Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China

The Chinese SARS Molecular Epidemiology Consortium*

*The complete list of authors and their affiliated institutions can be found in the supplementary text on *Science Online*.

Sixty-three SARS coronavirus genomic sequences derived from the early, middle and late phases of the severe acute respiratory syndrome (SARS) epidemic were analyzed. Genotypes characteristic of each phase were discovered and the earliest genotypes were similar to animal SARS-like coronaviruses. Major deletions were observed in the Orf8 region, both at the start and the end of the epidemic. The neutral mutation rate of the viral genome was constant but the amino acid substitution rate of the coding sequences slowed during the course of the epidemic. The spike protein initially showed the strongest responses to positive selection pressures followed by subsequent purifying selection and eventual stabilization.

Severe Acute Respiratory Syndrome (SARS) first emerged in Guangdong Province, China and subsequently SARS coronavirus (SARS-CoV) was identified as the causative agent (1–5). It remains a challenge to establish the relationship between observed genomic variations and the biology of SARS (4–8). Recent molecular epidemiological studies have identified characteristic variant sequences in SARS Co-V for tracking disease transmission (7, 9–11). Evidence suggests that SARS-CoV emerged from non-human sources (8, 12), while in this study, we sought epidemiological and genetic evidence for viral adaptation to human beings through molecular investigations of the characteristic viral lineages found in China (13).

Based on epidemiological investigations (14), we divided the course of the epidemic into early, middle and late phases (Fig. 1). The early phase is defined as the period between the first emergence of SARS to the first documented super-spreader event (SSE) (13). The middle phase refers to the ensuing events up to the first cluster of SARS cases in a hotel (Hotel M) in Hong Kong (15). Cases following this cluster fall into the late phase.

The early phase was initially characterized by a series of seemingly independent cases. Eleven index cases that had arisen locally in the absence of any contact history were identified from different geographical locations within Guangdong Province (fig. S1). This phenomenon was observed from the retrospectively identified SARS index patient from the city of Foshan (onset date, November 16,

2002) (13) through to an index patient from the city of Dongguan (onset date, March 10, 2003). All of these cases were confined to regions directly west of Guangzhou, the capital city of Guangdong Province, and to the city of Shenzhen in the south, with no cases being reported to the north or east of Guangzhou (Fig. 1) (fig. S1). This region, the Pearl River Delta, has enjoyed rapid economic development since the late 1970s, leading to the adoption of culinary habits requiring exotic animals. Seven of these 11 cases had documented contact with wild animals. In contrast to the apparently independent seeding of the earliest cases, the rest of the epidemic was characterized by SSEs and clusters of cases that were epidemiologically linked (Fig. 1) (fig. S1) (10–13, 15, 16).

The first major SARS outbreak occurred in a hospital, HZS-2, in the city of Guangzhou, beginning on January 31, 2003 where an SSE was identified to be associated with more than 130 primary and secondary infections, and in which, 106 were hospital-acquired cases. Doctor A, a nephrologist who worked in this hospital, visited Hong Kong and stayed in Hotel M on February 21, 2003. Other visitors to the hotel later became infected with SARS-CoV (13, 15). This has further led to the transmission of SARS to Vietnam, Canada, Singapore and the United States (17) with two further SSEs in Hong Kong, each resulting in the virus being transmitted to >100 contacts (10, 16).

Genomic sequence data for the SARS-CoV were largely derived from isolates linked to the Hotel M cluster (6) and thus, predominantly from the late phase of the epidemic. We determined 29 SARS-CoV genomic sequences obtained from 22 patients from Guangdong Province, with disease onset dates in all three phases of the epidemic, and from two patients from the late phase in Hong Kong. To eliminate mutational noise, we assumed that sequence variants associated with common ancestry, but not arising in cell culture, should be seen in multiple isolates (7). Meanwhile, critical genomic variations or complete genome sequences of certain virus isolates were verified by sequencing the reverse transcriptase-polymerase chain reaction (RT-PCR) products derived directly from patient specimens (14). The genomic sequences obtained were compared with 32 human SARS-CoV and 2 SARS-like-CoV sequences from the Himalayan

palm civets (*Paguma larvata*) available at GenBank as of the end of September, 2003 (Fig. 2).

Only two major genotypes predominated during the early phase of the epidemic. Five isolates were found to contain a 29-nucleotide (nt) sequence that is absent in most of the publicly available SARS-CoV sequences, while another four isolates showed a previously unreported 82-nt deletion in the same region of the genome, Orf8 (fig. S2 and table S1). The former sequence is represented by the GZ02 isolate [all GenBank accession numbers are listed in (14)], and is used as the reference for annotation throughout this study. All of the isolates exhibiting this sequence (GZ02, HGZ8L1-A, HSZ-A, HSZ-B, and HSZ-C, Fig. 2) were obtained from patients with contact histories traceable to some of the earliest independent cases in Guangzhou and was not detected in any of the later isolates. It is noteworthy that this sequence with the 29-nt segment is identical to the genomic sequence of coronaviruses isolated from animals in a Shenzhen live animal market (8).

Three of the SARS-CoV genome sequences (ZS-A, ZS-B, ZS-C, Fig. 2) with the 82-nt deletion were obtained from samples of very early cases from Zhongshan city. This 82-nt deletion was further confirmed by RT-PCR directly on an additional stool sample. A sequence with an identical 82-nt deletion has also been observed in coronaviruses isolated from farmed civets in Hubei Province, China (18). It is thus interesting to note that both sequences of the early phase were identified from other mammalian hosts. They provided a link to support the notion that early human infection of SARS-CoV may have originated from wild animals (8, 12).

In contrast to the early phase, a SARS-CoV sequence with the 29-nt deletion was observed during the middle phase that dominated the viral population for the rest of the epidemic (4, 5, 7). Although this shift in genome size might be due to chance, deletion events appeared to be overrepresented in the Orf8 region. A fourth sequence with the 82-nt deletion was obtained from a Guangzhou patient (HGZ8L1-B), who was infected in the same ward as one of the patients where the longest sequence was obtained (HGZ8L1-A) (see above). Furthermore, a lung biopsy of a patient from the middle phase was found to contain two SARS-CoV genotypes, with the 29-nt and the 82-nt deletions, respectively (fig. S3). Remarkably, another genotype with a 415-nt deletion resulting in the loss of the whole Orf8 region was isolated and confirmed in two Hong Kong patients with disease onset from mid-May, 2003 (Fig. 2) (fig. S2) (19).

As the majority of deletions observed in the SARS-CoV genome occurred in the Orf8 region with no apparent effect on the survival of the virus, it is tempting to suggest that this region is either non-coding or coding for a functionally unimportant putative protein (table S1). On the other hand, it is interesting to note that antiparallel reverse symmetrical

sequences were readily predicted around the deletion sites (fig. S2), which might account for the high deletion rates in this region. Whether such hairpin structures actually play a role in regulating either RNA replication or mRNA transcription in SARS-CoV is a subject for future studies.

Besides the deletion variants, 299 single nucleotide variations (SNVs) were detected among the 63 sequences. Eighty-five of these variant loci were seen in more than one of the human SARS-CoV sequences. Among them, 52 were predicted to cause amino acid changes (non-synonymous variations) (table S2). When the epidemiologically-determined transmission paths and SNV genotype data are combined, markers for genotypes characteristic of different lineages are evident (Fig. 2) (table S2).

Viruses of the early phase have the characteristic motif of **G:A:C:G:C** at the GZ02 reference nucleotide residues 17564, 21721, 22222, 23823, and 27827, with the bold SNVs matching the **C:G:C:C** motif identified previously (7) (Fig. 2). This motif is shared by almost all early Guangzhou and Zhongshan isolates together with the animal SARS-like coronavirus isolates (SZ3 and SZ16) (8). Along with the disappearance of viruses containing the 29-nt segment, the middle phase of the epidemic was characterized by the occurrence of genotypes with the **G:A:C:T:C** motif (Fig. 2). All of the middle phase genotypes demonstrate this common motif but can be further classified into two variant groups based on other SNVs (table S2). One group was represented by the isolates related to the Hospital HZS-2 outbreak (HZS2-A, HZS2-B, HZS2-C, HZS2-D, HZS2-E, HGZ8L-2). The other group was represented by the Hong Kong CUHK-W1 isolate that originated from Shenzhen (9) together with the early Beijing isolates BJ01, BJ02, and BJ03, traceable to Guangdong. The transition between the characteristic motifs of the early and middle phases represented a G→T transversion at 23823 and is predicted to cause a D→Y change at amino acid residue 778 of the spike (S) protein (fig. S4).

An additional A→G transition at nucleotide 21721 (Fig. 2) (fig. S4) was identified in one isolate from a secondarily infected patient from Hospital HZS-2 with disease onset on February 7, 2003 (HZS2-Fc) (Fig. 2). This sequence was additionally confirmed by direct sequencing of the RT-PCR product from an oropharyngeal swab of this patient (HZS2-Fb). This mutation is predicted to cause a D77G amino acid switch in the S protein (fig. S4) and the **G:G:C:T:C** motif is so far genotypically the closest sequence to that of the Hotel M outbreak (**T:G:T:T:T**) (Fig. 2) (15). Epidemiologically, this patient is potentially linked to the Hotel M outbreak through her contact with Doctor A during the first 3 days of illness. Thus, Doctor A was possibly infected with this viral variant.

Three additional C→T transitions at nucleotide residues 17564, 22222, and 27827 are observed in the Hotel M-associated SARS-CoV genotypes (Fig. 2) (table S2). These SNVs are predicted to cause amino acid switches in the nonstructural polyprotein (E1389D), the S protein (T244I) and Orf8a (R17C), respectively. This T:G:T:T:T motif is shared by the sequences of all isolates infected from and after the Hotel M cluster (7), including the Hong Kong Amoy Gardens isolates (10), and the more recent isolates from Zhejiang (ZJ01), Taiwan (11), and Guangdong (GZ-B, GZ-C, GZ-D) (Fig. 2) (table S2). This motif is also conserved in the late 415-nt deletion variant in Hong Kong with the exception of nucleotide 27827, which falls within the deleted segment (19). Thus, surprisingly few genotypes predominated during the late phase of the epidemic.

The characteristically high mutation rate of RNA viruses (20) may give rise to strains with increased virulence (21) that can either escape host defenses (22) or change their tissue tropism (23). We noticed that the neutral mutation rate for SARS-CoV during this epidemic was almost constant (fig. S5) (14), and was estimated to be 8.26×10^{-6} ($\pm 2.16 \times 10^{-6}$). This is similar to the values obtained for known RNA viruses and is about one third that for the human immunodeficiency virus (24, 25). In contrast to the constant rate of synonymous variations, the non-synonymous mutation rates were variable for the three epidemic phases (table S3) (14). It was the predicted domains of the S protein, responsible for viral-host receptor recognition or internalization (26) that underwent the most extensive amino acid substitutions (fig. S4).

Between the coronavirus sequences of the palm civets (SZ3 or SZ16) and each of the human SARS-CoV sequences, the ratios of the rates of non-synonymous to synonymous changes (Ka/Ks) for the S gene sequences were always greater than 1, indicating an overall positive selection pressure. However, pairwise analysis of the Ka/Ks for the genotypes in each epidemic group (fig. S6) (14) shows that the average Ka/Ks for the early phase was significantly larger than that for the middle phase, which in turn was significantly larger than the ratio for the late phase, which in fact was significantly less than one (table S3). These data indicate that the S gene showed the strongest positive selection pressures initially, with subsequent purifying selections and eventual stabilization. For Orf1a, we observed a similar pattern as that for S gene (table S3). In contrast, Orf1b (nt coordinate: 13398 – 21485) seems to be undergoing purifying selection during the whole course of the epidemic. In fact, it is the most conserved genomic region of the SARS-CoV (7).

Our analysis thus suggests that adaptive pressures operated on the SARS-CoV genome, but stabilized during the late epidemic phase with the emergence of a predominant genotype. Alternatively, sampling bias for cases related to SSEs (27) may distort the data. We believe that such a

sampling strategy may be justifiable from a public health perspective, as the viral genotypes associated with the SSEs are the most epidemiologically important. However, to explore the possibility of bias, we estimated the date for the most recent common ancestor of the samples available. Based on the observed neutral mutation rate, this date was estimated to lie in mid November 2002 (95% confidence interval: early June 2002 and late December 2002) (14). This result is consistent with the onset date of November 16, 2002 for the earliest index patient from Foshan (13) and supports the finding that the genotypes we studied from the early, middle and late phases represent different stages of evolution of the same viral lineage. This is further evident from the remarkable correlation between the molecular clustering and epidemiological grouping of the genotypes throughout the epidemic (Fig. 2) (table S2).

In tracing the molecular evolution of the SARS-CoV in China, we observed that the epidemic started and ended with deletion events, together with a progressive slowing of the non-synonymous mutation rates and a common genotype that predominated during the latter part of the epidemic. The mechanistic explanation for the selective adaptation and purification processes that led to such genomic evolutionary changes in the SARS-CoV requires further work (28). Nevertheless, this study has provided valuable clues to aid further investigation of this remarkable evolutionary tale.

We have sequenced the complete S gene (GenBank accession number AY525636) from an oropharyngeal swab sample (sampling date: Dec. 22, 2003) collected from the most recent index patient of the city of Guangzhou (onset date: Dec. 16, 2003; hospitalized on Dec. 20, 2003; www.wpro.who.int/sars/docs/pressreleases/pr_27122003.asp). Phylogenetic analysis of this S gene sequence with those from the human SARS-CoV and palm civet SARS-like CoV indicated that this most recent case of SARS-CoV is much closer to the SARS-like CoV of palm civet than any other human SARS-CoV ever detected in the previous epidemic (fig. S7 and table S4). Because it is evidently different from the recent laboratory infections in Singapore (www.who.int/csr/don/2003_09_24/en) and Taiwan (www.who.int/mediacentre/releases/2003/np26/en), it strengthens the argument for animal origin of the human SARS epidemic.

References and Notes

1. R. A. Fouchier *et al.*, *Nature* **423**, 240 (2003).
2. T. G. Ksiazek *et al.*, *N. Engl. J. Med.* **348**, 1953 (2003).
3. C. Drosten *et al.*, *N. Engl. J. Med.* **348**, 1967 (2003).
4. P. A. Rota *et al.*, *Science* **300**, 1394 (2003).
5. M. A. Marra *et al.*, *Science* **300**, 1399 (2003).
6. G. Vogel, *Science* **300**, 1062 (2003).
7. Y. J. Ruan *et al.*, *Lancet* **361**, 1779 (2003).
8. Y. Guan *et al.*, *Science* **302**, 276 (2003).

9. S. K. W. Tsui, S. S. C. Chim, Y. M. D. Lo, *N. Engl. J. Med.* **349**, 187 (2003).
10. S. S. C. Chim *et al.*, *Lancet* **362**, 1807 (2003).
11. R. W. K. Chiu, S. S. C. Chim, Y. M. D. Lo, *N. Engl. J. Med.* **349**, 1875 (2003).
12. J. S. Rest, D. P. Mindell, *Infect. Genet. Evol.* **3**, 219 (2003).
13. N. S. Zhong *et al.*, *Lancet* **362**, 1353 (2003).
14. Supporting materials are available on *Science Online*.
15. K. W. Tsang *et al.*, *N. Engl. J. Med.* **348**, 1977 (2003).
16. N. Lee *et al.*, *N. Engl. J. Med.* **348**, 1986 (2003).
17. Centers for Disease Control and Prevention, *Morb. Mortal. Wkly. Rep.* **52**, 241 (2003).
18. SARS-like coronaviruses were isolated from palm civets farmed domestically in Hubei Province, China, by Hu *et al.* at the Wuhan Institute of Virology, Chinese Academy of Sciences. Partial genome sequencing revealed an 82-nt deletion within the Orf8 region, which is identical to that found in human SARS-CoV isolates from the early patients of Zhongshan, Guangdong Province, China. Contamination can be ruled out since no human SARS-CoV isolate with the 82-nt deletion has ever been found in that institute or has been isolated in that region of China.
19. The SARS-CoV sequence with the 415-nt deletion (CUHK-LC2, CUHK-LC3, CUHK-LC4, and CUHK-LC-5) was obtained from two SARS patients whose disease was linked to a late cluster of SARS cases in Hong Kong. Both patients had disease onset in mid-May, 2003. The CUHK-LC2 sequence was initially obtained from the culture isolate of a throat wash specimen of an infected hospital health care worker and was later confirmed from the same specimen directly. CUHK-LC3, CUHK-LC4 and CUHK-LC5 were obtained from three different nasal swab specimens both directly and from the culture supernatants of an elderly patient who acquired SARS in the same hospital.
20. M. M. C. Lai, K. V. Holmes, in *Fields Virology*, D. M. Knipe, P. M. Howley, Eds. (Lippincott Williams & Wilkins, New York, ed. 4, 2001), chap. 35.
21. E. G. Brown, H. Liu, L. C. Kit, S. Baird, M. Nesrallah, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 6883 (2001).
22. S. H. Seo, E. Hoffmann, R. G. Webster, *Nature Med.* **8**, 950 (2002).
23. D. Rasschaert, M. Duarte, H. Laude, *J. Gen. Virol.* **71**, 2599 (1990).
24. W-H. Li, M. Tanimura, P. M. Sharp, *Mol. Biol. Evol.* **5**, 313 (1998).
25. J. W. Drake, J. J. Holland, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 13910 (1999).
26. Z. Luo, A. M. Matthews, S. R. Weiss, *J. Virol.* **73**, 8152 (1999).
27. R. M. Bush, C. B. Smith, N. J. Cox, W. M. Fitch, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6974 (2000).
28. P. W. Ewald, *J. Urban Health* **75**, 480 (1998).
29. See SOM Text at *Science Online* for acknowledgments.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1092002/DC1

Materials and Methods

SOM Text

References and Notes

Figs. S1 to S7

Tables S1 to S4

29 September 2003; accepted 14 January 2004

Published online 29 January 2004; 10.1126/science.1092002

Include this information when citing this paper.

Fig. 1. The tri-phasic SARS epidemic in Guangdong Province, China. The daily number of SARS cases reported in the Guangdong Province and in particular, the city of Guangzhou are presented. The early, middle and late phases of the epidemic are defined in the text. The inserted panel highlights the geographical distribution of cases belonging to the early phase by administrative districts of the Guangdong Province. The detailed data for individual cities are presented in fig. S1.

Fig. 2. Genotype clustering of the SARS-CoV during the course of the epidemic. An unrooted phylogenetic tree of the SARS-CoV is constructed from 61 human SARS-CoV genomes and 2 SARS-like-CoV sequences from palm civets. Only those variant sequences (including deletions) that were present in at least two independent samples were used for tree construction (table S2). The map distance between individual sequences represents the extent of genotypic difference. The 5-nt motifs (see text) that characterized the phylogenetically-related genotypes are indicated in boxes. The genomic sequences are named in concordance with their GenBank nomenclature and are represented in different colors according to the genotype clusters determined by our scoring method (table S2). Genotypes with major deletions are marked specifically (see text). All other genotypes (unmarked) had the 29-nt deletion. This 29-nt deletion was specifically marked for three genotypes, namely GZ-A, JMD and GZ50, to indicate their special clustering within the early phase isolates.

T G T T T

TC3 TC2
TWJ TWY TWK
TWS
TWH
CUHK-AG02
CUHK-AG03
TC1
CUHK-AG01

CUHK-SU10
CUHK-LC1

GZ-D

HKU-39849
TWJ1

ZJ01

Sin2679
HSR

Sin2774
Frankfurt

Sin2677
Sin2500

Sin2748
TWC

GZ-C
GZ-B

TOR2

T G T T -

CUHK-LC2

CUHK-LC5

CUHK-LC3
CUHK-LC4

415-nt deletion

G G C T C

HZS2-Fc
HZS2-Fb

HSZ2-A
HSZ2-B

HSZ2-C
HGZBL2

HZS2-D
HZS2-E

CUHK-W1

BJ04

BJ01

BJ03

BJ02

HGZBL1-B

JMD

GZ50
GZ-A

29-nt deletion

ZS-B
ZS-A

82-nt deletion

G A C C C

GZ02

GD01

SZ3

SZ16

HGZBL1-A

HSZ-Bb

HSZ-A

HSZ-Bc

HSZ-Cb

HSZ-Cc

Long genome with the 29-nt segment