Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: Molecular epidemiology and genome evolution

Shiou-Hwei Yeh*†, Hurng-Yi Wang‡, Ching-Yi Tsai†, Chuan-Liang Kao§, Jyh-Yuan Yang¶, Hwan-Wun Liu‖, Ih-Jen Su¶, Shih-Feng Tsai*, Ding-Shinn Chen†**††‡‡, Pei-Jer Chen†**††‡‡, and the National Taiwan University SARS Research Team§§

*Division of Molecular and Genomic Medicine, National Health Research Institutes, Taipei 115, Taiwan; †Hepatitis Research Center and **Department of Internal Medicine, National Taiwan University Hospital, Taipei 100, Taiwan; ‡Institute of Molecular Biology, Academia Sinica, Taipei 115, Taiwan; §Department of Medical Technology, National Taiwan University, Taipei 100, Taiwan; ¶Center for Disease Control, Taipei 115, Taiwan; ‖Institute of Preventive Medicine, National Defense Medical College, Taipei 237, Taiwan; and ††Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei 100, Taiwan

Communicated by Robert H. Purcell, Bethesda, MD, November 27, 2003 (received for review June 28, 2003)

Since early March 2003, the severe acute respiratory syndrome (SARS) coronavirus (CoV) infection has claimed 346 cases and 37 deaths in Taiwan. The epidemic occurred in two stages. The first stage caused limited familial or hospital infections and lasted from early March to mid-April. All cases had clear contact histories, primarily from Guangdong or Hong Kong. The second stage resulted in a large outbreak in a municipal hospital, and quickly spread to northern and southern Taiwan from late April to mid-June. During this stage, there were some sporadic cases with untraceable contact histories. To investigate the origin and transmission route of SARS-CoV in Taiwan's epidemic, we conducted a systematic viral lineage study by sequencing the entire viral genome from ten SARS patients. SARS-CoV viruses isolated from Taiwan were found closely related to those from Guangdong and Hong Kong. In addition, all cases from the second stage belonged to the same lineage after the municipal hospital outbreak, including the patients without an apparent contact history. Analyses of these full-length sequences showed a positive selection occurring during SARS-CoV virus evolution. The mismatch distribution indicated that SARS viral genomes did not reach equilibrium and suggested a recent introduction of the viruses into human populations. The estimated genome mutation rate was ≈0.1 per genome, demonstrating possibly one of the lowest rates among known RNA viruses.

The recently encountered severe acute respiratory syndrome (SARS) initially emerged in southern China in late 2002 and quickly spread worldwide after March 2003 (1–3). Globally, 8,098 people were infected and 774 people died in this SARS outbreak, with a mortality rate near 10%. (World Health Organization, www.who.int/csr/sars/country/table 2003_09_23/en/) (4). The SARS causative pathogen was first cultured in Vero E6 cells and found by electron microscopy to resemble a coronavirus (5–7). In a very short time, the whole genome of this virus has been completely sequenced, revealing it to be a new member of the *Coronaviridae* family, designated SARS-CoV (8, 9). The new virus bears a distinct phylogenetic pattern but similar genome organization when compared with three other groups of known coronaviruses, all containing a large, positive-sense RNA genome with a size around 30 kb (8–10).

To better understand the origin and the route of SARS-CoV transmissions, the molecular epidemiological approach, aided by viral sequencing analysis, has been conducted in several areas, including Hong Kong, Canada, Singapore, Vietnam, Germany, and China (11, 12). Sequence comparisons that support patient contact histories and help track infection routes have identified several viral genetic signatures useful in tracing the origins of the SARS virus in these areas (11, 12). The feasibility of this phylogenetic approach has been confirmed by inferring the following history: wide divergence among Hong Kong and Guangdong isolates suggested the earliest event in these areas. Subsequently, there were two routes of viral spreading, one to Beijing (Beijing cluster) and the other to the rest of the world, including Canada, Singapore, and Vietnam (Vietnam cluster). The latter route of transmission was mainly through an index case in each area with a contact history at Hotel M in Hong Kong (11).

In Taiwan, the SARS outbreak started from early March 2003 and resulted in 346 probable cases and 37 deaths by mid-June (World Health Organization, www.who.int/csr/sars/country/table 2003_09_23/en/). This epidemic can be divided into two stages (Fig. 1). In the first stage (stage I, from early March to mid-April), all SARS patients had a definite contact history either with travel to the affected areas or with an intrafamily or intrahospital exposure to SARS patients. The increase of probable cases was low (fewer than three cases a day), and the local transmission was limited in this stage. The contact history of patients did not show linkage with Hotel M, and the origin of the SARS infection remained to be determined.

A larger outbreak in a Taipei City Municipal Hospital H in late April marked the start of the second stage of SARS infection (stage II), which was far more serious than stage I (Fig. 1). The SARS patients or the contact persons spread the virus to other areas or hospitals in Taiwan until mid-June, resulting in 325 probable cases and 36 deaths. The origin of this large outbreak was undetermined by traditional epidemiological investigations. More importantly, this stage contained several sporadic cases from the community without any traceable contact or exposure histories.

Because identifying the origin of each affected individual is currently the prerequisite for an effective control of SARS-CoV spreading, we thus decided to conduct a systematic molecular epidemiological study in Taiwanese patients to trace the viral lineages. Because current sequence data did not identify any viral segments of SARS-CoV genomes as containing a hypervariable region, we decided to conduct whole-genome sequencing to obtain adequate genomic information for analysis. In addition, such se-

Abbreviations: SARS, severe acute respiratory syndrome; CoV, coronavirus.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY291451 and AY502923–AY502932).

‡‡To whom correspondence should be addressed at: Hepatitis Research Center, National Taiwan University Hospital, 7 Chung-Shan South Road, Taipei 100, Taiwan. E-mail: dschen@ha.mc.ntu.edu.tw or peijer@ha.mc.ntu.edu.tw.

§§The National Taiwan University SARS Research Team: Ding-Shinn Chen[a,b], Yuan-Teh Lee[c,d], Che-Ming Teng[e], Pan-Chyr Yang[b], Hong-Nerng Ho[f], Pei-Jer Chen[g], Ming-Fu Chang[h], Jin-Town Wang[i], Shan-Chwen Chang[b], Chuan-Liang Kao[i], Wei-Kung Wang[i], Cheng-Hsiang Hsiao[k], and Po-Ren Hsueh[l].

Offices of [a]the Dean and [e]Research and Development, and Departments of [b]Internal Medicine, [h]Biochemistry and Molecular Biology, [i]Microbiology, [i]Medical Technology, [k]Pathology, [l]Laboratory Medicine, and [g]Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei 100, Taiwan; and [c]Office of the Superintendent and Departments of [d]Internal Medicine and [f]Medical Research, National Taiwan University Hospital, Taipei 100, Taiwan.

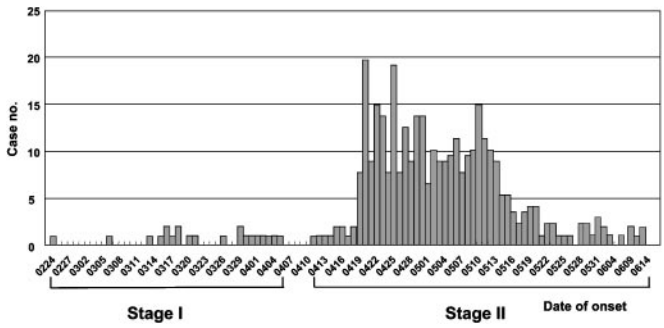

Fig. 1. Case number of SARS patients in Taiwan in an outbreak that lasted from March to June 2003. There were two stages of SARS infection in the epidemic. The second stage was much more serious.

quencing analysis in a series of endemic cases can help estimate the rate of viral genetic evolution and will possibly help reveal the host selection process on specific genes of the virus.

Finally, most current SARS-CoV genome sequencing was conducted on viral isolates cultured in Vero E6 cells instead of viruses directly isolated from clinical samples. Whether the virus isolates propagated in cell culture represent the major species in SARS patients remains to be clarified. To solve this problem, we also conducted sequencing analysis directly on the viruses in the primary specimens from SARS patients. Results of the present study may help clarify the transmission and the genomic evolution of SARS-CoV in the recent SARS epidemic on Taiwan.

Materials and Methods

Study Subjects. In total, we included 10 Taiwanese patients with SARS-CoV infection. All of them met the World Health Organization definitions as probable SARS cases, showing typical clinical symptoms and were confirmed by PCR with SARS-specific primers (5). The patients were from both stages of the epidemic, 4 from stage I and the remaining 6 from stage II. Patients 1 and 2 were infected in mid-March 2003 through familial or hospital contact with the first index SARS patient, who developed SARS around March 7 after returning to Taiwan from Guangdong. Patient 3 was an employee of an international construction company who developed symptoms after returning from Beijing (through Hong Kong) in late March. Patient 4 was the first fatal SARS case in Taiwan and was infected in early April by his visiting brother, who lived in Amoy Garden complex in Hong Kong.

The other 6 cases came from stage II of the epidemic. Patients 5–7 were from Hospital H where the SARS outbreak occurred in late April. Patient 8 was from a local clinic R and seemed to be infected in early May. Patients 9 and 10 were sporadic cases without apparent contacts and were reported from the Taipei metropolitan areas in mid-May. The number of all SARS patients was arranged in their chronological sequence of the disease onset (Table 1).

Both the clinical specimens and the virus isolate after passage in Vero E6 cells were collected from patients 1 and 2. For patients 3 to 7, we obtained viral isolates from culture supernatants only. Patients 8–10 provided clinical samples (throat swabs) only.

Viral Culture for SARS-CoV. Throat swab specimens were inoculated into Vero E6 cells, cultured, and monitored as described (13). Once the virus-induced cytopathic effects appeared, the culture cell supernatant was harvested and submitted to RNA extraction. All experiments involving viral culture and isolation were conducted in biosafety level 3 laboratories.

Extraction of SARS-CoV Genomic RNA, Reverse Transcription of SARS RNA, and PCR Amplification of SARS cDNA Fragments. The viral RNA was extracted with the High Pure Viral Nucleic Acid Kit (Roche Diagnostics Applied Science, Mannheim Germany), either from culture supernatant or from primary nasopharyngeal specimens as described (13).

We used the SuperScript cDNA system (Invitrogen) to reverse transcribe the RNA template into cDNA, which is used for subsequent PCR amplification. To sequence the whole viral genome, we designed 25 primer sets based on the cDNA sequence data from the TOR2 SARS isolate (accession no. NC_004718) (8). The sequence of the primers and the detailed PCR conditions have been described (13).

Direct Sequencing Analyses. The PCR products were used for direct sequencing analysis on ABI3730 sequencers (Applied Biosystems) with primers inward from both ends of the PCR fragments, and then analyzed with an ABI 3730 Genetics Analyzer. We used the SEQUENCHER package version 4.1.4 (Applied Biosystems) for processing all of the raw sequence data for base calling, assembly, and editing. Any nucleotide differences in the assembled genome sequences when compared with the first virus strain TOR2 (NC_004718) were all double-checked and confirmed. Sequences were deposited in the GenBank database.

Phylogenetic Construction and Data Analyses. Nucleotide sequences were aligned by using the default parameter of CLUSTAL W (14). A neighbor-joining (15) tree with 1,000 bootstrap replicates based on the number of mutations was constructed by using MEGA (16) to estimate phylogenetic relationships among sequences. The numbers of nucleotide positions were based on the TOR2 isolate (NC_004718).

For coding regions, we calculated both the number of synonymous changes per synonymous site (Ks) and the number of nonsynonymous changes per nonsynonymous site (Ka) (17) by using sequences of a coronavirus isolated from a palm civet (AY304486) as the outgroup (18). Assuming synonymous mutations as neutral variations, Ks is a measure of the mutation rate and Ka/Ks is a measure of the rate of protein evolution after controlling for the mutation rate.

Tajima's D (19), Fu and Li's D (20), and Fay and Wu's H (21) tests were applied to evaluate the deviations of the mutation frequencies of SARS-CoV from the standard neutral model. Tajima's (19) test examines whether the average number of pairwise nucleotide differences between sequences (θ_π) is larger than expected from the observed number of polymorphic sites (θ_w). The expected difference (D) between θ_π and θ_w is roughly zero under the standard neutral model. A positive value of D indicates possible balancing selection or population subdivision. A negative value suggests recent directional selection, a population bottleneck, or a purifying selection on deleterious alleles (19). Fu and Li's (20) test is based on the principle of comparing the number of mutations on internal branches with those on external branches. Compared with a neutral model of evolution, directional selection would result in an excess of external mutations, and balancing selection would result in an excess of internal mutations. Fay and Wu's (21) test compares the difference (H) between θ_π (which is influenced most by variants at intermediate frequencies) and θ_H (which is influenced most by high-frequency phylogenetically derived variants). A negative value reflects a relative excess of high-frequency derived alleles, as expected immediately after a selective sweep. Fay and Wu's H test was conducted on the website crimp.lbl.gov/htest.html.

Genomic Mutation Rate of SARS Coronaviruses. To analyze the mutation rate per generation per SARS-CoV genome, the model of recent population expansion was estimated to fit the current genome sequences. With the plot of a mismatch distribution, Tau (τ), the date of the population growth in units of mutational time, can be estimated. The estimations of the above population parameter including different estimators of θ and τ were carried out with DNASP software (22).

Table 1. Genetic variations in the genome of 28 completely sequenced SARS-CoV isolates

Position	TOR2	HSR1	Sin2679	Sin2748	Sin2500	Sin2677	Sin2774	Frankfurt1	HKU-39849	CUHK-Su10	TW1 (P11) (C)	TW2 (P11) (P)	TW3 (P12) (C)	TW4 (P12) (P)	TW5 (P13) (C)	TW6 (P15) (C)	TW7 (P16) (C)	TW8 (P17) (C)	TW9 (P18) (P)	TW10 (P19) (P)	TW11 (P10) (P)	TWC (P14) (C)	Urbani	CUHK-W1	BJ01	BJ02	BJ03	BJ04	Viral gene	aa change
601	G	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	leader	**E-->K**
623	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	R	-	-	-	-	-	-	-	-	-	leader	**N-->N/S**
1006	A	-	-	-	-	-	-	-	-	-	-	-	-	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	p65-like	**S-->S/R**
1180	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	p65-like	**V-->F**
1476	A	-	-	-	-	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	p65-like	R-->R
1782	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	p65-like	C-->C
2557	G	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	p65-like	**A-->T**
2601	T	-	-	-	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	p65-like	V-->V
3165	A	-	-	-	-	-	-	-	-	G	G	G	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp1	S-->S
3274	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp1	**N-->H**
3852	T	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	C	-	-	-	-	-	-	-	-	-	-	-	nsp1	S-->S
3916	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	-	-	-	-	-	-	-	-	-	-	nsp1	**T-->A**
4952	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	nsp1	**M-->K**
5427	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp1	A-->A
5548	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp1	**I-->L**
5591	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp1	**Q-->P**
5594	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp1	**Q-->P**
5681	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	nsp1	**G-->V**
6148	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	nsp1	**L-->I**
6172	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	-	nsp1	**H-->H/Y**
6406	G	-	-	-	-	-	-	-	-	-	R	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp1	**V-->V/I**
7746	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	nsp1	P-->P
7919	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	nsp1	**A-->V**
7930	G	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp1	**D-->N**
8387	G	-	-	-	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp1	**S-->T**
8417	G	-	-	-	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp1	**R-->T**
8572	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	-	-	nsp1	**V-->L**
9404	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	nsp1	**V-->A**
9479	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	nsp1	**V-->A**
9854	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	nsp1	**A-->V**
10550	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	nsp2	**Q-->P**
10587	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	nsp2	T-->T
10728	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	nsp2	**D-->E**
11448	C	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp3	I-->I
11493	C	-	-	-	-	-	-	-	-	-	-	-	T	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	nsp3	Y-->Y
11717	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp3	**N-->T**
11971	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	nsp4	**D-->N**
11974	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	nsp4	**I-->F**
13347	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	nsp7	D-->D
13494	G	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp9	**V-->I**
13495	T	-	-	-	-	-	-	-	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp9	**V-->G**
14979	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M	-	-	-	-	-	-	-	-	-	-	nsp9	**P-->P/T**
15235	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	nsp9	**C-->F**
16325	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	-	-	-	-	-	-	nsp10	P-->P
16622	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	nsp10	A-->A
17564	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	G	G	-	-	-	nsp10	**D-->E**
17798	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	nsp10	V-->V
17846	C	-	-	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	nsp10	R-->R
18065	G	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp11	K-->K
18282	C	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp11	**L-->I**
18965	T	-	-	-	-	-	A	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp11	I-->I
19064	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	-	-	-	-	-	nsp11	E-->E
19084	C	-	-	T	T	T	T	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nsp11	**T-->I**
19426	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	W	-	-	-	-	-	-	-	nsp11	**H-->H/L**
19838	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	G	G	nsp12	V-->V
20363	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	nsp12	**M-->I**
20781	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp13	**M-->L**
20848	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp13	**Q-->P**
20858	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	nsp13	P-->P
21072	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp13	**A-->P**
21239	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp13	**E-->D**
21333	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	nsp13	**K-->Q**
21488	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	non-transl.	(-)
21638	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	spike	S-->S
21674	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	spike	P-->P
21721	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	A	A	spike	**G-->D**
21921	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	spike	**M-->L**
22222	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	C	spike	**I-->T**
22422	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	spike	**G-->R**
22517	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	G	-	spike	R-->R
23174	C	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	spike	S-->S
23220	G	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	spike	**A-->S**
23792	C	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	spike	V-->V
24069	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	spike	**V-->L**
24072	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	spike	**S-->R**
24493	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	spike	**R-->M**
24872	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	spike	L-->L
24933	C	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	spike	**L-->F**
25298	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	unknown	**R-->G**
25299	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	A	-	unknown	**G-->E**
25341	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	-	-	unknown	**P-->P/L**
25569	T	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	unknown	**M-->K**
25673	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	unknown	**K-->Q**
25984	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	unknown	**T-->I**
26050	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	C	unknown	**Q-->P**
26203	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	-	envelope	V-->V
26428	G	-	-	-	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	membrane	**E-->K**
26477	T	-	-	-	-	-	-	-	-	-	G	-	-	-	-	G	G	G	G	G	G	G	-	-	-	-	-	-	membrane	**F-->C**
26600	C	-	-	-	-	-	-	-	T	T	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	-	membrane	**A-->V**
26734	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	-	-	-	membrane	**P-->S**
26857	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	-	-	-	membrane	**S-->P**
27067	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	d	-	-	-	-	-	non-transl.	(-)
27068	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	d	-	-	-	-	-	non-transl.	**(-)**
27091	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	-	-	-	-	-	-	unknown	D-->D
27111	A	-	-	-	-	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	unknown	**E-->G**
27243	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	unknown	**P-->L**
27782	A	-	-	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27783	A	-	-	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27784	A	-	-	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27785	C	-	-	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27786	T	-	-	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27787	T	-	-	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27807	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	d	-	-	-	-	-	non-transl.	(-)
27808	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	d	-	-	-	-	-	non-transl.	(-)
27810	C	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27811	T	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27812	C	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	T	-	-	-	-	non-transl.	(-)
27813	T	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27814	A	-	-	d	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	non-transl.	(-)
27827	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	C	C	C	non-transl.	(-)
28268	C	-	-	-	-	-	-	T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nucleocapsid	**T-->I**
28513	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	-	-	-	nucleocapsid	**V-->I**
28579	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	-	nucleocapsid	**N-->Y**
28696	G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	nucleocapsid	**G-->C**
29394	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	-	-	non-transl.	(-)

M, A/C heteroduplex; W, A/T heteroduplex; R, A/G heteroduplex; Y, C/T heteroduplex. The bold characteristics indicate the resulting nonsynonamous amino acid changes due to genetic variations (compared with the TOR2 isolate). P, Virus isolated from primary sample; C, virus isolated after passage in Vero cells; nontransl., the nontranslating region; unknown, the regions that are predicted to translate uncharacterized proteins; d, deletion.

Results

Full-Length Sequencing of SARS-CoV Isolates and Primary Samples. In our protocol of full-length sequencing, the SRAS-CoV genomes were divided into 25 fragments for PCR amplification and direct sequencing reactions (13). The sequence representing the dominant viral species was then derived. For virus isolates from Vero E6 cells, we were always successful in amplifying all fragments readily for sequencing work. However, for primary specimens (mainly from throat swabs), the success rate varied, probably depending on the viral titer in the samples. In fact, we succeeded in only 30–50% of the tested samples, all with viral titers over 100,000 copies per ml.

Comparison of Viral Sequences Obtained from Clinical Samples and Those Obtained After Passage in Vero E6 Cells. In nature infections, many RNA viruses exist as quasispecies with different extent of complexity. Therefore, the SARS virus isolated after passage in cell culture may or may not represent the major species in the host. To address this problem, we sequenced the paired virus isolates and viruses in the throat swabs of two patients from a clustered infection [patient 1 (Pt 1) and Pt 2). Pt 1 was the son of the first index case, and Pt 2 was the physician taking care of Pt 1. The whole-genome sequences of paired samples from these two patients were compared.

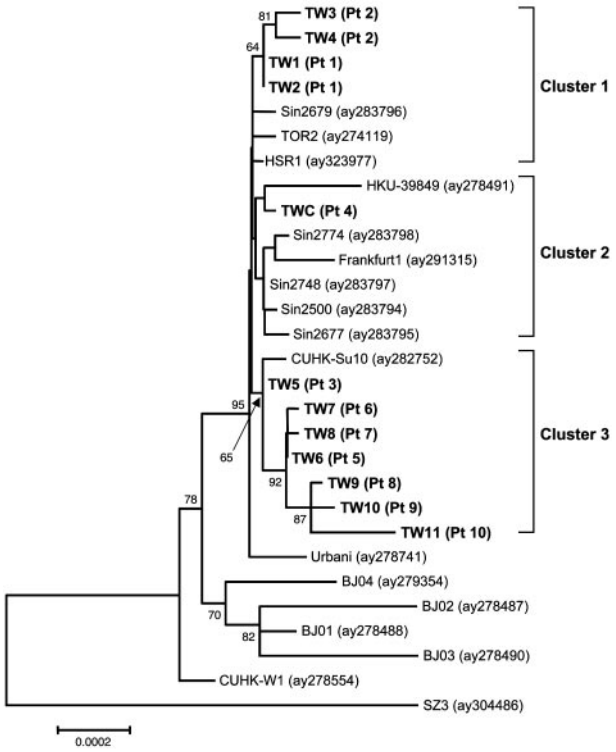

Fig. 2. Phylogenetic relationships of SARS virus isolates, including 12 isolates from Taiwan (TW), 16 isolates from other countries, and 1 isolate from a palm civet (SZ3). The neighbor-joining tree was constructed with bootstrap analysis based on the number of mutations in the viral genome, and the bootstrap values are added to the tree. The three clusters of transmission are indicated. The countries of origin of the sequences are as follows: TOR2, Canada; Sin2679, Sin2774, Sin2748, Sin2500, and Sin2677, Singapore; CUHK-Su10, HKU-39849, and CUHK-W1, Hong Kong; Urbani, Vietnam; BJ01, BJ02, BJ03, and BJ04, Beijing, China; HSR1, Italy; Frankfurt1, Germany; the others, Taiwan. Clusters 1 and 2 contain strains related to Hotel M origins.

For Pt 1, the two sequences (TW1 and TW2) showed homogeneous patterns, indicating no nucleotide polymorphism in any position of the viral genome. In addition, no sequence difference was present between the paired samples. For Pt 2, the genome from culture isolate (TW3) showed heterogeneity at two positions, with A/C polymorphism at position 1006 and C/T polymorphism at position 25341. It suggested a very mild degree of quasispecies in this viral isolate. However, both polymorphisms could not be detected in the corresponding primary specimen. Instead, there was a polymorphism of A/G at nucleotide 6404 in Pt 2's primary specimen.

Phylogenetic Analysis and Epidemiological Tracing of the Virus Origins. Because few genetic variations existed between primary samples and viruses isolated after limited passages of cultures, the viral sequences from either sample can be assumed to represent the major viral species present in patients. Therefore, in our attempt to clarify the origin of SARS-CoV in Taiwan by molecular epidemiological approaches, sequence data from both kinds of samples were included for further analysis.

In total, 12 full-length viral sequences of 10 patients from stage I and stage II of the Taiwan SARS epidemic were compiled. The phylogenetic tree analysis categorized them into three clusters, indicating that three independent infectious events had occurred in Taiwan (Fig. 2). Pts 1 and 2 belonged to the same cluster; Pt 4 belonged to another cluster. However, these three patients were located in the same lineage closer to the Hotel M lineages. The other patients (Pts 3 and 5–10) were in the third (also the major) cluster closer to another lineage of the SARS virus in Hong Kong (unrelated to Hotel M strains and represented by CUHK-Su10 isolate) (Fig. 2). The results supported the epidemiological observation that the SARS-CoVs in Taiwan originated from either Hong Kong or southern China.

Because there were some sporadic SARS cases in stage II of the outbreak without any traceable contact histories, two of such cases (Pts 9 and 10) were thus included for our analysis. We found that their viruses were most likely derived from the lineage of Hospital H. The subsequent transmission route could have been either through the clinic R (represented by Pt 8) or through some unidentified patients who got infected in Hospital H.

Nucleotide Variations in the SARS-CoV Genomes Suggest a Positive Selection. The sequences of our 12 virus isolates and the other 16 full-length virus isolates currently available from the public database (with accession numbers shown in the phylogenetic tree of Fig. 2) were compared. We also included the sequence of a coronavirus isolated from a palm civet, SZ3 (AY304486) for analysis. We summarize the genetic variations in Table 1, using the TOR2 isolate as a reference because it was the first SARS-CoV strain fully sequenced.

Patterns of nucleotide changes in different coding regions of the genome are listed in Table 2. By using the sequence of a palm civet

Table 2. Characterization of nucleotide substitutions in SARS-CoV isolates

		Between human and animal isolates			Within human isolates	
Genes	Sites	Ka, %	Ks, %	Ka/Ks	Syn change	Nonsyn change
orf1a	13,143	0.128	0.126	1.016	12	28
orf1ab_3′	8,067	0.055	0.353	0.156	9	14
spike	3,768	0.590	0.356	1.657	6	9
orf3	825	0.683	0.572	1.194	2	6
E	231	0.000	0.000	–	1	0
M	666	0.302	0.609	0.496	0	5
orf7	189	0.187	0.121	1.545	1	2
orf8	369	0.000	0.000	–	0	0
orf10	120	1.010	0.000	–	0	1
orf11	255	0.521	0.000	–	0	0
N	1,269	0.015	0.000	–	0	4
Total/average	28,902	0.183	0.238		32	69
Others	825				2	
Total	29,727				102	

Syn, synonymous; Nonsyn, nonsynonymous.

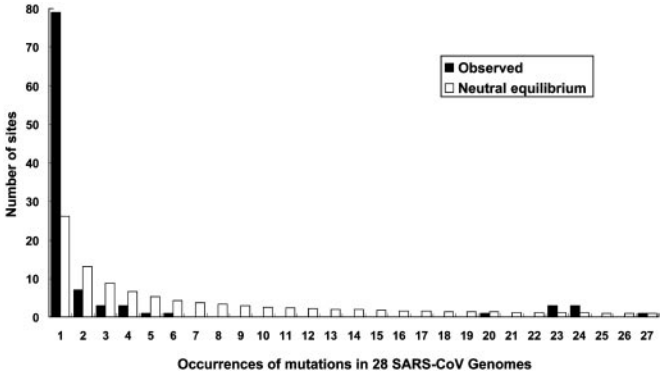

Fig. 3. The spectrum of mutation frequency in 28 SARS-CoV genomes. The derived nucleotide was inferred by reference to the sequence recovered from an animal (SZ3). The frequency of occurrences of these mutations in a sample of 28 SARS genomes is depicted on the *x* axis, whereas the *y* axis shows the number of sites with the corresponding mutations. The spectrum of mutation frequency showing the neutral equilibrium trend (open bars) is given by θi, where i is the number of occurrences (19); θ is the population parameter (2Nu) and is estimated by $\theta(1 + 1/2 + \cdots + 1/27) = 102$ (39).

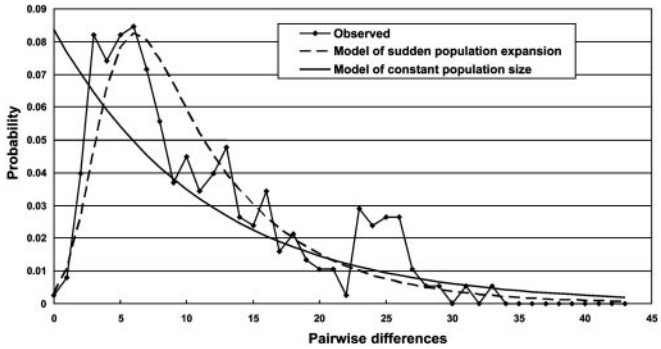

Fig. 4. The distribution of the observed and expected pairwise nucleotide site differences. The expected plot for constant (solid line) and growing (dashed line) population is shown with the observed distribution (solid line with squares).

coronavirus as the outgroup, the Ks and Ka values were calculated between human and animal isolates. Because of functional constraints, the synonymous mutation rate (Ks) is usually higher than the nonsynonymous mutation rate (Ka) for most of the protein coding genes. On the other hand, the reverse trend showing a higher nonsynonymous mutation rate often represents a sign of positive selection or adaptive evolution (23). For example, some genes associated with host–parasite interactions and male reproduction are found to have a higher nonsynonymous than synonymous mutation rate (24–28). In our analysis of the SARS-CoV genomes, 7 of 11 protein-coding regions exhibited a Ka higher than the Ks (Table 2). However, five ORFs (orf1a, orf7, orf10, orf11, and nucleocapsid) showed Ks values too low (smaller than average) for a conclusive comparison. Whereas the Ks of spike and orf3 were higher than the average, both of them exhibited the Ka > Ks and with Ka/Ks > 1, which strongly suggested that Darwinian selection had occurred on both genes. We also show the number of synonymous and nonsynonymous changes of individual genes within human isolates in Table 2.

The spectrum of mutation frequency of the 28 human SARS-CoV genome sequences, compared with the palm civet-derived SZ3, is illustrated in Fig. 3. Against the neutral equilibrium trend (open bar), the observed trend (filled bar) showed a significant excess of both low- and high-frequency mutations. The significance was examined with three neutrality tests. Tajima's D, which evaluates the normalized difference between θ_π and θ_ω, showed significant negative value ($D = -2.252, P < 0.01$). It indicated an excess of low-frequency polymorphisms and is expected after a selective sweep or a population bottleneck (19). Similar results were obtained with Fu and Li's D ($D = -3.67$, $P < 0.02$), which also measures the frequency distribution of polymorphisms and is sensitive to the number of singletons in the samples (20). Fay and Wu's H statistic (21) uses the frequency distribution of polymorphisms to test for an excess of high-frequency-derived variants compared with equilibrium neutral expectations. For SARS-CoV genomes, Fay and Wu's H test shows significant deviation from the neutral expectation ($P < 0.002$). The strong negative values obtain from the three tests confirmed an excess of both low- and high-frequency variants, evidently supporting a positive selection in SARS-CoV genomes (29).

SARS-CoV Equilibrium Curve and Mutation Rate. We next plotted the distribution of the observed pairwise nucleotide site differences (also called mismatch distribution) (Fig. 4). Clearly, the data fit poorly to the equilibrium curve. Instead of the smooth decline predicted by constant population size over time, the data exhibit a pronounced wave with a crest at roughly $\tau = 4.2$, the signature of sudden population explosion.

Because tau (τ) is the date of the growth or decline measured in units of mutational time ($\tau = 2ut$, where t is the time in generations and u is the mutation rate per sequence and per generation) (30), given the estimated generation time and date of the population expansion, we can estimate the mutation rate of the SARS genome. The generation time (defined as the time from release of a virion until it infects another cell and causes the release of a new generation of viral particles) of SARS virus is ≈2–3 days in Vero E6 cells. The outbreak of SARS started in early March, which is ≈2 months (or 60 days) before our last sampling in early May. According to the aforementioned estimation, the measurement of t is between 30 (60/2) to 20 (60/3). Thus, μg (mutation rate per genome) will be 0.11 to 0.07, which falls at the slowest end of the mutation rate of known RNA viruses (31).

Discussion

We have successfully determined the full-length sequence of the SARS-CoV genome in virus isolates from cell cultures as well as from primary clinical specimens. The sequence comparison between the culture isolate and primary isolate from the same patient revealed that most of the sequences were identical or with only a few variations. Accordingly, both kinds of samples can be used for sequencing, but primary samples directly taken from patients are preferred because they are readily available and the mutations occurring in the serial passages of cultures can be avoided.

When we compiled the sequences of all full-length SARS-CoV genomes for phylogenetic analysis, it seemed that three independent infection events had occurred in Taiwan. Two clusters were in the same lineage and were closer to the strains related to Hotel M in Hong Kong (Fig. 2, Pts 1, 2, and 4). The third cluster of patients was plausibly related to the strains from Hong Kong or Guangdong, but not linked to hotel M. For the first cluster, if we count the primary samples only, two new mutations were detected in the primary contact patient and then the infection stopped. Apparently, most SARS infections from either traceable or untraceable individuals in Taiwan belonged to the third cluster of patients derived from the same genetic origin. The molecular epidemiological analysis thus confirmed that the origin of the Taiwanese SARS epidemic was mainly from Hong Kong or Guangdong, rather than from Beijing. To prevent further outbreaks in the future, it will be critical to survey carefully people with a history of travel to SARS-affected areas.

When a SARS-CoV sequence recovered from an animal was used as the outgroup (18), the phylogenetic tree showed the Hong

Kong isolate CUHK-W1 (AY278554) to be located at the basal position. The phylogenetic tree further confirmed the point that the divergence among Hong Kong isolates was the earliest event followed by two routes of virus spreading, one to Beijing (Beijing cluster) and the other to the rest of the world, including Canada, Singapore, Taiwan, and Vietnam (Vietnam cluster). Interestingly, both Beijing and the Vietnam clusters were present in Hong Kong, and thus, the divergence existed before the spread of the disease. The history inferred above supports the epidemiological observation that SARS indeed originated from Hong Kong and its vicinity, although Ruan *et al*. (11) claimed that the Beijing cluster originated from Guangdong Province. However, because one Hong Kong isolate was placed at the basal position and both Beijing and Vietnam clusters were found in Hong Kong, the possibility that all of them actually originated from Hong Kong cannot be ruled out.

$Ka > Ks$ or $Ka/Ks > 1$ is the most stringent criterion of positive selection. Both spike and orf3 undoubtedly fit the criterion and thus indicated that both genes were subjected to Darwinian selection during virus evolution. Although the function(s) of orf3 is yet to be known, the spike protein is thought to be of particular importance in the infectious process, based on the studies of other coronaviruses because (*i*) it is the site for the virus to interact with the cognate receptor (32); (*ii*) it has fusion activities (33); and (*iii*) it contains sites against which major neutralizing antibodies are directed (34). The composition of this glycoprotein is therefore relevant to the ability of the virus to evade the host's immune system (35). Therefore, rapid amino acid change may help these molecules to evade the host immune response on the one hand and strengthen their ability to bind to cell surface antigens/receptors on the other.

When positive selection drives an advantageous mutation through a population to fixation, the neutral variation at linked sites is either eliminated (selection sweep) or increased (genetic hitchhiking) during the process. A population in recovery is characterized by an excess of new mutations at low frequency or linked variations at high frequency (19, 21). Thus, analyzing genetic variations provides a means to detect positive selection. Just as expected, the spectrum of the mutation frequency of the 28 SARS-CoV genome sequences showed an excess of both low- and high-frequency mutations significantly deviating from the neutral equilibrium curve (Fig. 3). Whereas the excess of low-frequency mutations might be solely an outcome of population bottleneck or purifying selection, the excess of high-frequency mutations is best explained by positive selection.

Clearly, our current data did not fit the equilibrium curve well (Fig. 4). Instead of the smooth decline predicted by constant population size over time, the data exhibit a pronounced wave, with a crest at roughly 4, the signature of sudden population explosion (30, 36). If the SARS virus had been associated with humans for a long time, the mismatch distribution would shift to the equilibrium curve of Fig. 4. Therefore, our present observation supports the notion that the current SARS virus was not present in humans until recently, which is consistent with the current serological studies (13, 37).

Furthermore, it is notable that the mutation rate of SARS-CoV is among the lowest of RNA viruses. Because the generation time of SARS-CoV has not been precisely defined *in vivo*, we calculated the mutation rate based on the generation time of the virus in Vero E6 culture: 2–3 days (13). Actually, the generation time seems not to deviate significantly from that in natural infections. Peiris *et al*. (37) followed the change of viral load in patients prospectively studied and showed 10^2- to 10^4-fold increases of the viral load in the nasopharyngeal aspirate from the 5th to the 10th day after onset of symptoms. This estimation is conservative in comparison with the generation time from other coronaviruses (6–8 h) (38). If we adopt the shorter generation time for calculation, the mutation rate would be even lower.

We observed limited nucleotide changes of the dominant viral species when sequences from cultures and from primary clinical specimens were compared (Pts 1 and 2). The data from the study of Tsui *et al*. (12) also supported this point: although only the spike gene was sequenced and compared, no additional mutations were detected in the seven viral samples they collected from the same SARS infection cluster. The low genomic mutation rate has to be confirmed in future studies. If our observations are true, this property would anticipate less difficulty than expected in future vaccine development against the SARS-CoV.

We thank C. K. James Shen for encouragement and kind help and Jennifer K. King for editing the English. H.-Y.W. is the recipient of a postdoctoral fellowship from Academia Sinica, Taiwan, and is also sponsored by the Ministry of National Defense, Taiwan. The study was supported by grants from the SARS Task Force and the National Research Program for Genomic Medicine (NSC92-2751-B-400-002-Y), National Science Council, Taiwan, and by National Health Research Institutes, Department of Health, Taiwan.

1. Lee, N., Hui, D., Wu, A., Chan, P., Cameron, P., Joynt, G. M., Ahuja, A., Yung, M. Y., Leung, C. B., To, K. F., *et al*. (2003) *N. Engl. J. Med*. **348,** 1986–1994.
2. Tsang, K. W., Ho, P. L., Ooi, G. C., Yee, W. K., Wang, T., Chan-Yeung, M., Lam, W. K., Seto, W. H., Yam, L. Y., Cheung, T. M., *et al*. (2003) *N. Engl. J. Med*. **348,** 1977–1985.
3. Poutanen, S. M., Low, D. E., Henry, B., Finkelstein, S., Rose, D., Green, K., Tellier, R., Draker, R., Adachi, D., Ayers, M., *et al*. (2003) *N. Engl. J. Med*. **348,** 1995–2005.
4. Parry, J. (2003) *BMJ* **326,** 999.
5. Drosten, C., Gunther, S., Preiser, W., van der Werf, S., Brodt, H. R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R. A., *et al*. (2003) *N. Engl. J. Med*. **348,** 1967–1976.
6. Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., *et al*. (2003) *N. Engl. J. Med*. **348,** 1953–1966.
7. Peiris, J. S., Lai, S. T., Poon, L. L., Guan, Y., Yam, L. Y., Lim, W., Nicholls, J., Yee, W. K., Yan, W. W., Cheung, M. T., *et al*. (2003) *Lancet* **361,** 1319–1325.
8. 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., *et al*. (2003) *Science* **300,** 1399–1404.
9. 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., *et al*. (2003) *Science* **300,** 1394–1399.
10. Lai, M. M. C., Holmes, K. V. (2001) in *Coronaviridae: The Viruses and Their Replication*, eds. Knipe D. M. & Howley P. M. (Lippincott Williams & Wilkins, London), pp. 1163–1186.
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., *et al*. (2003) *Lancet* **361,** 1779–1785.
12. Tsui, S. K., Chim, S. S. & Lo, Y. M. (2003) *N. Engl. J. Med*. **349,** 187–188.
13. Hsueh, P. R., Hsiao, C. H., Yeh, S. H., Wang, W. K., Chen, P. J., Wang, J. T., Chang, S. C., Kao, C. L. & Yang, P. C. (2003) *J. Emerging Infect. Dis*. **9,** 1163–1167.
14. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res*. **22,** 4673–4680.
15. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol*. **4,** 406–425.
16. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17,** 1244–1245.
17. Li, W. H. (1993) *J. Mol. Evol*. **36,** 96–99.
18. 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., *et al*. (2003) *Science* **302,** 276–278.
19. Tajima, F. (1989) *Genetics* **123,** 585–595.
20. Fu, Y. X. & Li, W. H. (1993) *Genetics* **133,** 693–709.
21. Fay, J. C. & Wu, C. I. (2000) *Genetics* **155,** 1405–1413.
22. Rozas, J. & Rozas, R. (1999) *Bioinformatics* **15,** 174–175.
23. Li, W.-H. (1997) in *Molecular Evolution* (Sinauer, Sunderland, MA).
24. Hughes, A. L. & Nei, M. (1988) *Nature* **335,** 167–170.
25. Wyckoff, G. J., Wang, W. & Wu, C. I. (2000) *Nature* **403,** 304–309.
26. Yang, Z. & Bielawski, J. P. (2000) *Trends Ecol. Evol*. **15,** 496–503.
27. Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F. & Aquadro, C. F. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 7375–7379.
28. Wang, H. Y., Tang, H., Shen, C.-K. J. & Wu, C. I. (2003) *Mol. Biol. Evol*. **20,** 1795–1804.
29. Ewen, W. J. (1979) in *Mathematical Population Genetics* (Springer, Berlin).
30. Rogers, A. R. & Harpending, H. (1992) *Mol. Biol. Evol*. **9,** 552–569.
31. Drake, J. W. & Holland, J. J. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 13910–13913.
32. Collins, A. R., Knobler, R. L., Powell, H. & Buchmeier, M. J. (1982) *Virology* **119,** 358–371.
33. De Groot, R. J., Van Leen, R. W., Dalderup, M. J., Vennema, H., Horzinek, M. C. & Spaan, W. J. (1989) *Virology* **171,** 493–502.
34. Jimenez, G., Correa, I., Melgosa, M. P., Bullido, M. J. & Enjuanes, L. (1986) *J. Virol*. **60,** 131–139.
35. La Monica, N., Banner, L. R., Morris, V. L. & Lai, M. M. (1991) *Virology* **182,** 883–888.
36. Rogers, A. R. & Jorde, L. B. (1995) *Hum. Biol*. **67,** 1–36.
37. Peiris, J. S. M., Chu, C. M., Cheng, V. C. C., Chan, K. S., Hung, I. F. N., Poon, L. L. M., Law, K. I., Tang, B. S. F., Hon, T. Y. W., Chan, C. S., *et al*. (2003) *Lancet* **361,** 1767–1772.
38. Hirano, N., Fujiwara, K. & Matumoto, M. (1976) *Jpn. J. Microbiol*. **20,** 219–225.
39. Watterson, G. A. (1975) *Theor. Popul. Biol*. **7,** 256–276.