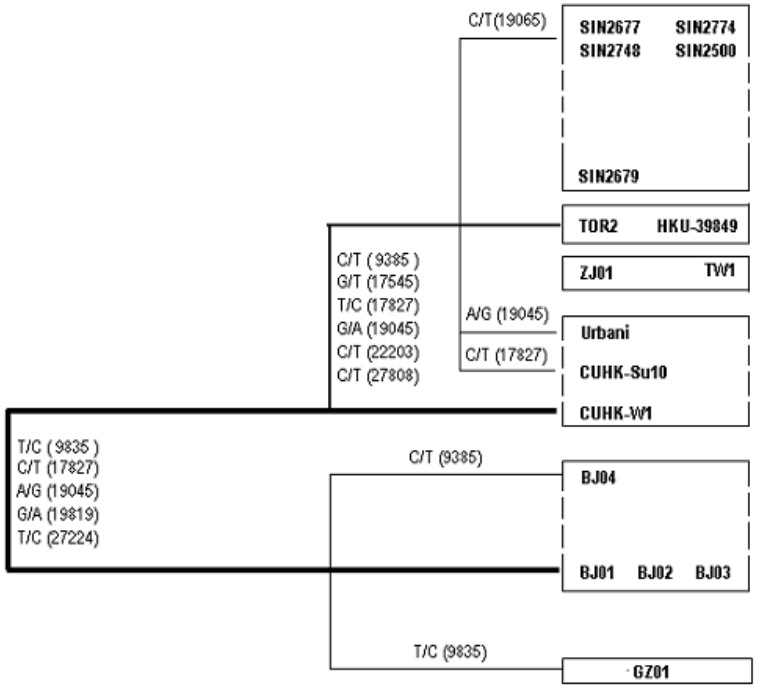

Fig. 5B (haplotypes)

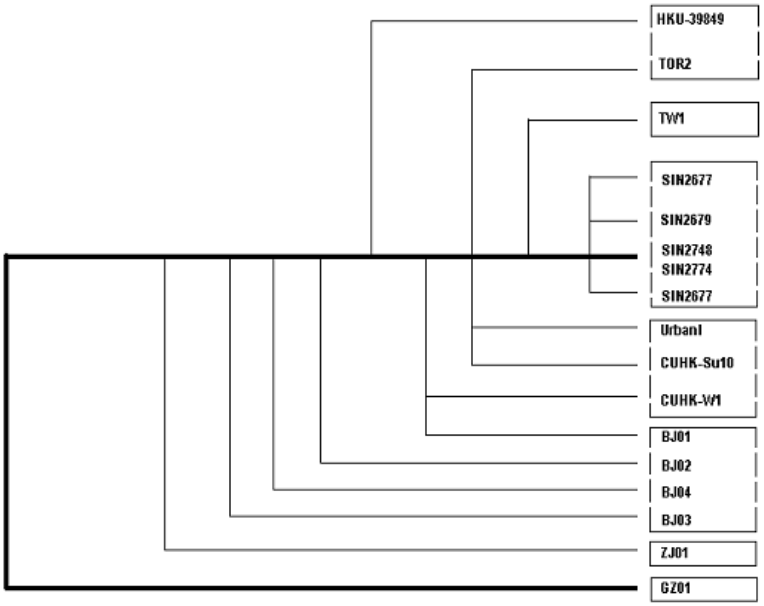

Fig. 5C (non-synonymous)

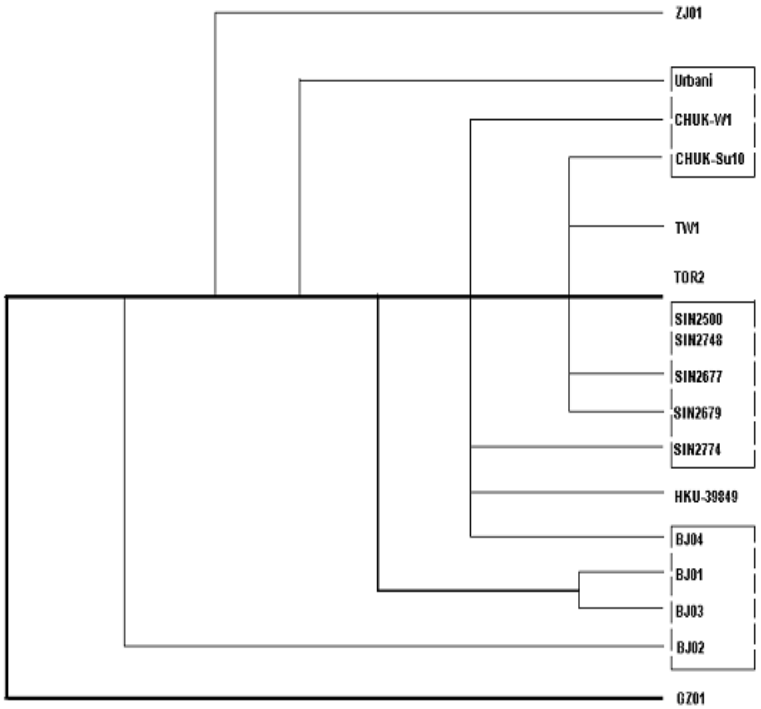

Fig. 5D (synonymous)

Fig. 5 Proposed rooted phylogenetic trees of the 17 isolates of the SARS-CoV based on all substitutions (A), haplotypes (B), non-synonymous (C), and synonymous (D) substitutions.

It is obvious that we are just in the early process of exploiting the information from the genomes of SARS-CoV isolates from patients of different countries and regions. The final picture of the infection route and the mutation spectra will be revealed in due time as long as we keep sequencing the many clinical isolates of the virus accurately and consistently. We have been doing so since the epidemic started, finding the unique insertion variant in the samples from Guangdong and now the haplotypes, and we will keep doing so until the next round of the infection if it does come in this fall.

Acknowledgements

We thank Ministry of Science and Technology of China, Chinese Academy of Sciences, and National Natural Science Foundation of China for financial support. We are indebted to collaborators and clinicians from Peking Union Medical College Hospital, National Center of Disease Control of China, and the Municipal Governments of Beijing and Hangzhou.

References

1. Lee, N., *et al.* 2003. A major outbreak of severe acute respiratory syndrome in Hong Kong. *N. Engl. J. Med.* 348: 1986-1994.
2. Rota, P.A., *et al.* 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394-1399.
3. Marra, M.A., *et al.* 2003. The genome sequence of the SARS-associated coronavirus. *Science* 300: 1394-1399.
4. 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.
5. Qin, E.D., *et al.* 2003. A genome sequence of novel SARS-CoV isolates: the genotype, GD-*Ins*29, leads to a hypothesis of viral transmission in South China. *Geno., Prot. & Bioinfo.* 1: 101-107.
6. Williamson, S. 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol. Biol. Evol.* 20: 1318-1325.
7. Hay, A.J., *et al.* 2001. The evolution of human influenza viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 356:1861-1870.
8. Graur, D. and Li, W.H. (ed.) 2000. *Fundamentals of molecular evolution.* Sinauer Press, Sunderland, USA.