Journal of General Virology (2006), **87**, 3355–3359 **DOI** 10.1099/vir.0.82220-0

Short Communication

Full-length genome sequences of two SARS-like coronaviruses in horseshoe bats and genetic variation analysis

Wuze Ren,[1,2] Wendong Li,[2,3] Meng Yu,[4] Pei Hao,[5] Yuan Zhang,[6] Peng Zhou,[1] Shuyi Zhang,[3] Guoping Zhao,[5] Yang Zhong,[6] Shengyue Wang,[5,6] Lin-Fa Wang[4] and Zhengli Shi[1]

[1]State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences (CAS), Wuhan, Hubei 430071, China

[2]Graduate School of CAS, Beijing 100039, China

[3]Institute of Zoology, CAS, Beijing 100080, China

[4]CSIRO Livestock Industries, Australian Animal Health Laboratory, Geelong, VIC 3220, Australia

[5]Shanghai Center for Bioinformation Technology, Shanghai 200235, China

[6]School of Life Sciences, Fudan University, Shanghai 200433, China

Correspondence
Zhengli Shi
zlshi@wh.iov.cn

Received 20 May 2006
Accepted 17 July 2006

Bats were recently identified as natural reservoirs of SARS-like coronavirus (SL-CoV) or SARS coronavirus-like virus. These viruses, together with SARS coronaviruses (SARS-CoV) isolated from human and palm civet, form a distinctive cluster within the group 2 coronaviruses of the genus *Coronavirus*, tentatively named group 2b (G2b). In this study, complete genome sequences of two additional group 2b coronaviruses (G2b-CoVs) were determined from horseshoe bat *Rhinolophus ferrumequinum* (G2b-CoV Rf1) and *Rhinolophus macrotis* (G2b-CoV Rm1). The bat G2b-CoV isolates have an identical genome organization and share an overall genome sequence identity of 88–92 % among themselves and between them and the human/civet isolates. The most variable regions are located in the genes encoding nsp3, ORF3a, spike protein and ORF8 when bat and human/civet G2b-CoV isolates are compared. Genetic analysis demonstrated that a diverse G2b-CoV population exists in the bat habitat and has evolved from a common ancestor of SARS-CoV.

Severe acute respiratory syndrome (SARS) is one of the most important emerging zoonotic diseases in the 21st century. A novel coronavirus, the SARS coronavirus (SARS-CoV), was identified as the aetiological agent of SARS (Fouchier *et al.*, 2003; Ksiazek *et al.*, 2003; Marra *et al.*, 2003; Peiris *et al.*, 2003; Rota *et al.*, 2003; Zhong *et al.*, 2003). The rapid identification of highly similar viruses in masked palm civet and racoon dog in the live-animal markets provided strong evidence of an animal origin of SARS-CoV and played an important role in the prevention of further outbreaks (Guan *et al.*, 2003). However, subsequent epidemiological studies on civets from market, farm and wild populations demonstrated that there was no widespread infection among wild or farmed civets, implying that wild animal(s) other than civets may serve as the natural reservoir(s) of SARS-CoV (Tu *et al.*, 2004; Kan *et al.*, 2005; Poon *et al.*, 2005).

Recently, we and another independent group have simultaneously reported the detection of SARS-like coronavirus (SL-CoV) or SARS coronavirus-like virus in different horseshoe bat species in the genus *Rhinolophus*, providing evidence that suggests bats as a natural reservoir of this group of viruses (Lau *et al.*, 2005; Li *et al.*, 2005b). Due to the close genetic and antigenic relationship of SARS-CoVs and SL-CoVs, this group of viruses has been named the SARS cluster coronaviruses or group 2b coronavirus (G2b-CoV) in differentiation from other group 2 coronaviruses in the genus *Coronavirus* (Gorbalenya *et al.*, 2004; Lau *et al.*, 2005; Li *et al.*, 2005b; Woo *et al.*, 2006). Molecular and serological studies indicated that at least five different horseshoe bat species in mainland China and Hong Kong harbour G2b-CoVs. They include *Rhinolophus sinicus*, *Rhinolophus pearsonii*, *Rhinolophus ferrumequinum*, *Rhinolophus macrotis* and *Rhinolophus pusillus*. Full-length genome sequences were published for three isolates, one from *R. pearsonii* (Rp3) and two from *R. sinicus* (HKU3-1 and HKU3-2). The sequences of the HKU3-1 and HKU3-2 genomes were almost identical and they probably represented different isolates of the same genotype. The Rp3 and HKU3 isolates share an overall nucleotide sequence identity of 92 and 88 % to the outbreak SARS-CoVs isolated from civets and humans, respectively.

In this paper, we describe the characterization of full-length genome sequences for two additional G2b-CoV isolates, Rf1 from *R. ferrumequinum* and Rm1 from *R. macrotis*, and present genome-comparison data of all known G2b-CoV genome types to demonstrate further the great genetic diversity among this group of novel coronaviruses and to identify potential genetic features that might be associated with host specificity, transmission in non-bat species and virus virulence. It should be noted that there seems to be a large number of different coronaviruses present in different bat species. At least seven other novel bat coronaviruses have been discovered among bat populations in Hong Kong (Poon *et al.*, 2005; Woo *et al.*, 2006). As these coronaviruses are not related to the G2b-CoVs, the focus of this study, and there were no full-length genome sequences available for them, they are not included in the current comparative study.

The collection, processing and storage of bat samples, as well as the determination of the full-length genome sequence, were conducted as described previously (Li *et al.*, 2005b). Sequence alignment was performed by using CLUSTAL_X version 1.83 (Thompson *et al.*, 1997) and corrected manually. Phylogenetic trees based on nucleotide sequence were constructed by using the neighbour-joining (NJ) method with a bootstrap of 1000 replicates implemented in MEGA version 3.1 (Kumar *et al.*, 2004). The mean non-synonymous substitution rate (K_a), synonymous substitution rate (K_s) and the ratio of K_a/K_s for four protein-coding sequences (ORF1a, ORF1b, ORF3a and S) were calculated by K-Estimator 6 (Comeron, 1999). The Kimura two-parameter substitution model was used and other parameters were as default settings in MEGA 3.1. Fisher's exact test of positive-selection analysis implemented in MEGA 3.1 and the CODEML program implemented in the PAML package (Yang & Swanson, 2002) were also used to detect potential positive selection for genes P1a, P1b, ORF3a and S of bat and human/civet G2b-CoV.

The full-length genomes of Rf1 and Rm1 are 29 690 and 29 733 nt [excluding the poly(A) tail], respectively. The genome organization and the predicted gene products of both viruses are similar to those of other characterized G2b-CoVs (Fig. 1; Table 1). However, Rf1 seems to have a unique feature that may represent an evolutionary intermediate between bat G2b-CoVs and human/civet G2b-CoVs. As shown in Fig. 1, there is an ORF3b of 154 aa (overlapping ORF3a) in the human/civet isolates that is absent from most bat G2b-CoVs. In the corresponding region in the Rf1 genome, there were two ORFs, of 113 and 32 aa. The four bat G2b-CoV genomes share a sequence identity of 88–90 % among themselves. Similar sequence identity, 88–92 %, exists between bat and human/civet isolates. Nucleotide variations are scattered along the whole genome, but the most variable regions were located in the genes encoding non-structural protein 3 (nsp3), S (the N-terminal S1 domain), ORF3a and ORF8. This is also true for deletion/insertion mutations in nsp3, S and ORF8. For nsp3 genes, the deletion/insertion mutations seem to be concentrated in the region encoding a unique domain originally identified by Snijder *et al.* (2003) that is present

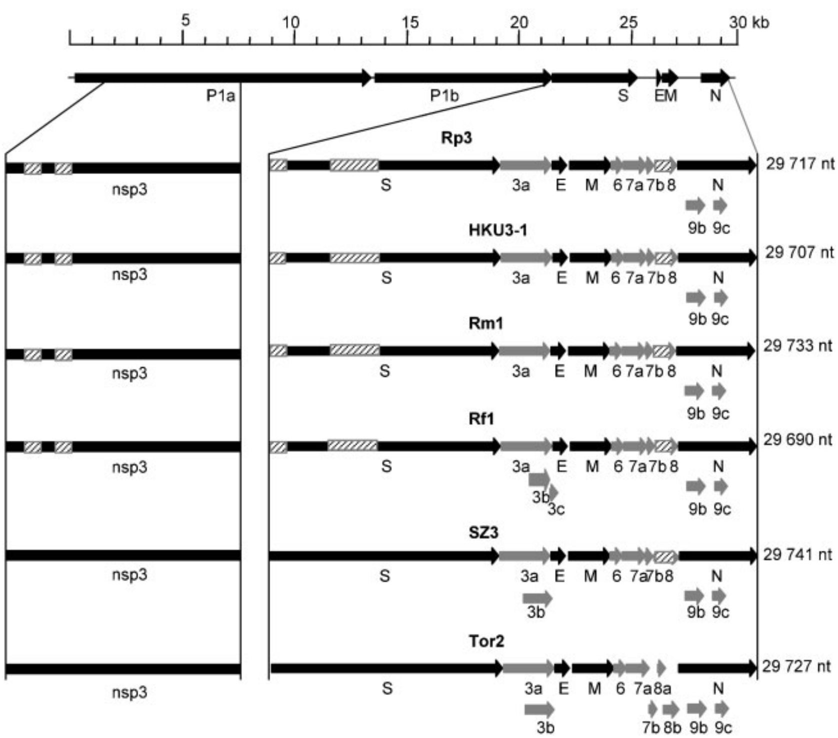

Fig. 1. Genome organization of isolates Rf1 and Rm1 and comparison with other G2b-CoV genomes. The nomenclature of genes and ORFs follows the recommendation by Spaan *et al.* (2005) and is similar to those used by others (Chinese SARS Molecular Epidemiology Consortium, 2004; Lau *et al.*, 2005; Snijder *et al.*, 2003). The genes present in all coronaviruses are shown in dark-shaded arrows and the G2b-CoV-specific ORFs in light-shaded arrows. The most variable regions are marked with hatched boxes. The drawing is not proportional for all regions of the genomes shown.

Table 1. Comparison of deduced gene-product size and protein sequence identity of different G2b-CoVs

NP, Not present; NA, not applicable.

Gene/ORF	Gene product size (aa)						Amino acid sequence identity with Tor2/SZ3 (%)*			
	Tor2	SZ3	Rf1	Rp3	Rm1	HKU3-1	Rf1	Rp3	Rm1	HKU3-1
P1a	4382	4382	4377	4380	4388	4376	94	96	93	94
P1b	2628	2628	2628	2628	2628	2628	98	99	98	98
S	1255	1255	1241	1241	1241	1242	76	78	78	78
(S1)†	680	680	666	666	666	667	63	63	64	64
(S2)†	575	575	575	575	575	575	92	96	96	93
ORF3a	274	274	274	274	274	274	86	83	83	81
ORF3b	154	154	113	NP	NP	NP	89	NA	NA	NA
ORF3c	NP	NP	32	NP	NP	NP	NA	NA	NA	NA
E	76	76	76	76	76	76	96	100	98	100
M	221	221	221	221	221	221	97	97	97	98
ORF6	63	63	63	63	63	63	93	92	92	93
ORF7a	122	122	122	122	122	122	91	95	93	94
ORF7b	44	44	44	44	44	44	90	93	93	93
ORF8a	39	NP	NP	NP	NP	NP	NA	NA	NA	NA
ORF8b	84	NP	NP	NP	NP	NP	NA	NA	NA	NA
ORF8	NP	122	122	121	121	121	80	35	35	33
N	422	422	421	421	420	421	95	97	97	96
ORF9b	98	98	96	97	97	97	81	85	90	87
ORF9c	70	70	70	70	70	70	80	91	91	88

*Tor2 was used for all similarity calculations with the exception of ORF8, which is absent in Tor2. The SZ3 ORF8 was used instead.
†S1, the N-terminal domain of the coronavirus S protein responsible for receptor binding; S2, the C-terminal domain responsible for membrane fusion.

in SARS-CoV, but absent in other coronaviruses (Fig. 1). The sequence identity of the S genes among four bat G2b-CoVs is 89–95 %. The sequence identity drops to 76–78 % between S genes of bat G2b-CoVs and human/civet G2b-CoVs, and even lower (63–64 %) for the putative S1 domain. There are one 6 aa insertion and three deletions of various lengths in the S1 domains of bat isolates in comparison to those of the human/civet isolates (Lau *et al.*, 2005; Li *et al.*, 2005b). Two deletion sites (5 and 12 aa, respectively) are located in the receptor-binding domain (RBD) region, and overlap with the so-called receptor-binding motif (RBM; aa 424–494 of the Tor2 S protein), which is identified as being critical for receptor binding (Li *et al.*, 2005a). Human G2b-CoV isolates are known to use angiotensin-converting enzyme-2 (ACE2) as the main receptor for cell entry (Li *et al.*, 2003). It is not known whether the bat G2b-CoVs are able to use the bat ACE2 homologue as receptor or whether they use an alternative receptor molecule for cell entry, as speculated by Li *et al.* (2006).

Phylogenetic trees based on the full-length genome sequences and individual genes of selected human and civet G2b-CoVs and four bat G2b-CoVs are shown in Fig. 2. Depending on the sequences used, several different phylogenetic patterns were observed. When the full-length genome sequences were used, bat isolate Rp3 grouped closer to the human/civet isolates than to other bat isolates, with a high bootstrap support (Fig. 2a). Similar observations were also made for trees based on P1a and P1b gene sequences (data not shown). When the full-length S genes were analysed, all bat G2b-CoVs clustered together and were separated from human/civet isolates (Fig. 2b). A third pattern was observed for trees based on ORF3a, M, ORF6 and ORF8 sequences (the representative tree of ORF3a is shown in Fig. 2c). In these trees, the Rf1 sequence does not group with other bat isolates; instead, it sits between the bat isolates and human/civet isolates, and for ORF8, the Rf1 sequence is related much more closely to human/civet isolates than to other bat isolates. Poorly resolved trees were observed for genes E, ORF7a, ORF7b and N among four different bat isolates (a representative tree of ORF7a is shown in Fig. 2d). These incongruent phylogenetic trees seem to suggest potential recombination events among these G2b-CoVs. However, when these sequences were analysed by using a recombination-detection program (RDP2; Martin *et al.*, 2005), we were unable to obtain conclusive evidence for any definitive recombination event (data not shown). We aim to collect more G2b-CoVs and related corona-viruses of bat to continue the search for recombination points in the G2b-CoV genomes.

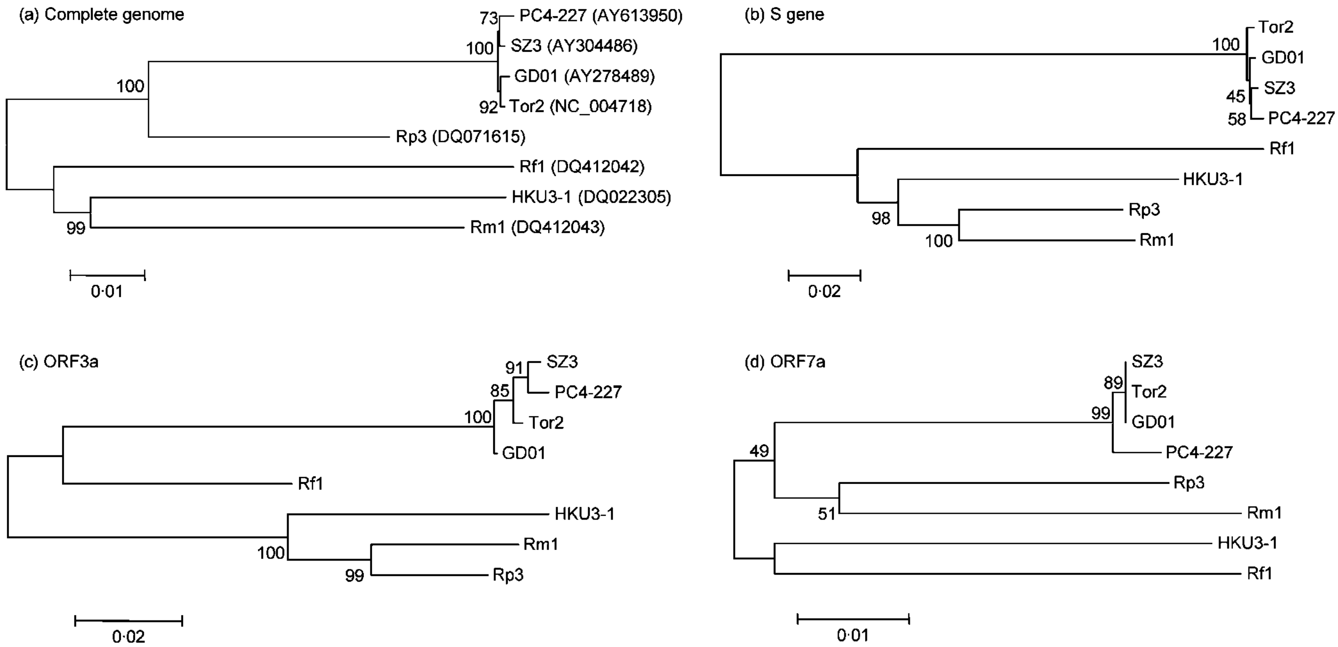

Fig. 2. Phylogenetic trees based on sequences of full-length genomes and different genes. Sequences used in this study are as follows: Tor2, human isolate from the late phase of the 2002–2003 outbreak; GD01, human isolate from the early phase of the 2002–2003 outbreak; SZ3, civet isolate from 2003; PC4-227, civet isolate from 2004; HKU3-1, bat isolate from *R. sinicus*; Rp3, bat isolate from *R. pearsonii*; Rf1, bat isolate from *R. ferrumequinum*; Rm1, bat isolate from *R. macrotis*. The phylogenetic trees were constructed by using the NJ algorithm in the MEGA 3.1 software with a bootstrap of 1000 replicates. The representative sequences used for different tree patterns are as follows: full-length genome sequence (a), S gene (b), ORF3a (c) and ORF7a (d). The GenBank accession number for each full-length genome sequence is given next to the isolate name in (a). Genetic variation scales are indicated for each tree and different genetic scales are used for different trees.

The synonymous and non-synonymous substitution rates (K_a and K_s, respectively) for genes P1a, P1b, ORF3a and S were used to estimate the selection pressure for bat and human/civet G2b-CoVs. The K_a/K_s ratio of these four genes among all bat isolates and between bat and human/civet isolates is <1. By contrast, the K_a/K_s ratios of human/civet isolates from different origins were different. For P1a and P1b, K_a/K_s is <1 among isolates of different origins, except for P1a between civet isolate SZ3 (isolated in 2003) and human isolate Tor2 (from a human patient in the late phase of the 2002–2003 outbreak). However, the K_a/K_s ratios were significantly greater than 1 for S and ORF3a sequences among civet isolates obtained from 2003 (SZ3) and 2004 (PC4-227) and human isolates from early (GD01) and late (Tor2) phases of the outbreak. These results indicate that G2b-CoVs in bats found to date have not experienced a positive-selection pressure and that these viruses have evolved independently for a relatively long time. In contrast, the human/civet isolates have undergone a strong positive selection during the transmission from animal to human (Song *et al.*, 2005), suggesting a recent species-crossing event.

Among the five complete bat isolates sequenced so far, HKU3-1 and HKU-2 were almost identical in genome sequence, which was not unexpected considering that they were isolated from the same species (*R. sinicus*) within a small geographical location in Hong Kong (Lau *et al.*, 2005). For that reason, we considered them to be of the same genome type. We noted that the genome sequence of Rf1 displayed a more distant evolutionary relationship to other bat isolates. Whether these different G2b-CoV genotypes from different bat species are linked to their host evolution needs further investigation when more G2b-CoVs from different bat species become available. Based on the current data, it can be hypothesized that there is a wide spectrum of genetically diverse G2b-CoVs present in their natural reservoir hosts, and viruses with a much closer evolutionary relationship to the SARS outbreak strains from civets and human may be present in different *Rhinolophus* species or other bat species in China or neighbouring countries.

Acknowledgements

This work was funded jointly by State Key Program for Basic Research grant 2005CB523004, State High Technology Development Program grant 2005AA219070 and a special grant for ‘Animal Reservoir of SARS-CoV’ from the Ministry of Science and Technology, People’s Republic of China, a special fund from the president of Chinese Academy of Sciences (no. 1009), the Sixth Framework Program ‘EPISARS’ from the European Commission (no. 51163) and the Australian Biosecurity CRC for Emerging Infectious Diseases (Project

1.007R). Our initial investigation, which led to the discovery of horseshoe bats as the reservoir host of G2B-CoVs, was conducted in collaboration with research activities co-funded by an NIH/NSF 'Ecology of Infectious Diseases' award (no. R01-TW05869) from the John E. Fogarty International Center and the V. Kann Rasmussen Foundation.

References

Chinese SARS Molecular Epidemiology Consortium (2004). Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China. *Science* **303**, 1666–1669.

Comeron, J. M. (1999). K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15**, 763–764.

Fouchier, R. A. M., Kuiken, T., Schutten, M. & 7 other authors (2003). Aetiology: Koch's postulates fulfilled for SARS virus. *Nature* **423**, 240.

Gorbalenya, A. E., Snijder, E. J. & Spaan, W. J. M. (2004). Severe acute respiratory syndrome coronavirus phylogeny: toward consensus. *J Virol* **78**, 7863–7866.

Guan, Y., Zheng, B. J., He, Y. Q. & 15 other authors (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* **302**, 276–278.

Kan, B., Wang, M., Jing, H. & 27 other authors (2005). Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. *J Virol* **79**, 11892–11900.

Ksiazek, T. G., Erdman, D., Goldsmith, C. S. & 23 other authors (2003). A novel coronavirus associated with severe acute respiratory syndrome. *N Engl J Med* **348**, 1953–1966.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.

Lau, S. K. P., Woo, P. C. Y., Li, K. S. M. & 7 other authors (2005). Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci U S A* **102**, 14040–14045.

Li, W., Moore, M. J., Vasilieva, N. & 9 other authors (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* **426**, 450–454.

Li, F., Li, W., Farzan, M. & Harrison, S. C. (2005a). Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864–1868.

Li, W., Shi, Z., Yu, M. & 14 other authors (2005b). Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679.

Li, W., Wong, S.-K., Li, F., Kuhn, J. H., Huang, I.-C., Choe, H. & Farzan, M. (2006). Animal origins of the severe acute respiratory syndrome coronavirus: insight from ACE2-S-protein interactions. *J Virol* **80**, 4211–4219.

Marra, M. A., Jones, S. J. M., Astell, C. R. & 56 other authors (2003). The genome sequence of the SARS-associated coronavirus. *Science* **300**, 1399–1404.

Martin, D. P., Williamson, C. & Posada, D. (2005). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21**, 260–262.

Peiris, J. S. M., Lai, S. T., Poon, L. L. M. & 13 other authors (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319–1325.

Poon, L. L. M., Chu, D. K. W., Chan, K. H. & 9 other authors (2005). Identification of a novel coronavirus in bats. *J Virol* **79**, 2001–2009.

Rota, P. A., Oberste, M. S., Monroe, S. S. & 32 other authors (2003). Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* **300**, 1394–1399.

Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C. & 7 other authors (2003). Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J Mol Biol* **331**, 991–1004.

Song, H.-D., Tu, C.-C., Zhang, G.-W. & 56 other authors (2005). Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc Natl Acad Sci U S A* **102**, 2430–2435.

Spaan, W. J. M., Brian, D., Cavanagh, D. & 8 other authors (2005). Family *Coronaviridae*. In *Virus Taxonomy: Eighth Report of the Committee on Taxonomy of Viruses*, pp. 947–964. Edited by C. M. Fauquet, M. A. Mayo, J. Maniloff, U. Desselberger & C. A. Ball. London: Elsevier Academic Press.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.

Tu, C., Crameri, G., Kong, X. & 11 other authors (2004). Antibodies to SARS coronavirus in civets. *Emerg Infect Dis* **10**, 2244–2248.

Woo, P. C. Y., Lau, S. K. P., Li, K. S. M. & 7 other authors (2006). Molecular diversity of coronaviruses in bats. *Virology* **351**, 180–187.

Yang, Z. & Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* **19**, 49–57.

Zhong, N. S., Zheng, B. J., Li, Y. M. & 13 other authors (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. *Lancet* **362**, 1353–1358.