Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human

Huai-Dong Song[a,b], Chang-Chun Tu[b,c], Guo-Wei Zhang[a,b], Sheng-Yue Wang[b,d], Kui Zheng[e], Lian-Cheng Lei[c], Qiu-Xia Chen[e], Yu-Wei Gao[c], Hui-Qiong Zhou[e], Hua Xiang[c], Hua-Jun Zheng[d], Shur-Wern Wang Chern[f], Feng Cheng[a], Chun-Ming Pan[a], Hua Xuan[c,g], Sai-Juan Chen[a,g], Hui-Ming Luo[b,e], Duan-Hua Zhou[b,h], Yu-Fei Liu[h], Jian-Feng He[e], Peng-Zhe Qin[h], Ling-Hui Li[e], Yu-Qi Ren[i], Wen-Jia Liang[e], Ye-Dong Yu[i], Larry Anderson[f], Ming Wang[g,h], Rui-Heng Xu[e,g], Xin-Wei Wu[b,h], Huan-Ying Zheng[b,e], Jin-Ding Chen[b,j], Guodong Liang[k], Yang Gao[h], Ming Liao[j], Ling Fang[e], Li-Yun Jiang[h], Hui Li[e], Fang Chen[h], Biao Di[h], Li-Juan He[h], Jin-Yan Lin[e,g], Suxiang Tong[f,g], Xiangang Kong[g,l], Lin Du[g,h], Pei Hao[b,m,n], Hua Tang[b,o], Andrea Bernini[b,p], Xiao-Jing Yu[m], Ottavia Spiga[p], Zong-Ming Guo[n], Hai-Yan Pan[n], Wei-Zhong He[n], Jean-Claude Manuguerra[q], Arnaud Fontanet[q], Antoine Danchin[q], Neri Niccolai[g,p], Yi-Xue Li[g,m,n], Chung-I Wu[g,o], and Guo-Ping Zhao[d,m,r,s]

[a]State Key Laboratory for Medical Genomics/Pôle Sino-Français de Recherche en Sciences du Vivant et Génomique, Ruijin Hospital Affiliated to Shanghai Second Medical University, 197 Rui Jin Road II, Shanghai 200025, China; [c]Changchun University of Agriculture and Animal Sciences, Changchun 130062, China; [d]Chinese National Human Genome Center, 250 Bi Bo Road, Zhang Jiang High Tech Park, Shanghai 201203, China; [e]Guangdong Center for Disease Control and Prevention, 176 Xingangxi Road, Guangzhou 510300, Guangdong, China; [f]Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333; [h]Guangzhou Center for Disease Control and Prevention, 23 Third Zhongshan Road, Guangzhou 510080, Guangdong, China; [i]Guangdong Provincial Veterinary Station of Epidemic Prevention and Supervision, Guangzhou 510230, China; [j]College of Veterinary Medicine, South China Agriculture University, Guangzhou 510246, China; [k]National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 100052, China; [l]National Key Laboratory of Veterinary Biotechnology, Harbin Veterinary Research Institute, Chinese Academy of Agriculture Sciences, Harbin 150001, China; [m]Bioinformation Center/Institute of Plant Physiology and Ecology/Health Science Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China; [n]Shanghai Center for Bioinformation Technology, 100 Qinzhou Road, Shanghai 200235, China; [o]Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637; [p]Biomolecular Structure Research Center and Department of Molecular Biology, University of Siena, Via A. Fiorentina 1, I-53100 Siena, Italy; [q]Institut Pasteur, 25, Rue du Docteur Roux, 75724 Paris Cedex 15, France; and [r]State Key Laboratory of Genetic Engineering/Department of Microbiology, School of Life Science, Fudan University, 220 Handan Road, Shanghai 200433, China

Communicated by Zhu Chen, Shanghai Institute of Hematology, Shanghai, People's Republic of China, December 22, 2004 (received for review November 20, 2004)

The genomic sequences of severe acute respiratory syndrome coronaviruses from human and palm civet of the 2003/2004 outbreak in the city of Guangzhou, China, were nearly identical. Phylogenetic analysis suggested an independent viral invasion from animal to human in this new episode. Combining all existing data but excluding singletons, we identified 202 single-nucleotide variations. Among them, 17 are polymorphic in palm civets only. The ratio of nonsynonymous/synonymous nucleotide substitution in palm civets collected 1 yr apart from different geographic locations is very high, suggesting a rapid evolving process of viral proteins in civet as well, much like their adaptation in the human host in the early 2002–2003 epidemic. Major genetic variations in some critical genes, particularly the *Spike* gene, seemed essential for the transition from animal-to-human transmission to human-to-human transmission, which eventually caused the first severe acute respiratory syndrome outbreak of 2002/2003.

The prompt identification of a novel human coronavirus (CoV) as the etiologic agent of severe acute respiratory syndrome (SARS) demonstrated the power deriving from coordinate integration of clinical investigation and molecular virology (1–4). SARS-CoV-like virus was isolated from a few Himalayan palm civets (*Paguma larvata*) and a raccoon dog (*Nyctereutes procyonoides*) at a Shenzhen food market during the SARS epidemic of 2002–2003 (May 7 and 8, 2003). Their genomic sequences displayed 99.8% identity with that of the human SARS-CoV (5). Together with the evidence of a significant high ratio of positive cases bearing the anti-SARS-CoV antibody in the population with a history of close contact to these animals, animal–human interspecies transmission of SARS-CoV was first proposed. Meanwhile, molecular epidemiological approaches were effectively conducted for better understanding the origin, route of transmission, and evolution of SARS-CoV (6, 7). Characteristic genotypes were identified for viruses of different transmitting lineages, and the disease episodes were categorized into different epidemiological phases based on the combination of classical epidemiology analysis and molecular phylogeny analysis using well represented viral genomic sequences. It was particularly interesting that critical intermediate single-nucleotide variations (SNVs) were found among isolates collected between connective phases along with their transmission paths. It also strongly suggested an animal origin of the human SARS-CoV and its viral adaptation to human hosts (7). However, direct evidence of animal-to-human infection has yet to be provided, and the molecular mechanism that enabled the virus to switch hosts has not been investigated.

After the first epidemic of SARS ended in July 2003, as announced by the World Health Organization (WHO) (www.who.int/csr/don/2003_07_05/en), scattered new cases were reported. Unlike the cases of laboratory infections reported from Singapore (www.who.int/csr/don/2003_09_24/en), Taiwan (www.who.int/csr/don/2003_12_17/en), and Beijing (www.who.int/csr/don/2004_05_18a/en), the four confirmed SARS patients of the 2003–2004 episode in the city of Guangzhou, China, were all community-infected cases without obvious human-to-human contact history related to SARS (see *Materials and Methods*). In this report, using the sequence data of viruses obtained from these human patients as well as from palm civets collected at the same period in the same region, we were able to delineate the characteristics of the cross-host evolution of SARS-

Abbreviations: SARS, severe acute respiratory syndrome; CoV, coronavirus; SNV, single-nucleotide variation; WHO, World Health Organization; CDSs, coding DNA sequences; S, Spike; MRCA, most recent common ancestor; *Ks*, number of synonymous substitution per synonymous site; A/S, ratio of nonsynonymous/synonymous substitution numbers.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession numbers are listed in Table 1, which is published as supporting information on the PNAS web site).

[b]H.-D.S., C.-C.T., G.-W.Z., S.-Y.W., H.-M.L., D.-H.Z., X.-W.W., H.-Y.Z., J.-D.C., P.H., H.T., and A.B. contributed equally to this work in performing the research.

[g]H.X., S.-J.C., M.W., R.-H.X., J.-Y.L., S.T., X.K., L.D., N.N., Y.-X.L., and C.-I.W. contributed equally to this work in organizing the research.

[s]To whom correspondence should be addressed. E-mail: gpzhao@sibs.ac.cn.

www.pnas.org/cgi/doi/10.1073/pnas.0409608102

CoV over a short period. This is an essential step for understanding the genetic process of the adaptation of an animal virus to a human host.

Materials and Methods

Epidemiological Investigation and Sample Collection. Official epidemiological records about SARS cases occurring during the 2003–2004 period from both the Guangdong Center for Disease Control and Prevention and the Guangzhou Center for Disease Control and Prevention were reviewed. These records were matched with the information released by WHO (www.who.int/csr/don/2004_01_05/en, www.who.int/csr/don/2004_01_27/en, and www.who.int/csr/don/2004_01_31/en).

Human patient samples were collected by the virologists of the Guangzhou Center for Disease Control and Prevention. Palm civet samples from the animal cage of the restaurant TDLR were collected by WHO experts, whereas those from the Guangzhou food market were collected by virologists of the SARS Consortium of the Minister of Agriculture of the Chinese Central Government.

Sequencing Strategy and Procedures. The sequencing strategy was basically the same as described (7). However, all new sequences we obtained during this study were derived directly from RT-PCR products of specimens from individual human patients or animals (or their cages, as indicated). Another set of nested PCR primers with shorter genomic fragments being amplified was used when the regular primer set failed to amplify the corresponding genomic regions. This strategy was successful in obtaining more genomic DNA fragments being amplified for sequencing. All of the International Nucleotide Sequence Database Collaboration/GenBank accession nos. for SARS-CoV sequences analyzed in the text are listed in Table 1, which is published as supporting information on the PNAS web site.

***In Silico* Sequence Analysis.** Whole-genome sequence alignments were generated by using CLUSTALW, Ver. 1.83 (www.ebi.ac.uk/clustalw) with the default DNA weight matrix for the 96 SARS-CoV genomic sequences analyzed in this study (91 from human patients and 5 from palm civets). The same method was used for the alignment analysis of *Spike* (*S*) genes from 14 animal samples (in addition to the five sequences from palm civet host used in whole-genome sequence alignment, seven sequences from other palm civet samples of the Guangzhou food market and two sequences, SZ1 and SZ13, from palm civet samples of the 2002–2003 epidemic were added) and 92 human SARS-CoV sequences. Compared with the 91 sequences used in the whole genome analysis, the previously sequenced *S* gene (GD03T13) from the first patient (GZ03-01) (7) and the newly sequenced *S* gene from the third patient (GZ03-03) of the 2003–2004 epidemic were added, whereas the GZ-D of the 2002–2003 epidemic was deleted due to the incompleteness of the sequence. The scoring algorithm used to determine the variant loci characteristic of the SARS-CoV genotypes and to allow the segregation of the SARS-CoV genotypes into major groups was previously described (7), and the outcome of this analysis is listed in Table 2, which is published as supporting information on the PNAS web site.

For purposes of illustration, we adopted the following nomenclature as shown in Fig. 1: PC for palm civet and HP for human patient. Both of them were suffixed with 03 or 04 to specify the 2002–2003 or 2003–2004 epidemics, respectively. Furthermore, the HP03 events are followed by E, M, or L, representing the early, middle, or late phases of the 2002–2003 epidemic (7).

Analysis of the Phylogenetic Relationship Among Different Transmission Lineages of the Early Samples of SARS-CoV Sequences. The consensus genomic nucleotide sequences for groups PC04, HP04, PC03 and individual transmission lineages of HP03E (GZ, HSZ, and ZS) were used to construct the neighbor-joining tree (8). Tajima's relative rate test (9) was then performed to see whether there is significant difference between the distance from PC03 to PC04 and that from PC04 to HP04.

Calculating the Average Number of Nucleotide Difference *D* Between Two Sample Groups. We used n_1 and n_2 to denote the sample sizes for groups 1 and 2. All of the singleton sites were excluded for the sequences between the two groups. The total number of the nucleotide difference $D_{i,j}$ ($i = 1, \ldots, n_1; j = 1, \ldots, n_2$) was then calculated for two genome sequences, i and j, one from each group. Therefore,

$$D = \frac{\sum_{i=1}^{n1}\sum_{j=1}^{n2} D_{i,j}}{n_1 \times n_2}.$$

Analysis of the Three Most Significantly Variable Protein Coding DNA Sequences (CDSs), S, sars3a, and nsp3, Among Palm Civets and Human Patients of the Two Epidemics. The phylogenetic tree of sequences in the four groups (PC03, PC04, HP03E, and HP04) of each gene was first constructed by the neighbor-joining method (8). Given the tree, we used maximum-likelihood analysis (10) for codon substitutions to estimate the number of nonsynonymous and synonymous changes in each branch as well as their rate ratio ω (= dN/dS) (10). The codon-substitution model (11) accounts for the genetic code structure, transition/transversion rate bias, and different base frequencies at each codon position. In the likelihood analysis, we applied the most general model, which implies an independent dN/dS ratio for each branch in the phylogeny (10). An ω value >1 is usually taken as evidence for the signature of positive selection (12).

Statistical Analysis for Estimation of the Neutral Mutation Rate and the Date for the Most Recent Common Ancestor (MRCA) and Construction of Rooted Phylogenetic Tree. The Pamilo–Bianchi–Li model was used to calculate the number of synonymous substitutions per synonymous site, *Ks*, for the concatenated five known major coding sequences (orf1ab, S, E, M, and N) of SARS-CoV, as we did previously (7). Taking GZ02 (7), the reference sequence of the HP03 epidemic, as the outgroup, the *Ks*s were calculated for two PC03 SARS-CoV (SZ16 and SZ3), three PC04 SARS-CoV (PC4-136, PC4-227, PC4-13), and two HP04 sequences (GZ03-01 and GZ03-02), to estimate the neutral mutation rate.

Based on the plot of Fig. 2, the intercept (β_0) of the fitted line is 0.0007806, with the corresponding sampling date 0, which is the end of year 2002. Let T denote the number of days ahead of January 1, 2003, for the MRCA of the PC03 and HP03 groups. Because we used the GZ02 as an outgroup whose sampling date is February 11, 2003 (i.e., 42 days after January 1, 2003), the estimated T will be $T=(\hat{\beta}0/\hat{\beta}1-42)/2=28$(days), which is equivalent to early December 2002.

The *Ks* between SZ16 and SZ3 is 0.001585. Therefore, the estimated date of MRCA for PC03 group is around the end of January 2003. (0.001585/0.000008/2 = 99 days ahead of May 7, 2003, which is the sampling date for SZ16 and SZ3).

The *Ks* between SZ16 and PC4-136, PC4-227, and PC4-13 are 0.003785, 0.003752, and 0.003782, respectively. Therefore, the estimated date of MRCA for PC03 and PC04 is ≈(0.00378/0.000008-244)/2 = 114 days ahead of May 7, 2003, which corresponds to the middle of January 2003.

Based on these estimates, a rooted phylogenetic tree for

EVOLUTION

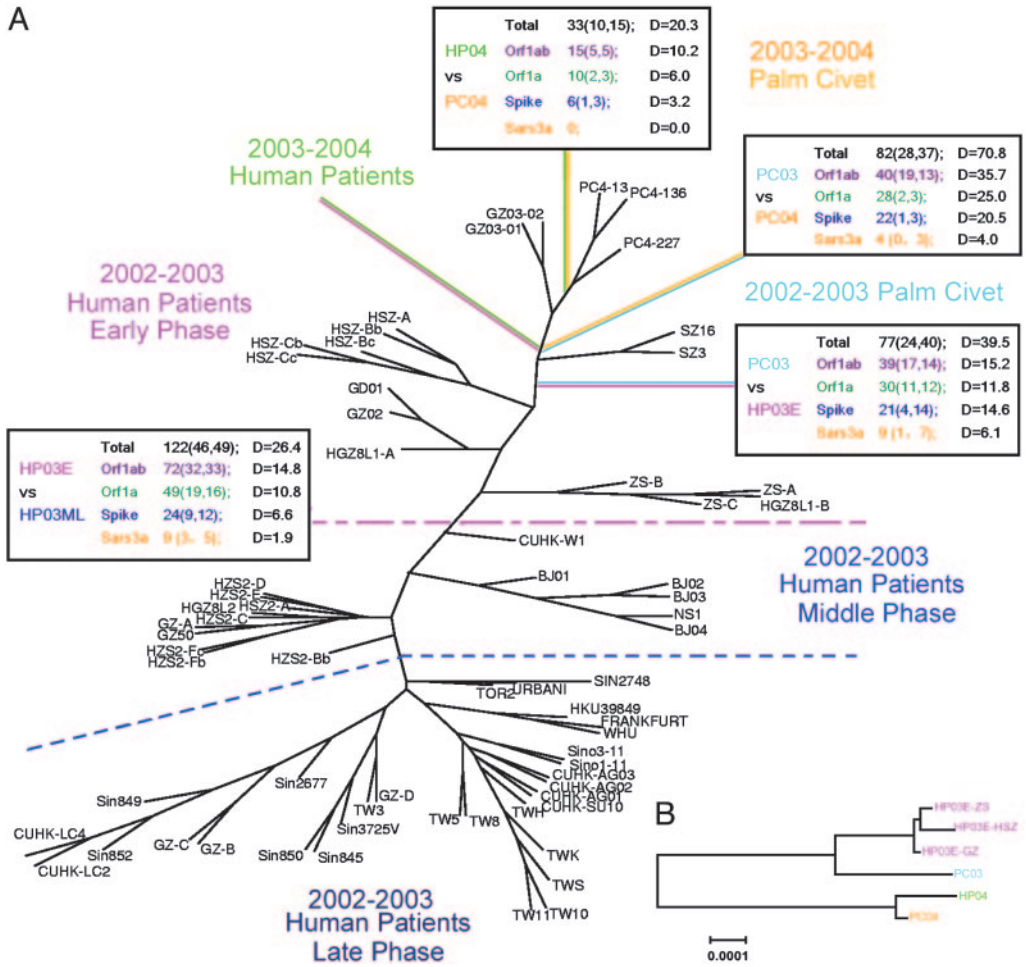

Fig. 1. Genotype clustering of SARS-CoV covering the epidemics from 2002 to 2004. It is illustrated by an unrooted phylogenetic tree constructed with complete SNVs and deletions of 91 sequences from the human patient-derived viruses (HP) and five sequences from the palm civet-derived viruses (PC) (*A*) and a neighbor-joining (N-J) tree for the consensus nucleotide sequences of PC and early individual transmission lineages of HP (*B*). In *A*, the division of the clusters and the corresponding nomenclatures was based on both the hosts of the viruses and the phases of the epidemic (7) (Table 2). The map distance between individual sequences represents the extent of genotypic difference. To highlight the variations between two neighboring clusters, the number of SNVs [total (synonymous, nonsynonymous causing drastic amino acid changes)] occurring among the genomic sequences of both groups and the average number of nucleotide difference *D* between the two sample groups (see *Materials and Methods*) were shown in the boxes. Besides the SNVs of the whole genome (Total), those occurring in ORF1AB (particularly in ORF1A, which is part of Orf1ab), S, and sars3a are listed in the same manner as the total SNVs. These SNVs were present in at least two independent samples of all the sequences used for this analysis. In *B*, consensus nucleotide sequences were derived from each PC and HP data set. For HP03E, consensus nucleotide sequences were individually derived from three primary transmission lineages, based on their direct epidemiological connections and high genomic sequence similarities, and were represented as HP03EGZ (Guangzhou), HP03EHSZ (Shenzhen), and HP03EZS (Zhongshan). These six consensus nucleotide sequences were used to construct the N-J tree (8) in MEGA2 (23), and the Kimura 2-parameter model was assumed. The branch lengths are the estimates of genetic distances.

SARS-CoV isolates from palm civet (PC03 and PC04) and early human patients (HP03E and HP04) is constructed (Fig. 3).

Estimating the Coevolution Coefficients Among SARS-CoV Proteins (Identified and Hypothetical) Based on Amino Acid Substitution Rates. The value of the linear correlation coefficient (r) of the amino acid substitution rates between two proteins of SARS-CoV indicates their level of coevolution (13). We first conducted multiple sequence alignment for each of the SARS proteins (among 72 samples with 21 assigned or predicted protein CDSs without gap in the coding areas) and then used them to build matrices containing the distances between all possible protein pairs. Distances were calculated as the average value of the residue similarities taken from the McLachlan amino acid homology matrix (14). The outcome of this study is listed in Table 3, which is published as supporting information on the PNAS web site.

Results and Discussion

Contact History and Clinical Symptoms of the Four Confirmed SARS Patients (2003–2004) Provide Direct Evidence of Animal-to-Human Infection. The epidemiology information collected by the Guangdong Center for Disease Control and Prevention and the Guangzhou Center for Disease Control and Prevention indicated that between December 16, 2003, and January 8, 2004, a total of four patients were independently hospitalized in the city of Guangzhou, Guangdong Province, China, with flu-like syndromes later diagnosed as confirmed SARS cases (see *Materials and Methods*). Although none of these patients had a contact history with the other previously documented SARS cases, they all had direct or indirect contact history with wild animals in geographically restricted areas. The second patient worked in a local restaurant, TDLR, and the fourth patient dined in the same restaurant where palm civet and other exotic dishes were served, whereas

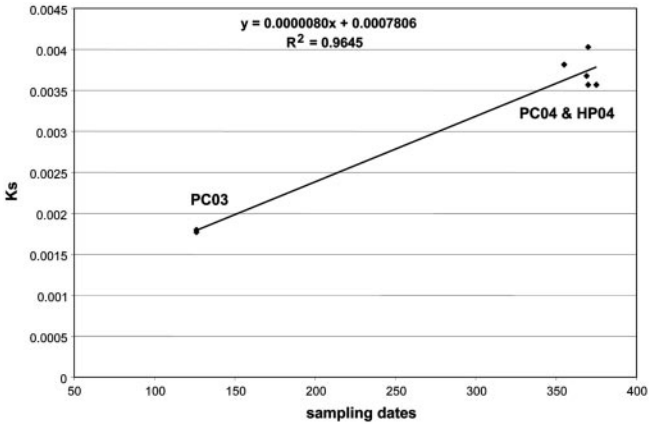

Fig. 2. A plot of the number of synonymous substitutions per synonymous site, *Ks*, for the concatenated coding sequences vs. the sampling dates. The *Ks* calculation and samples used are described in *Materials and Methods*. The sampling dates are measured as the number of days away from Jan. 1, 2003. The slope (β_1) of the fitted line from the linear regression model gives the estimation of the neutral rate, 8.00×10^{-6} per site per day.

the third patient dined in a neighboring restaurant, SJR. These restaurants are located near two major hospitals in Guangzhou where many SARS patients were treated in the previous epidemic, and the first patient, the only patient with no contact with TDLR or SJR, visited one of the hospitals in February 2003. This index patient also contacted house rats in his apartment a few days before disease onset. It is important to emphasize that, unlike most SARS patients during the 2002–2003 epidemic, these four new patients clinically presented very mild symptoms, and neither of them had close contacts who were infected (15).

Genomic Sequences of SARS-CoV from both the Human Patients and the Market Palm Civet of the 2003–2004 Outbreak Are Almost Identical. Among the specimens collected during the 2003–2004 outbreak in Guangzhou (see *Materials and Methods*), we were able to sequence nearly completely the SARS-CoV viral genome from the first two of the four human patients, the two palm civets of the Guangzhou food market, and one sample from the palm civet cage at the restaurant TDLR. These genomic sequences were characterized and phylogenetically analyzed by comparison with 89 human SARS-CoV and two SARS-like-CoV sequences from the Himalayan palm civets available at GenBank as of the end of September 2004, using the *in silico* analysis methodology adopted previously. A total of 202 SNVs with multiple occurrences were identified, among which 200 were in the CDSs. Among the 128 nonsynonymous mutations, 89 led to a predicted radical amino acid changes (Table 2 and Fig. 1*A*).

Besides the individual sequence-based analysis, we further analyzed the data based on comparisons between groups of samples. Although both the nomenclature for sample groups and the analytical methods are described in *Materials and Methods*, an abbreviated nomenclature will be redefined in the text on first occurrence.

All of the HP04 and PC04 (human patient and palm civet, 2003–2004) SARS-CoV isolates retained the 29-nt segment marker in orf8a as in the viruses of PC03 (palm civet 2002–2003) and the Guangzhou primary transmission lineages of HP03E (human patient 2002–2003, early phase). The genomes of the SARS-CoV from HP04 were almost identical to those of the SARS-CoV-like viruses from PC04 (Fig. 1*A*). There were 33 SNVs detected among the viral genomic sequences from PC04 and HP04, which accounts for 0.11% of the viral genome. The average total number of nucleotide differences in the whole genome between the two groups is 20.33. In contrast, between genomic sequences of HP03E and PC03, the average number of nucleotide differences is 39.5, and a total of 77 SNVs was detected, accounting for nearly 0.26% of the viral genome (Fig. 1*A*). Although 17 of the 202 SNVs were polymorphic in the palm civets only, no signature SNVs are shared by all members of palm civet isolates distinguishable from all members of the human isolates (Table 2).

The phylogenetic relationship among different transmission lineages of the early samples of SARS-CoV sequences were also analyzed on the basis of consensus of each epidemiological phase/primary transmission lineage (7) (Fig. 1*B*). Tajima's relative rate test was performed based on the phylogenetic analysis of consensus nucleotide sequence of PC04 as the root, the χ^2 was 39.72 with one degree of freedom ($P = 0.000$), i.e., the distance between PC04 and PC03 is significantly larger than that between PC04 and HP04. Thus, structurally, there is little difference to distinguish the genomic sequences of the SARS-CoV and SARS-CoV-like viruses and functionally, concerning the animal contact history of the current patients, it is likely that the same virus can infect both palm civet and human.

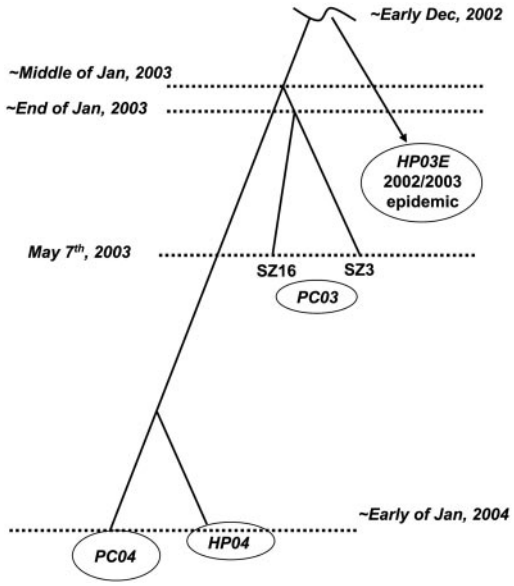

Fig. 3. A rooted phylogenetic tree for SARS-CoV isolates from palm civet (PC03 and PC04) and early human patients (HP03E and HP04) based on MRCA estimations. All data are described in *Materials and Methods* except that for HP03E, which was from previous work (7). The branch length is proportional to the time interval.

The Estimation of the Neutral Mutation Rate and the Date for the MRCA Illustrated the Evolving SARS-CoV in both Palm Civet and Human. We used the concatenated five major CDSs (orf1ab, S, E, M, and N) of SARS-CoV from PC03, PC04, and HP04 to estimate the neutral mutation rate during SARS-CoV transmission in palm civets and HP04 (Fig. 2). The total length of the concatenated sequence accounts for 91.25% of the whole genome. The estimate turned out to be $\approx 8.00 \times 10^{-6}$ nt^{-1}·day^{-1}, which is almost the same as that estimated in our previous work based on 10 samples in HP03 group (8.26×10^{-6} nt^{-1}·day^{-1}) (7). These two independent estimates are almost identical, and thus it supports well the previous conclusion. On the other hand, because samples of PC03, PC04, and HP04 were collected 1 yr apart from different geographic locations, this new estimate should be more accurate. This relatively long-term evolutionary analysis once again strongly suggested that SARS-CoV evolves at a relatively constant neutral rate both in human and palm civet. Furthermore, the date estimates of the MRCAs for PC03, HP03E, PC04, and HP04 were obtained (see

EVOLUTION

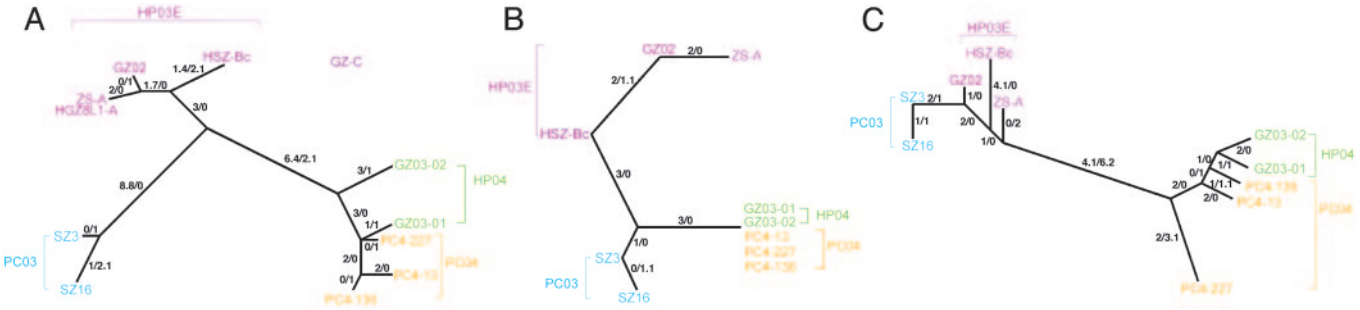

Fig. 4. Phylogeny of the most variable genes, S (*A*), sars3a (*B*), and nsp3 (*C*) in the SARS-CoV samples from the early cases of the epidemic 2002–2003 and the new cases of the 2003–2004 outbreak. Samples of the early cases of the 2002–2003 epidemic were selected based on two criteria: the completeness of the sequences and their representativity in each of the epidemiology lineages. The two numbers shown along each branch are the maximum-likelihood estimates of the numbers of synonymous and nonsynonymous substitutions for each entire gene along the branch (*Materials and Methods*). In each tree, a different dN/dS ratio is assumed for each branch. The branch length is proportional to the total number of estimated synonymous and nonsynonymous substitutions occurring in that branch.

Materials and Methods), which enabled us to derive a rooted phylogenetic tree (Fig. 3). It clearly indicated that PC03 and PC04 are not in the same primary transmission lineage. The viral transmission from animal to human occurred independently in these two instances. PC03 and PC04 further diverged around January 2003, i.e., after the PC03 and HP03 groups bifurcated. Given the relatively long divergence time since their MRCA, it is no surprise to observe an average of 70.83 total nucleotide difference between the viral genome PC03 and PC04 (Fig. 1*A*), higher than that observed for PC03 and HP03E (see above). Because a higher viral load of PC04 was suggested in palm civets from Guangzhou food market during the 2003–2004 outbreak based on the fact that it was much easier to obtain SARS-CoV samples for genomic sequencing than that during the 2002–2003 epidemic (laboratory experience; C.-C.T., H.X., and J.-D.C.), PC04 might have evolved to be more virulent in or better adapted to palm civet. This further demonstrated that SARS is a zoonotic disease from still-unknown origin that has been evolving not only in human but also in palm civet hosts.

The Three Most Significantly Variable Protein CDSs, S, sars3a, and nsp3, Evolved Differently Among Palm Civets and Human Patients of the Two Epidemics. The phylogeny relationship among palm civets and human patients of the two epidemics was further analyzed by using the maximum-likelihood method (10) based on the three most significantly variable CDSs, S, sars3a, and nsp3 (Fig. 4). In the *S* gene (Fig. 4*A*), from the ancestor node of PC03 to the node of PC04, the ratio of nonsynonymous/synonymous substitution numbers (A/S) is 18.2/2.1, i.e., $\omega = 2.68$ (ω = dN/dS: ratio of nonsynonymous and synonymous rates), indicating a positive selection pressure during animal-to-animal transmission. Furthermore, the ancestor nodes of PC04 and HP04 in the *S* gene were the same, indicating that unlike during the 2002/2003 epidemic, HP04 viruses did not have a chance to diverge for enough time, although in the patient GZ03-02, they already accumulated some amino acid changes (A/S = 6/1). In contrast, the A/S from the ancestor node of PC03 to the node of HP03E in *S* gene is 11.8/0, which corresponds to $\omega = \infty$ (no synonymous variations). This is consistent with our previous conclusion that, during the virus transmission from palm civet to human, the *S* gene experienced strong positive selection and improvement to adapt to its human host. Within the HP03E, in most branches, we observed a very high A/S, again suggesting that the *S* gene was still evolving, having not yet reached its maximum adaptation to human.

It has been shown that the sars3a CDS encodes a minor structural protein associated with the S protein on the surface of the SARS-CoV viral envelope (16). Interestingly, the sars3a CDS evolved in synergy with the S protein (Table 3). Therefore, it is no surprise that it evolved adaptively, as did the *S* gene, as a trifurcating tree for the four epidemic groups (Fig. 4*B*). The A/S is 4/0 between PC03 and HP03E, 4/0 between PC03 and PC04 (HP04), and 6/0 between HP03E and HP04. In contrast, there is no single variation among palm civets and human beings of the current epidemic. Although the coevolving process between S and sars3a is likely due to the need of maintaining their necessary interaction, amino acid changes in the sars3a protein might also be critical, as are those in the S protein, to modulate the host switch of SARS-CoV.

The phylogenetic tree of nsp3 is largely different from that of S or sars3a (Fig. 4*C*). The PC03 is very close to HP03E but relatively more divergent from those of new cases. This suggests that nsp3 may be under different evolutionary pressure from that for the *S* and *sars3a* genes. In the lineage connecting the ancestor node of HP03E and HP04 (or PC04), the A/S is only 4.1/6.2 ($\omega = 0.227$), which does not show any positive selection signature. It is worth pointing out that in the new cases, there is one mutation at nucleotide 6295 leading to a stop codon in the nsp3 CDS of the orf1a. Considering the unique alterations of nsp3 CDS structure in SARS-CoV compared with other CoVs (17), we propose this special mutation might account for the mild clinical symptoms and apparent weak infectivity of this episode.

Major Genetic Variations in the *S* Gene Seem Essential for the Transition from Animal-to-Human Transmission to Human-to-Human Transmission. The S protein is responsible for binding to the angiotensin-converting enzyme 2 (ACE2) receptor (18) and thus is the fastest-evolving protein of SARS-CoV in the epidemic from animal to human. Besides the *S* gene sequences available from whole-genome data (Table 2, except for GZ-D of the 2002–2004 epidemic, which was deleted due to the incompleteness of the sequence), we were able to add more *S* gene sequences for alignment analysis, one from a human sample (GZ03-03) and seven from palm civet samples of the Guangzhou food market, all for the 2003–2004 epidemic. Two sequences, SZ1 and SZ13, from palm civet samples of the 2002–2003 epidemic publicly available were also included in the analysis. Because the 3D structure of the S protein was successfully simulated (Protein Data Bank ID code 1T7G) (19), it was used for a better understanding of the molecular mechanism driving the mutations of the *S* gene over the course of the epidemic. Table 4, which is published as supporting information on the PNAS web site, lists all 49 SNVs observed in >1 of the 103 S CDS sequences, i.e., two more SNVs observed than that using whole-genome sequences (Table 2), because more sequences were added for the analysis. One of them (nucleotide 22220) causes synonymous variation in the amino acid residue 243D, which is predicted to be partially exposed at the top of the S1 domain,

although the other (nucleotide 23163) causes a nonsynonymous variation of amino acid residue 558F/I, which is predicted to be exposed at the side of the S1 domain but not in the predicted receptor-binding region (20).

Of the 17 SNVs observed in animals, 10 are located in the *S* gene. Among them, seven were observed in the current epidemic, one in the previous epidemic, and two in both. With more *S* gene sequences from samples of the third patient of the current epidemic and of seven palm civets from the Guangzhou food market added for analysis, no further changes were found in the SNV patterns.

Although mutations are dispersed over the whole protein, the majority of the mutations are located in the S1 domain (31 of 48 total SNVs), particularly in the region (residues 318–510) predicted to constitute the ACE2 receptor-binding site (20), 11 SNVs corresponding to 10 amino acid residues. Among them, except for two synonymous variations, seven of the nine nonsynonymous mutations may cause radical amino acid changes. Two of them (nucleotides 22422 and 22549) occurred in HP03 only, whereas the remaining five fell into three categories. First, mutations at the second and third nucleotides (22927 and 22928) of codon 479 may cause changes corresponding to three different amino acid residues (K, R, or N). Although all of these codons were found in the palm civet samples, only the aat codon for N was found in all of the human samples as well as some 2003–2004 palm civet samples (PC04). Second, the c→t switch of nucleotide 22570 causing the S→F mutation of codon 360 distinguishes the virus of the 2002–2003 epidemic (HP03) from all other viruses isolated from palm civet (PC03 and PC04) as well as human patients of the 2003–2004 outbreak (HP04). Third, the g→a switch of nucleotide 22930 causing the G→D mutation of codon 480 distinguishes the virus of 2002–2003 (PC03 and HP03) from those of 2003–2004 (PC04 and HP04), regardless of the sources.

Outside of the predicted receptor-binding peptide, we observed another two-substitution codon (19) 609 (nucleotides 23316 and 23317), which is predicted to be buried at the interface of the S1 and S2 domains (19). This tta→gca switch causing an L→A mutation is one of a few nonsynonymous mutations that nearly distinguishes the virus of 2002–2003 from those of 2003–2004, disregarding either their human or animal sources.

Given such low variations in the SARS-CoV genome among all available samples, the probability of having a multiple substitution codon is almost zero, especially for the two-substitution codon involving two nonsynonymous changes, codons 479 and 609 for the S gene. The latter event is the more remarkable, because it also goes in the direction of G + C enrichment, a feature usually extremely rare in viruses, for metabolic reasons (21). Thus, it is another good representative site showing the signature of positive selection (22).

The unfortunate recurrence of SARS at the end of the year 2003 gave us the opportunity to witness the variation/adaptation behavior of the etiological agent of the disease. The new SARS-CoV derived not from the preceding episode but very likely from a common ancestor, which does not harbor the 29-nt deletion that marks most of the virulent forms of SARS-CoV for the 2002–2003 epidemic. The fates of the virus inside the human host and in palm civets are similar, i.e., the virus has not yet adapted to its new host, making it evolve fast (and possibly into highly contagious and/or virulent forms), and in general the infection is mild. Therefore, humans working with wild animals are often seropositive for the SARS-CoV without noticeable symptoms (5). All of this points to a common source of disease lingering in the environment, presumably adapted to its nature host that can come in contact with humans and/or animals. It may have a fairly high probability of mutating under favorable conditions to a form causing SARS in humans. This situation is expected to yield an unusual epidemic pattern, because a proportion of humans may have been immunized against an innocuous form of the virus, so that distribution of the disease, when it happens, is expected to be highly uneven. This should prompt support for more research on the discovery of CoVs in animals, in particular in the Guangdong region.

We are grateful for the critical technical assistance supplied by Yan Sheng, Yi Chen, Zheng Ruan, Guo-Wen Peng, Ai-Ping Deng, Ji-Ya Dai, Hao-Jie Zhong, Xin Zhang, Li-Mei Diao, and Yan-Hua Ao. We sincerely thank Hong-Wei Gao for providing the necessary laboratory facilities and Zhao-An Xin for providing coordination support. We thank InforSense (San Diego) for providing KDE BioScience software for data analysis. We appreciate the strong support of the Ministry of Science and Technology and the Ministry of Agriculture of the Central Government of China and the governments of Guangdong Province and Shanghai Municipality. This work was supported by State High Technology Development Program Grant 2003AA208407, State Key Program for Basic Research Grants 2003CB514101 and 2003CB715904 (to P.H.), the Guangdong Provincial Program on SARS Prevention and Treatment (Project No. 2003FD02-06), and European Commission Grant EPISARS (no. 511063). H.-D.S., G.-W.Z., F.C., C.-M.P., and S.-J.C. are partly supported by a SARS Research Grant from the Bank of BNP PARIBAS. P.H., Z.-M.G., H.-Y.P., W.-Z.H., and Y.-X.L. are partly supported by the Shanghai Commission of Science and Technology. H.T. is supported by a National Institutes of Health grant (to C-.I.W.).

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. Chinese SARS Molecular Epidemiology Consortium (2004) *Science* **303,** 1666–1669.
8. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
9. Tajima, F. (1993) *Genetics* **135,** 599–607.
10. Yang, Z. (1998) *Mol. Biol. Evol.* **15,** 568–573.
11. Goldman, N. & Yang, Z. (1994) *Mol. Biol. Evol.* **11,** 725–736.
12. Li, W. H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
13. Pazos, F. & Valencia, A. (2001) *Protein Eng.* **14,** 609–614.
14. McLachlan, A. D. (1971) *J. Mol. Biol.* **61,** 409–424.
15. Liang, G., Chen, Q., Xu, J., Liu, Y., Lim, W., Peiris, J. S., Anderson, L. J., Ruan, L., Li, H., Kan, B., *et al*. (2004) *Emerg. Infect. Dis.* **10,** 1774–1781.
16. Zeng, R., Yang, R. F., Shi, M. D., Jiang, M. R., Xie, Y. H., Ruan, H. Q., Jiang, X. S., Shi, L., Zhou, H., Zhang, L., *et al*. (2004) *J. Mol. Biol.* **341,** 271–279.
17. Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L. M., Guan, Y., Rozanov, M., Spaan, W. J. M. & Gorbalenya, A. E. (2003) *J. Mol. Biol.* **331,** 991–1004.
18. Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., Somasundaran, M., Sullivan, J. L., Luzuriaga, K., Greenough, T. C., *et al*. (2003) *Nature* **426,** 450–454.
19. Bernini, A., Spiga, O., Ciutti, A., Chiellini, S., Bracci, L., Yan, X., Zheng, B., Huang, J., He, M-L., Song, H-D., *et al*. (2004) *Biochem. Biophys. Res. Commun.* **325**, 1210–1214.
20. Wong, S. K., Li, W., Moore, M. J., Choe, H. & Farzan, M. (2004) *J. Biol. Chem.* **279,** 3197–3201.
21. Rocha, E. P. & Danchin, A. (2002) *Trends Genet.* **18,** 291–294.
22. Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S. & Kondrashov, A. S. (2004) *Nature* **429,** 558–562.
23. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics* **17,** 1244–1245.