Molecular evolution of the SARS-coronavirus during the course of the SARS epidemic in China

The Chinese SARS molecular epidemiology consortium

Supporting online materials

Materials and Methods

A. Epidemiological investigations

Official epidemiological records of the Guangdong Center for Disease Control and Prevention (GDCDCP), which represented an aggregate of the regular SARS epidemiology reports submitted by the local Centers for Disease Control and Prevention of individual cities, were reviewed. The contact and clinical histories of all of the early seemingly independent index cases and several key cases (*e.g.* HZS2-F) were reconfirmed either by review of hospital patient records or direct interview with the patients and/or the physicians-in-charge. In particular, eleven index cases from seven cities located in the Pearl River Delta region of Guangdong Province (Fig.1 and fig. S1), which occurred prior to the first superspreader event of a Guangzhou hospital, HZS-2, were investigated in detail.

Majority of the specimens were collected by the virologists of GDCDCP, with the remaining samples collected by the staffs of local hospital or Guangzhou Center for Disease Control and Prevention.

B. Sequencing strategy and procedures

We isolated the viral RNA templates either from the culture supernatants of VeroE6 cells that showed cytopathic effects or directly from patients' specimens of SARS cases (including serum, stool, oropharyngeal swabs, nasal pharyngeal aspirates or autopsy lung tissues). RNA was extracted with the QIAamp viral RNA mini kit (Qiagen, Valencia, CA, USA) or TRIZOL Reagent (GIBCOBRL). The double-strand cDNA was synthesized with the SuperScript II cDNA system (Invitrogen, Carlsbad, CA, USA) or RNA PCR Kit (AMV) Ver.2.1 (Takara, Dalian China). To amplify the genomic sequences of the SARS-CoV, 53 sets of nested primers were designed based on the TOR2 sequence. The nested PCR fragments were directly sequenced in both forward and reverse directions on the ABI-3700 DNA sequencer (Applied Biosystems, Foster City, CA, USA) with 2- to 4-fold redundancy. For GZ02, PCR primers were designed to cover the whole genome in every 1kb interval with 200bp overlap with the adjacent fragment based on the TOR2 sequence. PCR products were sequenced using ABI BigDye Terminator Cycle Sequencing Kit on ABI-377. All of the nucleotide sequence variations of GZ02, which differ from that of the human SARS-CoV sequences available at GenBank as of June 2003, in particular, TOR2 and GZ01 (the sequence of an independent viral isolate from the same patient as GZ02 and currently renamed as GD01) sequences (including the 29-bp segment), were re-sequenced from RNA extractions from the same lung tissue specimen of that patient and the 5' end sequence was completed. The PHRED/PHRAP/CONSED software (University of Washington, Seattle, WA, USA; http://www.phred.org) was used for base calling, assembly, and editing. The assembled genome sequence was checked manually for accuracy and the regions with poor quality were re-sequenced. For data analysis, the nucleotide coordinate of GZ02 was used as a reference.

C. List of GenBank accession numbers for sequences mentioned in the text and SOM:

Sequences generated by this study	**GenBank accession number**	**Sequences previously available**	**GenBank accession number**
GD03T13 (S gene)	AY525636	SZ16 (palm civet)	AY304488
		SZ3 (palm civet)	AY304486
GZ02	AY390556	GD01 (GZ01)	AY278489
HGZ8L1-A	AY394981		
HSZ-A	AY394984		
HSZ-B (b, c)	AY394985, AY394994		
HSZ-C (b, c)	AY394986, AY394995		
ZS-A	AY394997	gz43 (S gene)	AY304490
ZS-B	AY394996	gz60 (S gene)	AY304491
ZS-C	AY395003		
GZ-A	AY394977		
JMD	AY394988		
HGZ8L1-B	AY394982		
HZS2-A	AY394983	CUHK-W1	AY278554
HZS2-Bb	AY395004	BJ04	AY279354
HZS2-C	AY394992	BJ01	AY278488
HZS2-D	AY394989	BJ02	AY278487
HZS2-E	AY394990	BJ03	AY278490
HGZ8L-2	AY394993		
HZS2-Fc	AY394991		
HZS2-Fb	AY394987		
CUHK-LC1	AY394998		
GZ-B	AY394978	TOR2	AY274119
GZ-C	AY394979	ZJ01	AY297028
GZ-D	AY394980		
CUHK-LC2	AY394999	CUHK-AG01	AY345986
CUHK-LC3	AY395000	CUHK-AG02	AY345987
CUHK-LC4	AY395001		
CUHK-LC5	AY395002		

Supporting Online Text S1

Statistical analysis for (A) the estimation of the neutral mutation rate and the date for the most recent common ancestor (MRCA); (B) calculation of the average Ka/Ks for three coding sequences (S, Orf1a, Orf1b) of the SARS-CoV genome within the three epidemic phases

(A) Estimation of the neutral mutation rate and the date for the most recent common ancestor (MRCA)

(1) **Selection of samples**

Culture artifacts may potentially introduce apparent sequence variations in the SARS-CoV genome. Therefore, we used sequences that are derived directly from the patients' clinical specimens for the present statistical analysis. Sequences generated from specimens collected more than 4 weeks after disease onset were also excluded. Among all of the available sequences, 10 (GZ02, CUHK-AG01, CUHK-AG02, GZ-C, GZ-D, HZS2-A, HZS2-Fb, HSZ-A, HSZ-Bb, HSZ-Cb) met all the criteria. We used GZ02 as the out-group, since it is the most divergent from all of the remaining 9 sequences (see Fig. 2 in the text).

(2) **Estimation of the neutral mutation rate**

The Pamilo-Bianchi-Li model (*S1–2*) was used to calculate the Ks for the 6 known concatenated coding sequences (orf1a, orf1b, S, E, M, and N) of the SARS-CoV genome. We plotted the Ks versus the sampling dates, defined as the number of days away from January 1, 2003 (fig S5).

There is a positive correlation between the Ks and the sampling dates (correlation

coefficient: 0.82). Thus, the synonymous substitution rate appears to be relatively constant throughout the sampling period. As there are less selective constraints for synonymous mutations in general, we proposed to use the synonymous substitution rate per site per day for the concatenated coding sequences to represent the neutral mutation rate of the SARS-CoV genome.

A simple linear regression model was used to estimate the neutral base substitution rate. The rate was estimated from the slope (β_1) of the fitted linear regression line and is found to be 8.26×10^{-6} ($\pm 2.16 \times 10^{-6}$) $nt^{-1}day^{-1}$. This estimated mutation rate is quite similar to the values obtained for other known RNA viruses (*S3–4*). Specifically, the rate is about one third that for human immunodeficiency virus (*S3*).

The relationship can thus be described as, $Y = \beta_0 + \beta_1 X$ where Y is the Ks (using GZ02 as out-group) and X is the sampling date, which is measured by the number of days away from January 1, 2003.

(3) **Estimation of the date of MRCA for the available samples**

The intercept (β_0) of the fitted line is 1.055×10^{-3}, which equates to a sampling date of 0 and thus, corresponded to a calendar date in the end of year 2002. We used the GZ02 sequence as the out-group, which was sampled on February 11, 2003 (*i.e.*, 42 days after January 1, 2003). If T denotes the number of days before January 1, 2003 for the occurrence of the MRCA, then $T = \frac{\hat{\beta}_0 / \hat{\beta}_1 - 42}{2} = 46(days)$, which is equivalent to mid-November of 2002. The 95% confidence interval for T is estimated to be 5.5 - 201.5 days (Supporting References and Notes: Appendix 1) (*S5*), meaning that the date for the MRCA is estimated to range from early June, 2002 to end of December, 2002. We know from the phylogenetic tree generated from all of the SARS-CoV sequences compared in this study (refer to Fig. 2 of the text and fig. S6) that the ancestor node of

the 10 samples we used to estimate the neutral mutation rate is the same as that for all of the other available human SARS-CoV sequences. Thus the estimated date of the MRCA should be applicable to all of the studied human SARS-CoV sequences.

(B) Calculation of the average Ka/Ks for three coding sequences (S, Orf1a, Orf1b) of the SARS-CoV genome within the three epidemic phases

The 61 human SARS-CoV genotypes were divided into three groups according to the three different phases of the epidemic (fig. S6: green for early phase, red and purple for middle phase and blue for late phase). The Ka/Ks for the S gene, Orf1a and Orf1b sequences were calculated for each group. The results are shown in table S3. All ratios were calculated in a similar manner and here we illustrate the method of calculation using the S gene. We included all the different S coding sequences in the calculation. For samples with identical sequences, one was randomly selected.

Within each group, Ka/Ks ratios were first calculated in a pairwise manner (2 sequences at a time) according to the Pamilo-Bianchi-Li model (*S1-2*). The average Ka/Ks and its standard error (table S3) were obtained using all these pairwise Ka/Ks values except those equal to infinite which is due to Ks=0. One-sided unpaired two-sample t-test was used to test whether the average Ka/Ks ratios for each of the studied coding region of the SARS-CoV sequences during the different epidemic phases were significantly different (table S3).

Supporting Online Text S2:

Acknowledgements

We appreciate the strong support from the Minister of Science and Technology of the Central Government of China, the Governments of Guangdong Province, Shanghai Municipality, and the Hong Kong Special Administrative Region. This work was supported by the State High Technology Development Program (863, Grant No. 2003AA208407), the State Key Program for Basic Research (973, Grant No. 2003CB514101), and a grant from the Guangdong Government for SARS research (2003FD02). RWKC, JST and YMDL are supported by the Hong Kong Research Grants Council Special Grant for SARS Research (CUHK4508/03M). HDS, GWZ, BWG, SJC, and ZC are partly supported by a SARS research grant from the BNP PARIBAS. PH, WZH and YXL are partly supported by the Shanghai Commission of Science and Technology. SYZ is supported by a special grant on SARS reservoir of the Institute of Zoology, Chinese Academy of Sciences, and a grant (30340035) from the National Science Foundation of China (NSFC). XNW and JLH are supported by the “973” grants (2001CB510008 and 2003CB514113) and the NSFC Fund for Distinguished Young Scholars. XJZ is supported by the "863" grants (2002AA214151 and 2003AA2Z347A). HT is supported by the NIH grant to C.I.W. We are grateful to the critical technical assistance supplied by Hui-Qiong ZHOU[1], Ping HUANG[1], Li-Mei DIAO[1], Qiu-Xia CHEN[1], Yan SHENG[8], Yi CHEN[8], Zheng RUAN[8], Chun-Lei JIANG[8], Yu LIU[8], Wei SHEN[14], and Li WANG[14]. We thank the Institute of Biochemistry and Cell Biology, SIBS, CAS for supplying VeroE6 cells. We thank Li Ruan, Wuchun Cao, and Ruifu Yang for sharing their unpublished information with us. We would also like to express our special appreciation to Shou-Yi Yu, Antione Danchin and Linfa Wang for helpful suggestions and comments on our research strategy and the manuscript preparation.

*All affiliations are the same as those listed in the main text.

Supporting References and Notes

Supporting references

S1. P. Pamilo, N. O. Bianchi, *Mol. Biol. Evol.* **10,** 271 (1993).

S2. W. H. Li, *J. Mol. Evol.* **36,** 96 (1993).

S3. W. H. Li, M. Tanimura, P. M. Sharp, *Mol. Biol. Evol.* **5,** 313 (1988).

S4. J. W. Drake, J. J. Holland, *Proc. Natl. Acad. Sci. U. S. A.* **96,** 13910 (1999).

S5. J. A. Nelder, P. McCullagh, *Generalized Linear Models* (Chapman and Hall, London, ed. 2, 1989)

S6. N. Saitou, M. Nei, *Mol. Biol. Evol.* **4,** 406 (1987).

S7. M. Kimura, *J. Mol. Evol.* **16,** 111 (1980).

Supporting notes:

Appendix 1: Calculation of 95% confidence interval for T:

From the linear model, we obtained $\hat{\beta}_0 = 1.055$ x 10^{-3} and $\hat{\beta}_1 = 8.257$ x 10^{-6}

Let's denote $\theta = \hat{\beta}_0 / \hat{\beta}_1$. So $\frac{\hat{\beta}_0 - \theta\hat{\beta}_1}{\hat{\sigma}\sqrt{Q(\theta)}} \sim t_{df=7}$, where $Q(\theta) = c_{11} - 2\times\theta\times c_{12} + \theta^2 c_{22}$.

From the covariance matrix of the fitted linear model, we can get $\hat{\sigma}^2 \times c_{11} = 2.2658$ x 10^{-8}, $\hat{\sigma}^2 \times c_{12} = -3.0558$ x 10^{-10} and $\hat{\sigma}^2 \times c_{22} = 4.7013$ x 10^{-12}. So the 95% confidence of θ will satisfy,

$$\Pr\{|\frac{1.055\times10^{-3} - \theta\times8.257\times10^{-6}}{\sqrt{2.266\times10^{-8} + 2\theta\times3.0558\times10^{-10} + \theta^2\times4.7013\times10^{-12}}}| < t_{7,0.025}\} = 0.05.$$

When we solved the inequality, we obtained the 95% confidence interval (CI) for θ ($\hat{\beta}_0 / \hat{\beta}_1$) which is (53, 445.1). Since T is equal to $\frac{\hat{\beta}_0 / \hat{\beta}_1 - 42}{2}$, the 95% CI for T will be estimated as (5.5, 201.55) (*S5*).

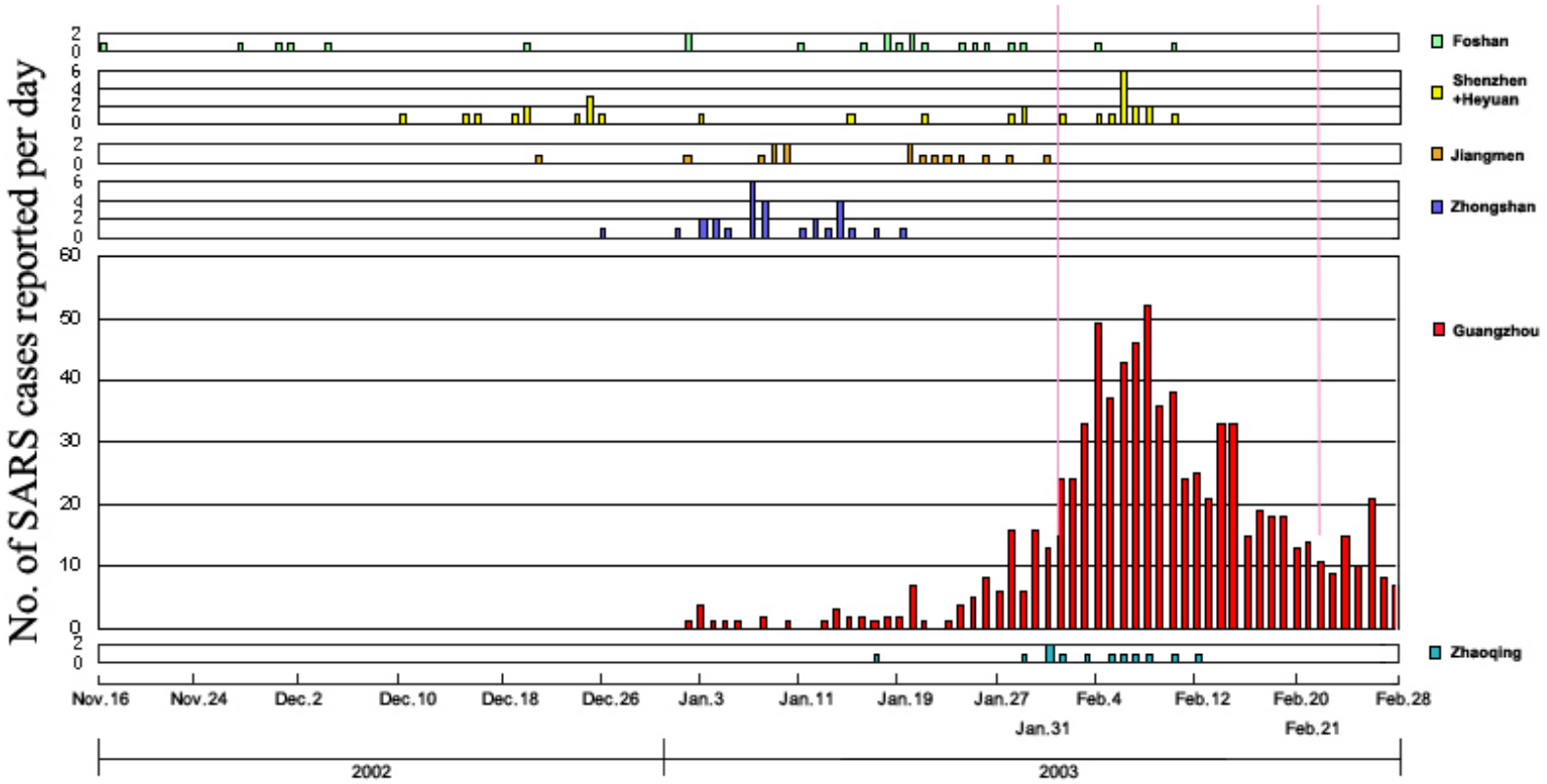


Fig. S1. Number of daily documented SARS cases reported from individual cities of the Guangdong Province, China, up to February 2003. Original epidemiological data were collected and analyzed by the Guangdong Center for Disease Control and Prevention. We combined the cases reported from the cities of Heyuan and Shenzhen because the Heyuan index case became infected in Shenzhen and after this nosocomial infection, no additional infections were reported in Heyuan. The order of the cities is arranged from top to bottom based on the disease onset date of their respective index cases, starting from the earliest to the latest dates of onset.

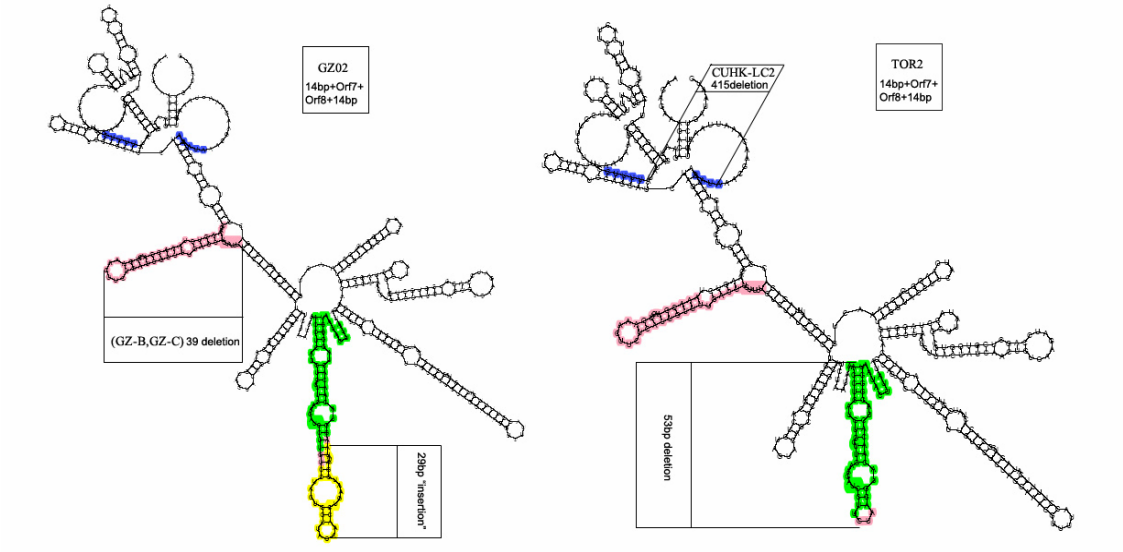


Fig. S2. Predicted RNA secondary structure of the Orf7b-Orf8 region of the SARS-CoV genome. SARS-CoV genotypic variations caused by major deletion events were observed on a number of occasions during the epidemic. All such deletions were confined to the Orf7b-Orf8 region. The genomic locations of the major deletions observed in this study are indicated on the predicted RNA secondary structures of the longest SARS-CoV genotype (left panel) and the genotype with the 29-nt deletion (right panel). The former genotype is represented by GZ02 while the latter is represented by TOR2. This latter genotype predominated the remainder of the epidemic from the middle phase onwards. For both panels, the illustrated region starts from 14 nucleotides upstream to the start of the predicted Orf7 to 14 nucleotides downstream to the end of Orf8. The illustrated region corresponds to nucleotide positions 27288 to 28161 on GZ02 and nucleotide positions 27259 to 28132 on TOR2. The prediction was made using the VIENNARNA:RNAfold software (http://bioweb.pasteur.fr/). GZ-B and GZ-C are two genotypes obtained from two Guangzhou patients with disease onset from mid-March but demonstrated a 39-nt deletion.

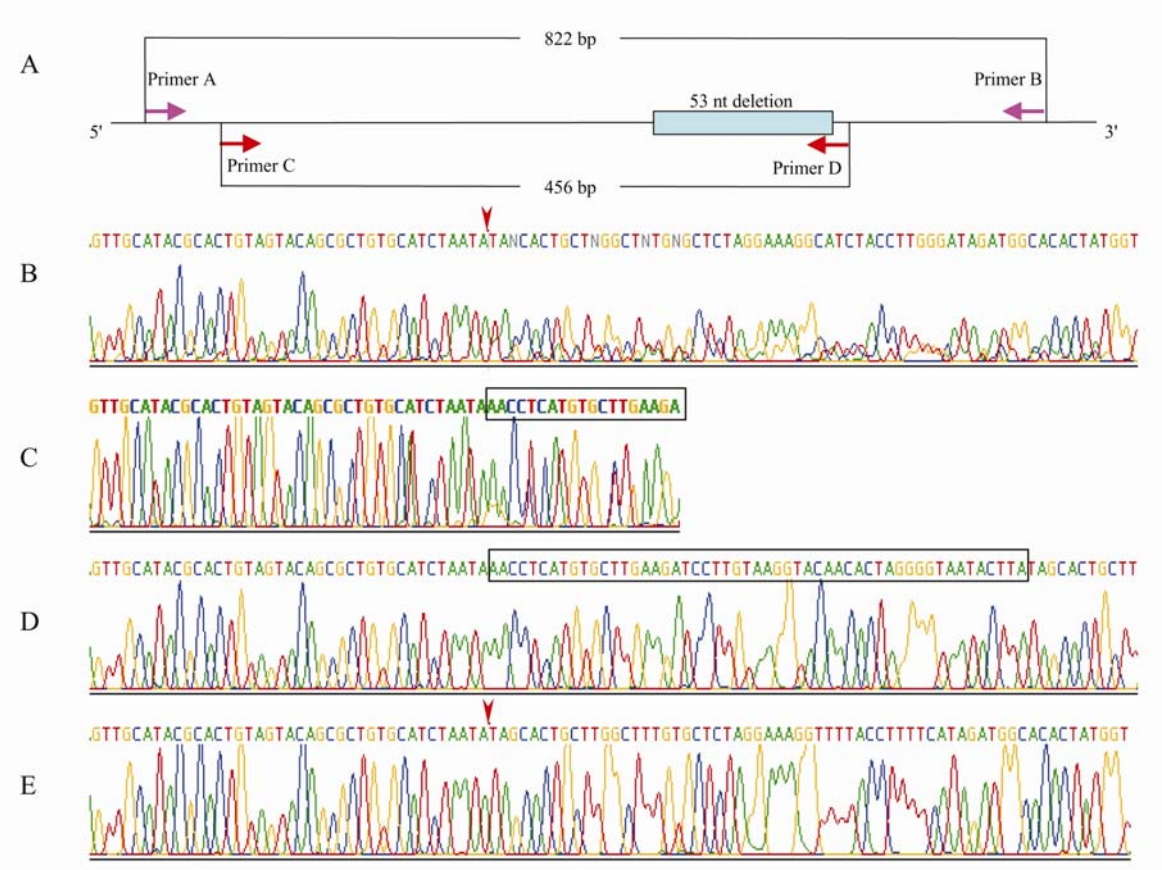
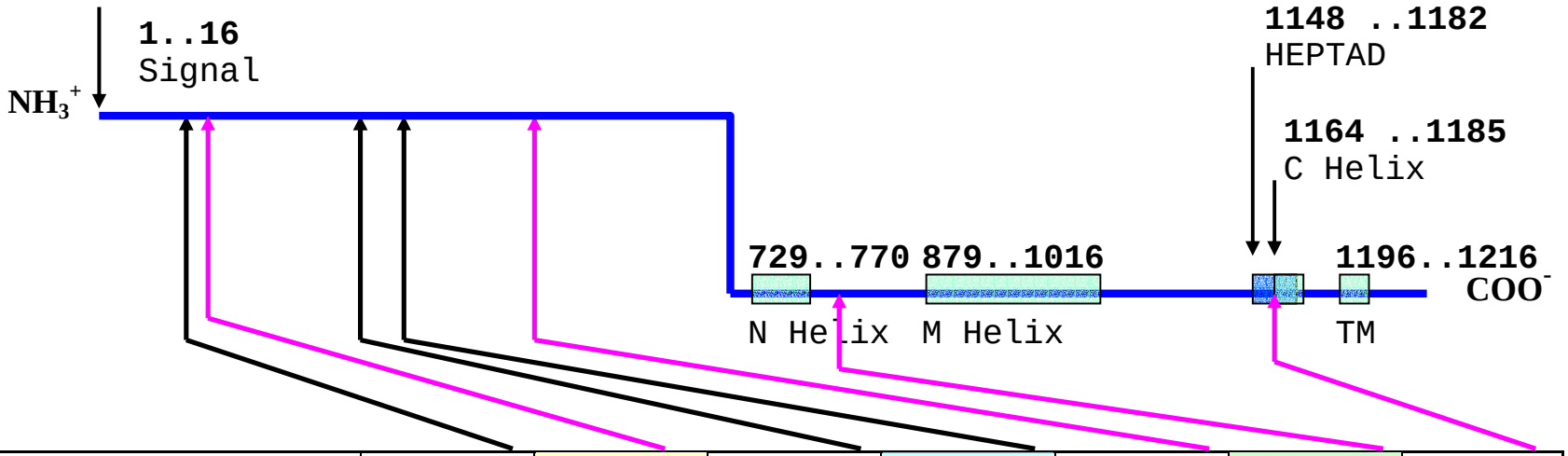


Fig. S3. Co-existence of two SARS-CoV genotypes, with or without a 53-nt deletion, in a lung biopsy specimen of patient HZS2-B. (**A**) Schematic illustration of the locations of the primers used. The filled box denotes the 53-nt segment. When primers A (5'-tagcacacactttgcttttg-3') and B (5'-cagtattattgggtaaaccttgg-3') were used, both an 822-nt and a 769-nt fragment could be amplified, corresponding to the genotypes with or without the 53-nt segment, respectively. Direct sequencing of this PCR product illustrated in (**B**) showed a row of mixed sequence to the right side of the red arrow head. If primers C (5'-ttttgcttgtgctgacggtac-3') and D (5'-cagtgctataagtattacccctagtg-3') were used, only the sequence without the 53-nt deletion was amplified because primer D is complementary to the sequence of the deleted region, as the partial sequence of this region illustrated in (**C**). For additional confirmation, the PCR amplicons obtained using primers A and B were cloned into the PGEM-T vector. Among the 27 randomly picked clones for sequencing, 10 clones had the same sequence as other viral isolates of the middle phase (**D**), while the other 17 clones had sequences with the 53-nt deletion (**E,** deletion site marked by arrow head). The nucleotides corresponding to the deleted segment are boxed in (**C**) and (**D**).



Epidemic phase	GZ02 nucleotide coordinate	21715	21721	22207	22222	22422	23823	24978
	S gene nucleotide coordinate	224	230	716	731	931	2,332	3,487
	S gene amino acid coordinate	75	77	239	244	311	778	1,163
Early	S gene codon (amino acid)	aGg (Arg)	gAc (Asp)	tTa (Leu)	aCt (Thr)	Aga (Arg)	Gat (Asp)	Gaa (Glu)
Late	S gene codon (amino acid)	aCg (Thr)	gGc (Gly)	tCa (Ser)	aTt (Ile)	Gga (Gly)	Tat (Tyr)	Aaa (Lys)

Fig. S4. The predicted amino acid residue alterations in the S protein caused by non-synonymous SNVs observed in the epidemic. The amino acid residue alterations predicted are listed and mapped to the approximate regions of the modeled S protein. Nucleotides 21721, 22222 and 23823 are the loci included in the 5-nt motif used for the classification of the major genotypes. These loci are shaded in the same colors as in Fig. 2 or Table S1.

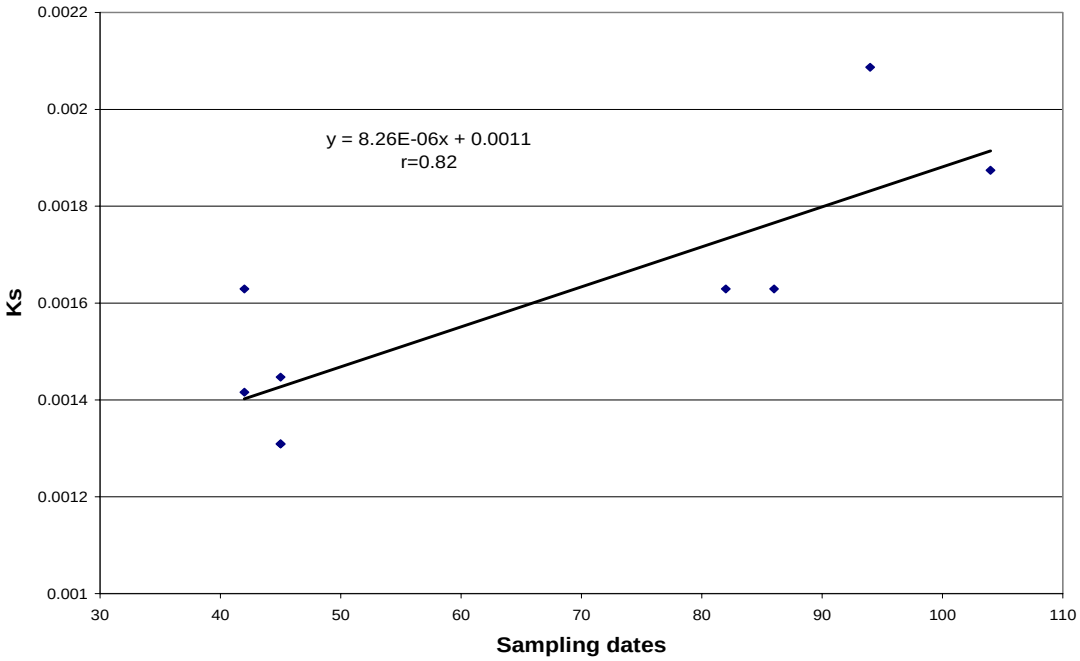


Fig. S5. Synonymous substitution rate, Ks. for the concatenated coding sequences versus sampling dates. The principles for sample selection and the statistical analysis methods are described above. The relatively most divergent sequence GZ02 was used as the out-group and Ks was determined for 9 representative human SARS-CoV sequences. The sampling dates are measured as the number of days away from January 1, 2003. A simple linear regression model was used to estimate the neutral base substitution rate.

Fig. S6: Neighbor-joining tree for the 60 SARS-CoV genomic sequences. ZJ01 was omitted for its low quality sequence and the tree was constructed using the Kimura two-parameter distance (*S 7*).

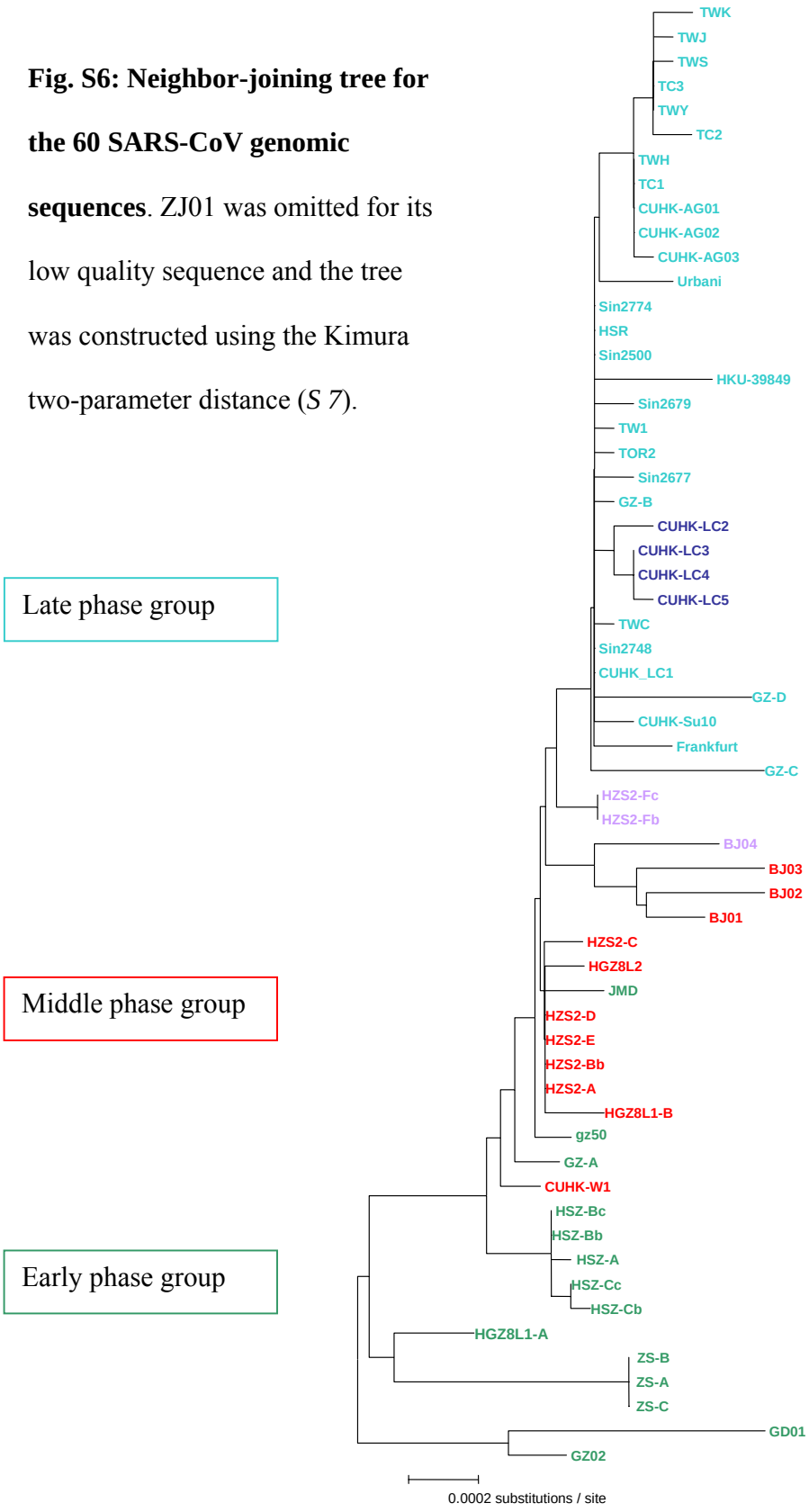
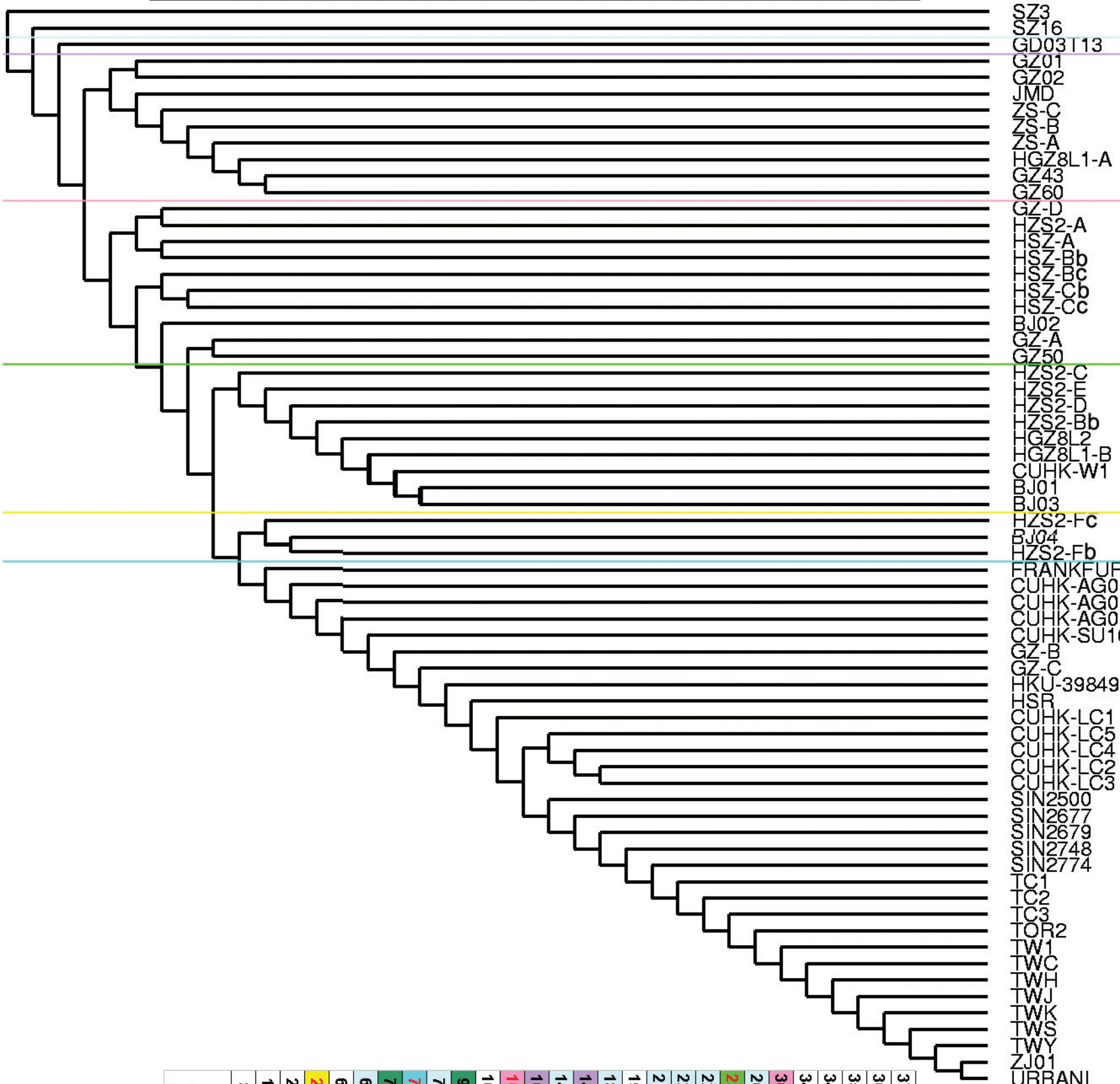


Fig. S7. Genotype clustering of the S genes from human SARS-CoV and palm civet SARS-like-CoV of the 2002/2003 epidemic and the 2003/2004 Guangdong index patient. The rooted phylogenetic tree for the nucleotide sequences of S genes from 2 palm civet SARS-like-CoV sequences (SZ16 and SZ3) and 64 human SARS-CoV sequences (61 as those used in the whole genome analysis of this project, in addition to two more S gene sequences, gz43 and gz60 of the 2002/2003 epidemic and the S gene sequence of the 2003/2004 Guangdong index patient, GD03T13 [Materials and Methods]). Only those variant sequences that were present in at least two independent samples were used for tree construction (total of 28 SNPs, synonymous and non-synonymous; Table S4). The map distance between individual sequences represents the extent of genotypic difference. A 6-nt motif that characterized the major phylogenetically-related genotypes are indicated in boxes. The nucleotide 24566, corresponding to the S gene nucleotide 3075 was not included in the characteristic motif because it only caused a synonymous variation. The sequences are named in concordance with their GenBank nomenclature.

S-nt	81	146	224	230	654	681	716	731	782	931	1026	1031	1079	1437	1460	1819	1994	2102	2227	2261	2332	2680	3075	3442	3487	3536	3623	3739
SZ3	T	C	C	A	T	A	T	C	A	G	G	G	C	A	G	C	C	T	G	T	G	G	C	C	G	T	C	C
GD03T13	T	C	C	A	T	C	C	C	C	G	G	G	C	T	G	T	C	C	A	C	G	A	C	C	G	T	C	C
GZ02	T	C	C	**A**	**C**	C	**T**	**C**	C	**A**	**G**	**G**	T	T	C	T	T	C	A	C	**G**	A	**C**	C	**G**	T	C	C



S-nt	81	146	224	230	654	681	716	731	782	931	1026	1031	1079	1437	1460	1819	1994	2102	2227	2261	2332	2680	3075	3442	3487	3536	3623	3739
URBANI	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C

Table S1. The predicted coding sequence changes caused by the major deletion events in the Orf7b-Orf8 region of the SARS-CoV genome.

	Orf7b	Orf8a	Orf8b	Nucleocapsid Protein
GZ02	27635:27769 MNELTLIDFYLCFLA FLLFLVLIMLIIFWF SLEIQDLEEPCTKV*	27776: 28144 (Sars8) MKLLIVLTCISLCSCI**R**TVVQRCASNKPHVLEDPCPT GYQPEWNIRYNTRGNTYSTAWLCALGKVLPFHRWHTM VQTCTPNVTINCQDPAGGALIARCWYLHEGHQTAAFR DVLVVLNKRTN*		28146: 29414 sars9a nucleocapsid protein MSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTALTQHGKEELRFPRGQGVPINTNSGPDDQIGYYR RATRRVRGGDGKMKELSPRWYFYYLGTGPEASLPYGANKEGIVWVATEGALNTPKDHIGTRNPNNNAATVLQLPQGTTLPKGFYAEGSR GGSQASSRSSSRSRGNSRNSTPGSSRGNSPARMASGGGETALALLLLDRLNQLESKVSGKGQQQQGQTVTKKSAAEASKKPRQKRTATK QYNVTQAFGRRGPEQTQGNFGDQDLIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYHGAIKLDDKDPQFKDNVILLNK HIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADMDDFSRQLQNSMSGASADSTQA*
TOR2 29 dels	27635:27769 MNELTLIDFYLCFLA FLLFLVLIMLIIFWF SLEIQDLEEPCTKV*	27776..27924 sars8a unknown (sars8a) MKLLIVLTCISLCSC I**C**TVVQRCASNKPHV LEDPCKVQH	27861: 28144 sars8b unknown (sars8b) MCLKILVRYNTRGNTYSTA WLCALGKVLPFHRWHTMVQ TCTPNVTINCQDPAGGALI ARCWYLHEGHQTAAFRDVL VVLNKRTN*	28146: 29414 sars9a nucleocapsid protein MSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTALTQHGKEELRFPRGQGVPINTNSGPDDQIGYYR RATRRVRGGDGKMKELSPRWYFYYLGTGPEASLPYGANKEGIVWVATEGALNTPKDHIGTRNPNNNAATVLQLPQGTTLPKGFYAEGSR GGSQASSRSSSRSRGNSRNSTPGSSRGNSPARMASGGGETALALLLLDRLNQLESKVSGKGQQQQGQTVTKKSAAEASKKPRQKRTATK QYNVTQAFGRRGPEQTQGNFGDQDLIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYHGAIKLDDKDPQFKDNVILLNK HIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADMDDFSRQLQNSMSGASADSTQA*
HGZ8L1-B 82 dels	27635:27769 MNELTLIDFYLCFLA FLLFLVLIMLIIFWF SLEIQDLEEPCTKV*	27776..27963 MKLLIVLTCISLCSC I**R**TVVQRCASNIALL GFVL*	*	28146: 29414 sars9a nucleocapsid protein MSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTALTQHGKEELRFPRGQGVPINTNSGPDDQIGYYR RATRRVRGGDGKMKELSPRWYFYYLGTGPEASLPYGANKEGIVWVATEGALNTPKDHIGTRNPNNNAATVLQLPQGTTLPKGFYAEGSR GGSQASSRSSSRSRGNSRNSTPGSSRGNSPARMASGGGETALALLLLDRLNQLESKVSGKGQQQQGQTVTKKSAAEASKKPRQKRTATK QYNVTQAFGRRGPEQTQGNFGDQDLIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYHGAIKLDDKDPQFKDNVILLNK HIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADMDDFSRQLQNSMSGASADSTQA*
CUHK-LC2 415 dels	27635:27769 MNELTLIDFYLCFLA FLLFLVLIMLIIFWF SLEIQDLEEPCTKV*	*	*	27635: 29414 MNELTLIDFYLCFLAFLLFLVLIMLIINEQIKMSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTAL TQHGKEELRFPRGQGVPINTNSGPDDQIGYYRRATRRVRGGDGKMKELSPRWYFYYLGTGPEASLPYGANKEGIVWVATEGALNTPKDH IGTRNPNNNAATVLQLPQGTTLPKGFYAEGSRGGSQASSRSSSRSRGNSRNSTPGSSRGNSPARMASGGGETALALLLLDRLNQLESKV SGKGQQQQGQTVTKKSAAEASKKPRQKRTATKQYNVTQAFGRRGPEQTQGNFGDQDLIRQGTDYKHWPQIAQFAPSASAFFGMSRIGME VTPSGTWLTYHGAIKLDDKDPQFKDNVILLNKHIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADMDDFSRQLQNSMS GASADSTQA*
GZ-C,GZ-B 39 dels +29 dels	27638:27844 MNELTLIDFYLCFLA FLLFLVLIMLIIFWF SLEIQDLEEPCTKVS LCSCICTVVQRCASN KPHVLEDP	*	*	28146: 29414 sars9a nucleocapsid protein MSDNGPQSNQRSAPRITFGGPTDSTDNNQNGGRNGARPKQRRPQGLPNNTASWFTALTQHGKEELRFPRGQGVPINTNSGPDDQIGYYR RATRRVRGGDGKMKELSPRWYFYYLGTGPEASLPYGANKEGIVWVATEGALNTPKDHIGTRNPNNNAATVLQLPQGTTLPKGFYAEGSR GGSQASSRSSSRSRGNSRNSTPGSSRGNSPARMASGGGETALALLLLDRLNQLESKVSGKGQQQQGQTVTKKSAAEASKKPRQKRTATK QYNVTQAFGRRGPEQTQGNFGDQDLIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYHGAIKLDDKDPQFKDNVILLNK HIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADMDDFSRQLQNSMSGASADSTQA*

The amino acid sequences of the Orf7b, Orf8 (8a and 8b) and N protein as predicted for the major SARS-CoV deletion variants are listed in the table. Amino acid sequence changes due to the deletion events are labeled in red. Corresponding nucleotide coordinates for each predicted open reading frame are based on the GZ02 sequence.

Table S2. List of all SNVs with multiple occurrences.

Sequence alignments were generated using CLUSTALW 1.83 with the Gonnet nuclear acid comparison matrix for the 63 SARS-CoV genomic sequences analyzed in this study (61 human SARS-CoV and 2 plam civet SARS-like-CoV sequences). The names of the genomic sequences (as from GenBank) are listed on the first column based on clusters determined by our scoring algorithm (see below). Characteristics of the SNVs observed are listed in the first row, including the variant nucleotides observed at the loci, the affected codon sequence, the resultant amino acids, the amino acid coordinate and the nucleotide coordinate (based on GZ02, shaded horizontally in blue). The relative genomic positions of the listed SNVs are indicated by the predicted open reading frames as shaded bars at the top of the Table.

Variant loci that are characteristic of and allow the segregation of SARS-CoV genotypes into major groups were determined by a method developed for this work. First, for a given mutation site i, sequences are sorted into two groups according to the nucleotide on this site. Second, The number of mutations on the other mutation site k of group j is counted to give $N_k^j(i)$. Then the resulting counts are summed up to give the score S (i) for site i.

$$S(i) = \sum_{k=1}^{C} \left\{ \sum_{j=1}^{2} N_k^j(i) \right\}$$

Where C is the number of mutation sites. Then the scores for all mutation sites are obtained by reiterating the above steps. Finally, the mutation site m that carries the smallest score is chosen to be the primary clustering marker and the genotypes are initially clustered based on this primary marker.

$$m = \arg\min_{1 \le i \le C} \{S(i)\}$$

Once the first level of clustering is determined, the above process can be applied recursively until all of the sub-clustering is completed. Further fine adjustment in the resultant clusters / groups of genotypes was performed by integrating information of the epidemiological relationship of genotypes and sequence quality. The same data set was used to generate an unrooted phylogenetic tree with the PHYLIP software package (Fig. 2).

The table lists all of the 107 SNVs that are observed in more than one of the 63 genomic sequences. Only 85 of these SNVs are seen in more than one of the human-derived sequences. Fifty-two of these SNVs were predicted to cause amino acid changes (non-synonymous variations).

SNVs that contribute to the grouping of genotypes based on the predominant clustering criteria described above were further highlighted in the first row with different color shading, except for:

1. The 22 nucleotide sites exhibiting a sequence only observed in the SARS-like-CoV sequences from palm civets or with a single variation observed in human SARS-CoV sequences. These SNVs are shaded with orange color.

2. Synonymous variations.

3. Non-synonymous variations causing similar amino acid alterations, with the exception of the 2 SNVs that were shown to be significantly associated with the transition between the epidemic phases, namely, nt 9404 and nt 17564. The nt 22522 was also highlighted for its association with some minor phase variation events.

4. Some SNVs with the following features:

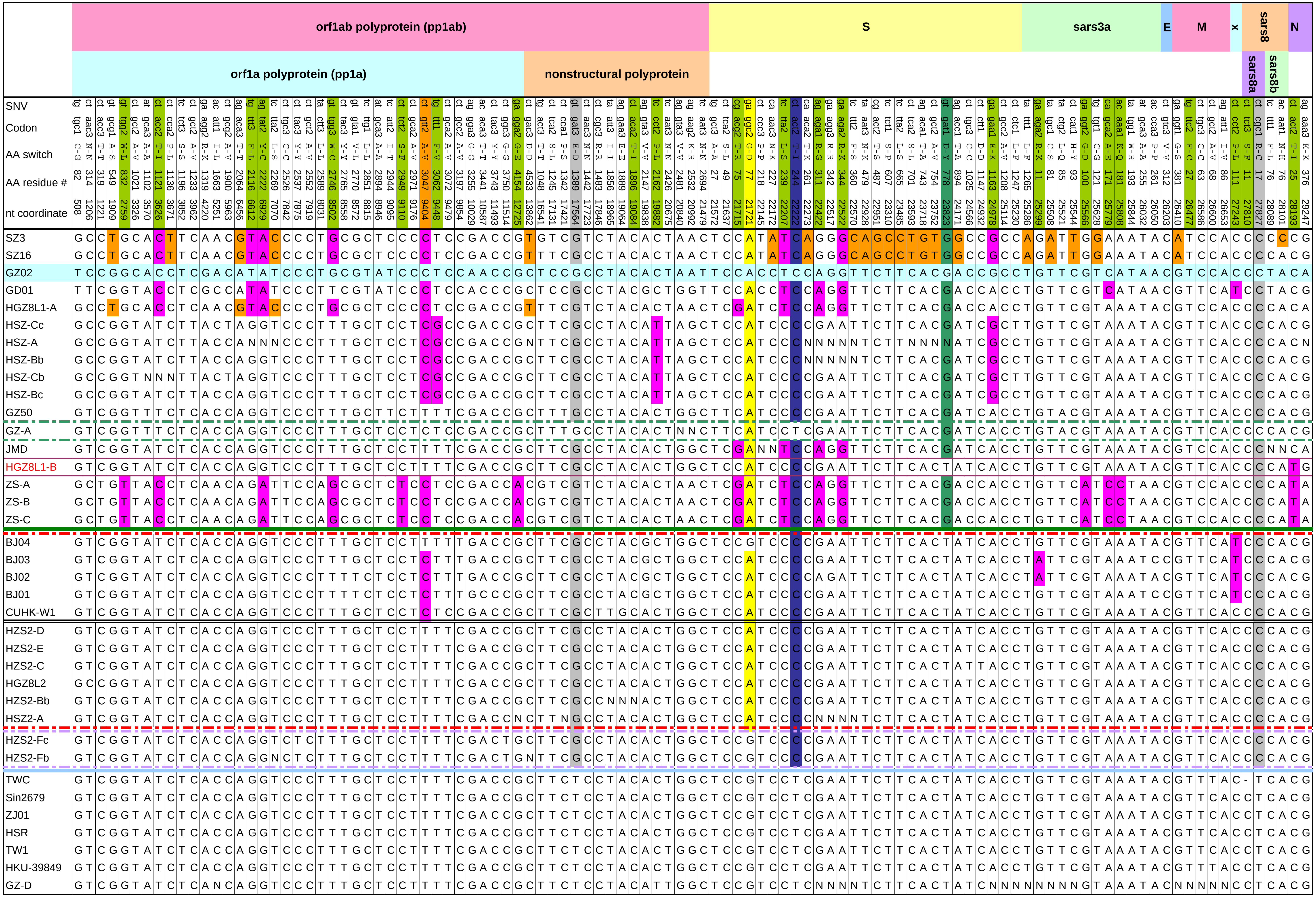
i. Variations that were only shown in samples of the same patient (*e.g.*, nts 508, 17131, and 28089 for GZ02/GZ01[GD01]; nt 25320 for HSZ-Cc/HSZ-Cb).

ii. The variations were only shared by two sequences, one or both of the sequences belonged to a minor genotype group (*e.g.*, nts 17421, 21637, and 25521 for the genotypes gz50 and GZ-A).

iii. Variations that were not consistent within one transmission path (*e.g.*, nt 9095 for GZ02/GD01 and gz50; nt 25844 for GZ02/GD01 and the ZS group).

The positions of the 5-nt motif used to classify the major genotypes are shaded, including the 2 loci external to the S gene (17564 and 27827, shaded in grey) and the three loci in the S gene, namely positions 21721 (yellow), 22222 (blue) and 23823 (green). These S gene residues are shaded in the same color scheme as used in Figs. 2 and S4. All other SNVs that contribute to the clustering of genotypes are highlighted in pink.

The listed genomic sequences are clustered into groups based on our scoring method. Major genotype clusters were demarcated by solid horizontal lines, where a green line separates the early and middle phases, while a blue line separates the middle and late phases. The further sub-classification of genotypes within the major groups is demarcated by dashed horizontal lines.



orf1ab polyprotein (pp1ab)
S
sars3a
E
M
x
sars8
N
orf1a polyprotein (pp1a)
nonstructural polyprotein
sars8a
sars8b
SNV
Codon
AA switch
AA residue #
nt coordinate
SZ3
SZ16
GZ02
GD01
HGZ8L1-A
HSZ-Cc
HSZ-A
HSZ-Bb
HSZ-Cb
HSZ-Bc
GZ50
GZ-A
JMD
HGZ8L1-B
ZS-A
ZS-B
ZS-C
BJ04
BJ03
BJ02
BJ01
CUHK-W1
HZS2-D
HZS2-E
HZS2-C
HGZ8L2
HZS2-Bb
HSZ2-A
HZS2-Fc
HZS2-Fb
TWC
Sin2679
ZJ01
HSR
TW1
HKU-39849
GZ-D

	orf1ab polyprotein (pp1ab)																																			
	orf1a polyprotein (pp1a)																																			
SNV	tg	ct	ct	gt	gt	ct	at	ct	ct	tc	ct	ga	ac	ct	ag	tg	ag	tc	ct	ct	ct	ta	gt	ct	gt	tc	at	tc	ct	ct	ct	tg	ct	ct	ag	ac
Codon	tgc1	aac3	acc3	gcg1	tgg2	gct2	gca3	acc2	cca2	tct3	gct2	agg2	att1	gcg2	aca3	ttt3	tat2	tta2	tgc3	tac3	gct2	ctt3	tgg3	tac3	gta1	ttg1	gca3	att2	tct2	gca2	gtt2	ttt1	gca2	gcc2	gga3	aca3
AA switch	C-G	N-N	T-T	A-S	W-L	A-V	A-A	T-I	P-L	S-S	A-V	R-K	I-L	A-V	T-T	F-L	Y-C	L-S	C-C	Y-Y	A-V	L-L	W-C	Y-Y	V-L	L-L	A-A	I-T	S-F	A-V	A-V	F-V	A-V	A-V	G-G	T-T
AA residue #	82	314	319	549	832	1021	1102	1121	1136	1196	1233	1319	1663	1900	2064	2116	2222	2269	2526	2537	2552	2589	2746	2765	2770	2851	2894	2944	2949	2971	3047	3062	3072	3197	3255	3441
nt coordinate	508	1206	1221	1909	2759	3326	3570	3626	3671	3852	3962	4220	5251	5963	6456	6612	6929	7070	7842	7875	7919	8031	8502	8558	8572	8815	8946	9095	9110	9176	9404	9448	9479	9854	10029	10587
Urbani	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	T	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
Sin2748	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
Sin2677	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
Sin2500	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
Frankfurt	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
Sin2774	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-Su10	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-LC1	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-AG01	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-AG02	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-AG03	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TWH	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TC1	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TWY	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TWS	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TWK	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TWJ	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TC3	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
TC2	G	T	C	G	G	T	A	T	C	C	C	A	C	C	A	G	G	T	C	C	T	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
GZ-B	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
GZ-C	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	C	T	T	C	G	A
TOR2	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-LC2	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-LC3	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-LC4	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A
CUHK-LC5	G	T	C	G	G	T	A	T	C	T	C	A	C	C	A	G	G	T	C	C	C	T	T	T	G	C	T	C	C	T	T	T	T	C	G	A

																				S																
				nonstructural polyprotein																																
SNV	ct	ct	ga	ct	tg	ct	ct	gt	ct	ct	ta	ag	ct	ag	tc	tc	ag	ag	tc	tg	ct	cg	ga	ct	ca	tc	ct	ca	ag	ga	ga	tc	ta	cg	tc	tc
Codon	tac3	ggc3	gga2	gac3	act3	tca2	cca1	gat3	aac3	cgc3	att3	gaa3	aca2	gta3	cct2	aat3	gta3	aag2	aat3	gct3	tca2	acg2	ggc2	ccc3	aac3	tta2	act2	aca2	aga1	agg3	aga2	ttt2	aat3	act2	tct1	tta2
AA switch	Y-Y	G-G	G-E	D-D	T-T	S-L	P-S	E-D	N-N	R-R	I-I	E-E	T-I	V-V	P-L	N-N	V-V	K-R	N-N	A-A	S-L	T-R	G-D	P-P	N-k	L-S	T-I	T-K	R-G	R-R	R-K	F-S	N-K	T-S	S-P	L-S
AA residue #	3743	3750	4154	4533	1048	1245	1342	1389	1442	1483	1856	1889	1896	2147	2162	2426	2481	2532	2694	27	49	75	77	218	227	239	244	261	311	342	344	360	479	487	607	665
nt coordinate	11493	11514	12725	13862	16541	17131	17421	17564	17723	17846	18965	19064	19084	19838	19882	20675	20840	20992	21479	21572	21637	21715	21721	22145	22172	22207	22222	22273	22422	22517	22522	22570	22928	22951	23310	23485
Urbani	C	C	G	C	T	T	C	T	C	C	T	G	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
Sin2748	C	C	G	C	T	T	C	T	C	C	T	A	T	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
Sin2677	C	C	G	C	T	T	C	T	C	C	T	A	T	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
Sin2500	C	C	G	C	T	T	C	T	C	C	T	A	T	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
Frankfurt	C	C	G	C	T	T	C	T	C	C	A	A	T	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
Sin2774	C	C	G	C	T	T	C	T	C	C	A	A	T	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-Su10	C	C	G	C	T	T	C	T	C	T	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-LC1	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-AG01	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-AG02	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-AG03	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TWH	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TC1	T	C	G	C	T	T	C	T	C	C	T	G	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TWY	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TWS	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TWK	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TWJ	T	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TC3	T	C	G	C	T	T	C	T	C	C	T	G	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TC2	T	C	G	C	T	T	C	T	C	C	T	G	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
GZ-B	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
GZ-C	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
TOR2	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-LC2	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	T	G	G	C	G	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-LC3	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	C	G	G	C	G	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-LC4	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	C	G	G	C	G	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T
CUHK-LC5	C	C	G	C	T	T	C	T	C	C	T	A	C	A	C	C	G	G	C	G	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T

											sars3a												E	M					sars6	sars8				N	
																														sars8a		sars8b			
SNV	ct	ag	ct	gt	ag	ct	ct	ga	ct	ct	ta	ga	ta	ta	ct	ga	tg	ca	ac	ta	at	ac	ct	ga	tg	ct	ct	ag	ct	ct	ct	tc	ac	ct	ag
Codon	tca2	aca1	gct2	gat1	acc1	tgc3	ctt1	gaa1	gcc2	ctc1	ttt1	aga2	tgc1	ctg2	cat1	ggt2	tgt1	gca2	aaa1	tgg1	gca3	cca3	gtc3	ggt1	ttc2	tgc3	gct2	att1	cct2	tct3	cgc1	ttt1	aat1	act2	aaa3
AA switch	S-L	T-A	A-V	D-Y	T-A	C-C	D-D	E-K	A-V	L-F	L-F	R-K	C-S	L-Q	H-Y	G-D	C-G	A-E	K-Q	W-R	A-A	P-P	V-V	G-S	F-C	C-C	A-V	I-V	P-L	S-F	R-C	F-L	N-H	T-I	K-K
AA residue #	701	743	754	778	894	1025	1147	1163	1208	1247	1265	11	81	85	93	100	121	171	181	193	255	261	312	381	27	63	68	86	111	11	17	76	76	25	376
nt coordinate	23593	23718	23752	23823	24171	24566	24932	24978	25114	25230	25286	25299	25508	25521	25544	25566	25628	25779	25808	25844	26032	26050	26203	26410	26477	26586	26600	26653	27243	27810	27827	28089	28101	28193	29247
Urbani	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	C	T	C	A	C	G
Sin2748	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	-	T	C	A	C	G
Sin2677	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	C	T	C	A	C	G
Sin2500	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	C	T	C	A	C	G
Frankfurt	C	A	C	T	A	T	T	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	T	A	C	C	T	C	A	C	G
Sin2774	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	C	T	C	A	C	G
CUHK-Su10	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	A	C	G
CUHK-LC1	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	A	C	G
CUHK-AG01	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	A	C	G
CUHK-AG02	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	A	C	G
CUHK-AG03	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	C	C	G
TWH	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	A	C	G
TC1	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	G	T	C	A	C	C	T	C	A	C	G
TWY	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	T	G	G	T	C	A	C	T	T	C	A	C	G
TWS	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	T	G	G	T	C	A	C	T	T	C	A	C	G
TWK	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	T	G	G	T	C	A	C	T	T	C	A	C	G
TWJ	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	T	G	G	T	C	A	C	T	T	C	A	C	G
TC3	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	T	G	G	T	C	A	C	T	T	C	A	C	G
TC2	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	T	G	G	T	C	A	C	T	T	C	A	C	G
GZ-B	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	G	C	C	T	C	A	C	G
GZ-C	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	G	C	C	T	C	A	C	G
TOR2	C	A	C	T	A	T	C	A	C	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	C	T	C	A	C	G
CUHK-LC2	C	A	C	T	A	T	C	A	T	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	-	-	-	-	C	G
CUHK-LC3	C	A	C	T	A	T	C	A	T	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	-	-	-	-	C	G
CUHK-LC4	C	A	C	T	A	T	C	A	T	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	-	-	-	-	C	G
CUHK-LC5	C	A	C	T	A	T	C	A	T	C	T	G	T	T	C	G	T	A	A	A	T	A	C	G	T	T	C	A	C	-	-	-	-	C	G

Table S3. Statistical analysis for the change of Ka/Ks ratios for different coding regions of the SARS-CoV sequences during the different epidemic phases.

Proteins	Epidemic phases	$\overline{Ks}(10^{-3})$	$\overline{Ka}(10^{-3})$	$\overline{Ka/Ks}$	$s.e.(\overline{Ka/Ks})$	H_1*	P-value†
Spike	early	1.698	1.997	1.248	0.081	Ka/Ks (early) > Ka/Ks (middle)	2.3 x 10^{-7}
	middle	2.377	0.898	0.410	0.087	Ka/Ks (middle) > Ka/Ks (late)	0.034
	late	1.355	0.257	0.219	0.043		
Orf1b	early	1.494	0.570	0.562	0.145	Ka/Ks (early) >Ka/Ks (late)	0.091
	middle	1.048	0.240	0.315	0.108		
	late	0.577	0.159	0.344	0.048		
Orf1a	early	1.264	1.030	0.923	0.124	Ka/Ks (early) > Ka/Ks (late)	7.4 x 10^{-5}
	middle	0.526	0.565	1.293	0.202		
	late	0.557	0.139	0.369	0.060		

*H_1 means the alternative hypothesis.

†One-sided unpaired two-sample t-test was use

Table S4. List of all SNVs in S genes with multiple occurrences.

Sequence alignments were generated by the method described in the notes for Table S2. Sequences used were described in the legend of Fig. S7. SNVs that contribute to the grouping of genotypes based on the predominant clustering criteria previously described were further highlighted in the rows with different color shading:

1. Green is for the SNV characteristically distinguishable between the early versus the middle phase viral isolates. Yellow is for the SNV that was characteristic for the motif transition between the middle phase and the late phase. Blue is for the SNV characteristically distinguishable between the middle versus the late phase viral isolates. These 3 coloring systems are consistent with that used for the whole genome analysis (Table S2 and Fig. 2)

2. Pink is for the SNVs characteristically distinguishable between groups of the early phase isolates, including synonymous and non-synonymous variations. Dark green is for the SNVs causing non-synonymous variations among significant portions of the early isolates. These variations were less systematic than the pink shaded SNVs with respect to their correlation with the epidemiological data.

3. Light blue indicates the palm civet specific SNVs known so far, while purple indicates the SNVs shared by palm civet SARS-like-CoV, SZ16 and SZ3, and the most recent human SARS-CoV, GD03T13.

nt coordinate	21572	21637	21715	21721	22145	22172	22207	22222	22273	22422	22517	22522	22570	22928	22951	23310	23485	23593	23718	23752	23823	24171	24566	24932	24978	25027	25114	25230
SNV	tg	ct	cg	ga	ct	ca	tc	ct	ca	ag	ga	ga	tc	ta	cg	tc	tc	ct	ag	ct	gt	ag	ct	ct	ga	tc	ct	ct
Codon	gct3	tca2	acg2	ggc2	cct3	aac3	tca2	att2	aca2	gga1	aga3	aaa2	ttt2	aat3	act2	tct1	tta2	tca2	aca1	gct2	tat1	acc1	tgt3	ctt1	aaa1	ctc2	gcc2	cta1
AA switch	A-A	S-L	T-R	G-D	P-P	N-k	L-S	T-I	T-K	R-G	R-R	R-K	F-S	N-K	T-S	S-P	L-S	S-L	T-A	A-V	D-Y	T-A	C-C	D-D	E-K	L-P	A-V	L-F
AA residue #	27	49	75	77	218	227	239	244	261	311	342	344	360	479	487	607	665	701	743	754	778	894	1025	1147	1163	1179	1208	1247
S-nt	81	146	224	230	654	681	716	731	782	931	1026	1031	1079	1437	1460	1819	1994	2102	2227	2261	2332	2680	3075	3441	3487	3536	3623	3739
SZ16	T	C	C	A	T	A	T	C	A	G	G	G	C	A	G	C	C	T	G	T	G	G	C	C	G	T	C	C
SZ3	T	C	C	A	T	A	T	C	A	G	G	G	C	A	G	C	C	T	G	T	G	G	C	C	G	T	C	C
GD03T13	T	C	C	A	T	C	C	C	C	G	G	G	C	T	G	T	C	C	A	C	G	A	C	C	G	T	C	C
GZ02_HBR	T	C	C	A	C	C	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	C	C	G	T	C	C
GZ01	T	C	C	A	C	C	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	C	C	A	T	C	C
gz43	T	C	G	A	T	C	T	C	C	A	G	G	T	T	C	T	C	C	A	C	G	A	C	C	A	C	C	C
gz60	T	C	G	A	T	C	T	C	C	A	G	G	T	T	C	T	C	C	A	C	G	A	C	C	A	C	C	C
ZS-B	T	C	G	A	T	C	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	C	C	A	T	C	C
ZS-A	T	C	G	A	T	C	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	C	C	A	T	C	C
HGZ8L1-A	T	C	G	A	T	C	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	C	C	A	T	C	C
ZS-C	T	C	G	A	T	C	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	C	C	A	T	C	C
JMD	T	C	G	A	N	N	T	C	C	A	G	G	T	T	C	T	T	C	A	C	G	A	T	C	A	T	C	C
BJ02	T	C	C	A	T	C	C	C	C	A	G	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
GZ50	T	T	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	A	T	C	C
GZ-A	T	T	C	A	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	A	T	C	C
HSZ-A	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	G	T	C	C
HSZ-Bb	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	G	T	C	C
HSZ-Cb	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	G	T	C	T
HSZ-Bc	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	G	T	C	C
HSZ-Cc	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	G	A	T	C	G	T	C	T
CUHK_W1	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HGZ8L1-B	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
BJ03	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
BJ01	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HZS2-D	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HZS2-E	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HZS2-C	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	T	A	T	C	C
HGZ8L2	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HZS2-Bb	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HSZ2-A	T	C	C	A	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
BJ04	T	C	C	G	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HZS2-Fb	T	C	C	G	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
HZS2-Fc	T	C	C	G	T	C	C	C	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TOR2	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
ZJ01	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TWY	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TWS	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TWK	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TWJ	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TWH	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TC3	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C
TC2	T	C	C	G	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	T	A	T	C	A	T	C	C

nt coordinate	21572	21637	21715	21721	22145	22172	22207	22222	22273	22422	22517	22522	22570	22928	22951	23310	23485	23593	23718	23752	23823	24171	24566	24932	24978	25027	25114	25230
SNV	tg	ct	cg	ga	ct	ca	tc	ct	ca	ag	ga	ga	tc	ta	cg	tc	tc	ct	ag	ct	gt	ag	ct	ct	ga	tc	ct	ct
Codon	gct3	tca2	acg2	ggc2	cct3	aac3	tca2	att2	aca2	gga1	aga3	aaa2	ttt2	aat3	act2	tct1	tta2	tca2	aca1	gct2	tat1	acc1	tgt3	ctt1	aaa1	ctc2	gcc2	ctc1
AA switch	A-A	S-L	T-R	G-D	P-P	N-k	L-S	T-I	T-K	R-G	R-R	R-K	F-S	N-K	T-S	S-P	L-S	S-L	T-A	A-V	D-Y	T-A	C-C	D-D	E-K	L-P	A-V	L-F
AA residue #	27	49	75	77	218	227	239	244	261	311	342	344	360	479	487	607	665	701	743	754	778	894	1025	1148	1163	1179	1208	1247
S-nt	81	146	224	230	654	681	716	731	782	931	1026	1031	1079	1437	1460	1819	1994	2102	2227	2261	2332	2680	3075	3442	3487	3536	3623	3739
TC1	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
HSR	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
TWC	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
SIN2774	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
SIN2748	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
SIN2679	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
SIN2677	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
SIN2500	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
FRANKFURT	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	**T**	A	T	C	C
TW1	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
CUHK_SU10	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
URBANI	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
HKU_39849	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
A11S	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
A7N	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
STL2	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
GZ-D	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
GZ-C	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
GZ-B	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
HK1	T	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	C	C
HK2	**G**	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	**T**	C
HK3	**G**	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	**T**	C
HK4	**G**	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	**T**	C
HK5	**G**	C	C	**G**	T	C	C	T	C	G	A	A	T	T	C	T	T	C	A	C	**T**	A	T	C	A	T	**T**	C