SARS-CoV Genome Polymorphism: A Bioinformatics Study

Gordana M. Pavlović-Lažetić[1]*, Nenad S. Mitić[1], Andrija M. Tomović[2], Mirjana D. Pavlović[3], and Miloš V. Beljanski[3]

[1] *Faculty of Mathematics, University of Belgrade, 11001 Belgrade, Serbia and Montenegro;* [2] *Friedrich Miescher Institute for Biomedical Research, CH-4058 Basel, Switzerland;* [3] *Institute of General and Physical Chemistry, 11001 Belgrade, Serbia and Montenegro.*

A dataset of 103 SARS-CoV isolates (101 human patients and 2 palm civets) was investigated on different aspects of genome polymorphism and isolate classification. The number and the distribution of single nucleotide variations (SNVs) and insertions and deletions, with respect to a "profile", were determined and discussed ("profile" being a sequence containing the most represented letter per position). Distribution of substitution categories per codon positions, as well as synonymous and non-synonymous substitutions in coding regions of annotated isolates, was determined, along with amino acid (a.a.) property changes. Similar analysis was performed for the spike (S) protein in all the isolates (55 of them being predicted for the first time). The ratio Ka/Ks confirmed that the S gene was subjected to the Darwinian selection during virus transmission from animals to humans. Isolates from the dataset were classified according to genome polymorphism and genotypes. Genome polymorphism yields to two groups, one with a small number of SNVs and another with a large number of SNVs, with up to four subgroups with respect to insertions and deletions. We identified three basic nine-locus genotypes: TTTT/TTCGG, CGCC/TTCAT, and TGCC/TTCGT, with four subgenotypes. Both classifications proposed are in accordance with the new insights into possible epidemiological spread, both in space and time.

Key words: SARS Coronavirus, single nucleotide polymorphism, insertions, deletions, spike protein, phylogenesis

Introduction

Severe acute respiratory syndrome (SARS), potentially fatal atypical pneumonia, first appeared in Guangdong province of China in November 2002 and soon afterward, within six months, spreaded all over the world (30 countries including China, Singapore, Vietnam, Canada, and USA), killing more than 700 people (*1*). In less than four weeks after the global outbreak, a novel member of Coronaviridae family, namely SARS Coronavirus (SARS-CoV), was identified in the blood of respiratory specimens and stools of SARS patients, and confirmed as the causative agent of disease according to the Koch postulates (*2*). Soon afterwards, first fully sequenced genomes of viral isolates were published (*3*, *4*). In 2005 the number of fully sequenced viral isolates exceeds one hundred (http://www.ncbi.nlm.nih.gov/entrez).

*** Corresponding author.**
E-mail: gordana@matf.bg.ac.yu

SARS-CoV probably originated due to genetic exchange (recombination) and/or mutations between viruses with different host specificities (*5*, *6*). Since coronaviruses are known to relatively easily jump among species, it was hypothesized that the new virus might have originated from wild animals. The analysis of SARS-CoV proteins supports and suggests possible past recombination event between mammalian-like and avian-like parent viruses (*6*). Common sequence variants define three distinct genotypes of the SARS-CoV: one linked with animal [palm civet (*Paguma larvata*)] SARS-like viruses and early human phase, the other two linked with middle and late human phases, respectively (*7*, *8*). SARS-CoV has a deleterious mutation of 29 nucleotides relative to the palm civet virus, indicating that if there was direct transmission, it went from civet to human, because deletions occur probably more easily than insertions (*5*). However, more recent reports indicate

that SARS-CoV is distinct from the civet virus and it has not been answered so far whether the SARS-CoV originated from civet, or civet was infected from other species (*9*, *10*). The genome is relatively stable, since its mutation rate has been determined to be between 1.83×10^{-6} and 8.26×10^{-6} nucleotide substitutions per site per day (*11*).

The SARS-CoV genome is approximately 30 Kb positive single strand RNA that corresponds to polycistronic mRNA, consisting of 5′ and 3′ untranslated regions (UTRs), 13 to 15 open reading frames (ORFs), and about 10 intergenic regions (IGRs) (*9*, *12*, *13*). Its genome includes genes encoding two replicate polyproteins (RNA-dependant-RNA-polymerase, *i.e.*, pp 1a and pp 1ab), encompassing two-thirds of the genome, and a set of ORFs at 3′ end that code for four structural proteins: surface spike (S) glycoprotein (1,256 a.a.), envelope (E, 77 a.a.), matrix (M, 222 a.a.), and nucleocapsid (N, 423 a.a.) proteins. It also encodes for additional 8–9 predicted ORFs whose protein product functions are still under investigation (*14*; http://www.ncbi.nlm.nih.gov/entrez).

The S protein is the main surface antigen of the SARS-CoV and is involved in virus attachment on susceptible cells using mechanism similar to those of class I fusion proteins. The receptor for the SARS-CoV S protein is identified as angiotensin-converting enzyme 2 (ACE-2), which is a metallopeptidase (*15*). The receptor-binding domain (RBD) has been determined to lay between a.a. postions 270–625 in recent studies (*16–20*).

Several epitope sites, defined by polyclonal or monoclonal antibodies, have been identified on the S protein, depending on experimental conditions, all lying within wide or narrow regions between a.a. 12–1,192 (*20–31*). Defining conserved immunodominant epitope regions of the S protein is of crucial importance for future anti-SARS vaccine development.

The main goal of this work was twofold: to perform mutation analysis of SARS-CoV viral genomes, with special attention to the S protein; and to group them according to different aspects of sequence similarity, eventually pointing to phylogeny and epidemiological dynamics of SARS-CoV.

Results and Discussion

Nucleotide content

Nucleotide content of SARS-CoV isolates favors T and A nucleotides. The corresponding percentages of letters in non-UTR regions of all the 96 isolates were found to be as follows: **T** (30.7940%), **A** (28.4246%), **G** (20.8121%), **C** (19.9535%), **N** (G, A, T, C; 0.0143%), **R** (Pur; 0.0005%), **K** (G or T; 0.0001%), **M** (A or C; 0.0002%), **S** (G or C; 0.0001%), **W** (A or T; 0.0002%), and **Y** (Pyr; 0.0004%). The overall ratio of (A,T)/(G,C) in the dataset was almost 3:2 (1.45). The ratio of Pur vs. Pyr nucleotides was almost 1 (0.97).

The distribution of nucleotides (nt) over sequences of length 250 nt is given in Figure S1 (Supporting Online Material). It exhibits three peak-regions of T nucleotide in the second quarter of the genome (ORF 1a), and rather stable behavior in the third quarter of the genome (ORF 1b), as also observed by Pyrc *et al* (*32*) for a group of coronaviruses (HCoV-NL63, HCoV-229E, SARS-CoV, and HCoV-OC43). Deviation of percentage of nucleotides over 250-nt blocks from the corresponding percentage in the whole dataset is given in Figure S2. Except for 3′ UTR where T nucleotide is underrepresented with even about −13%, the highest excess from the average is about +10% in four peaks, which is exhibited again by T nucleotide, three of them being between positions 7,000 and 11,000 (ORF 1a), complementary with the nucleotide A represented with −10%, and the fourth one in the S protein. Otherwise the nucleotides' offset oscillates rather regularly between −5% and +5% from the average.

Genome polymorphism

All the isolates had high degree of nucleotide identity (more than 99% pair wise). Still, they could be differentiated on the basis of their genome polymorphism, *i.e.*, the number and sites of SNVs and insertions and deletions (INDELs). Analysis of genomic polymorphism of the isolates resulted in the following two facts (Tables 1, S1, and S2). Firstly, two isolates, HSR 1 and AS, coincided with the "profile" on all the "non-empty" positions (see Materials and Methods) up to the poly-A sequence. Secondly, three isolates had large number of undefined nucleotides (N), either as contiguous segments (Sin3408 in ORFs 8a, 8b; Sin3408L in ORF 1b), or as scattered individual nucleotides or short clusters (SinP2) (Table S2). Isolate Sin3408 was the only one that has a 34-nt longer 5′ UTR as compared with the "profile". Thus these three isolates were not considered to be reliably compared with others.

Table 1 SARS-CoV Genome Polymorphism

Identification		SNVs					INDELs						Classification	
Label	ID	Total	Genes	5'/3' UTR	IGR	5' del	longIns	longDel	shortIns	shortDel	3' del	3' poly-A	Type	Group
1,2	Tor2	2	2	-/-	-	-	-	-	-	-	-	24	TTTT/TTCGT	A1
3	Urbani	5	5	-/-	-	-	-	-	-	-	-	-	TTTT/TTCGT	A1
4	CUHK-W1	9	8	-/-	1	15	-	-	-	-	-	24	CGCC/TTCAT	B1
5	BJ01	12	11	-/-	1	19	-	-	-	-	-	17	CGCC/TTCAT	B1
6	BJ02	22	22	-/-	-	-	-	-	-	-	-	18	CGCC/TTCAT	B1
7	BJ03	22	22	-/-	-	4	-	-	-	-	-	17	CGCC/TTCAT	B1
8,9	NS-1(BJ04)	15	14	-/-	1	16	-	-	-	-	-	21	TGCC/TTCGT	B1
10	GD01	49	49	-/-	-	16	29	-	-	-	-	17	CGCC/TCCAT	B2
11	HKU-39849	9	9	-/-	-	-	-	-	-	-	-	15	TTTT/TTCGT	A1
12	CUHK-Su10	2	1	-/-	1	15	-	-	-	-	-	24	TTTT/TTCGG	A1
13	Sin2500	2	2	-/-	-	16	-	-	-	-	-	-	TTTT/TTCGT	A1
14	Sin2679	2	2	-/-	-	16	-	-	-	-	-	-	TTTT/TTCGT	A1
15	Sin2774	4	4	-/-	-	16	-	-	-	-	-	-	TTTT/TTCGT	A1
16	Sin2677	3	3	-/-	-	16	-	-	-	1x6	-	-	TTTT/TTCGT	A1
17	Sin2748	1	1	-/-	-	16	-	-	-	1x5	-	-	TTTT/TTCGT	A1
18	Frankfurt 1	7	7	-/-	-	-	-	-	-	-	-	-	TTTT/TTCGT	A1
19	FRA	7	7	-/-	-	-	-	-	-	-	-	13	TTTT/TTCGT	A1
20	ZJ01	23	23	-/-	-	14	-	-	7x1	2x1	3	-	TTTT/TTCGT	B4
21	SZ3	54	53	-/1	-	15	29	-	-	-	-	-	CGCC/TCCAT	B2
22	SZ16	55	55	-/-	-	15	29	-	-	-	10	-	CGCC/TCCAT	B2
23	GZ50	11	10	-/1	-	15	-	-	-	-	-	8	TGCC/TTCAT	B1
24	GD69	21	21	-/-	-	-	-	-	1x1,1x10	-	-	16	TTTT/CTTGG	A1
25	TWC	2	2	-/-	-	-	-	-	-	1x2	-	-	TTTT/TTCGT	A1
26	HSR 1	0	0	-/-	-	-	-	-	-	-	-	24	TTTT/TTCGT	A1
27	Taiwan TC1	4	4	-/-	-	69	-	-	-	-	85	-	TTTT/CTTGG	A1
28	Taiwan TC2	9	9	-/-	-	69	-	-	-	-	85	-	TTTT/CTTGG	A1
29	Taiwan TC3	7	6	-/-	1	69	-	-	-	-	85	-	TTTT/CTTGG	A1
30,31	CUHK-AG01(02)	3	3	-/-	-	15	-	-	-	-	-	24	TTTT/CTTGG	A1
32	CUHK-AG03	5	4	-/-	1	15	-	-	-	-	-	24	TTTT/CTTGG	A1
33	PUMC01	3	2	-/-	1	13	-	-	-	-	-	24	TTTT/TTCGG	A1
34	PUMC02	2	1	-/-	1	14	-	-	-	1x2	-	27	TTTT/TTCGG	A1
35	PUMC03	4	3	-/-	1	14	-	-	-	1x3	-	35	TTTT/TTCGG	A1
36	ZMY 1	78	77	1/-	-	-	-	-	24x1	2x1,1x2	-	2	TTTT/TTCGT	B4
37,38	TWH,TWC2	4	4	-/-	-	-	-	-	-	-	-	-	TTTT/CTTGG	A1
39	TWK	7	7	-/-	-	-	-	-	-	-	-	-	TTTT/CTTGG	A1
40	TWS	6	6	-/-	-	-	-	-	-	-	-	-	TTTT/CTTGG	A1
41	TWY	6	6	-/-	-	-	-	-	-	-	-	-	TTTT/CTTGG	A1
42	TWC3	3	3	-/-	-	-	-	-	-	-	-	-	TTTT/CTTGG	A1
43	TWJ	6	6	-/-	-	-	-	-	-	1x2	-	-	TTTT/CTTGG	A1
44	GZ02	39	39	-/-	-	-	29	-	-	-	-	4	CGCC/TCCAT	B2
45	WHU	15	15	-/-	-	-	-	-	-	1x2	-	3	TTTT/TTCGT	A1
46	HZS2-D	5	5	-/-	-	15	-	-	-	-	-	24	TGCC/TTCAT	A1

Table 1 *Continued*

Identification		SNVs				INDELs							Classification	
Label	**ID**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' del**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3' del**	**3' poly-A**	**Type**	**Group**
47	HZS2-E	5	5	-/-	-	15	-	-	-	-	-	24	TGCC/TTCAT	A1
48	HZS2-Fc	6	6	-/-	-	15	-	-	-	-	-	24	TGCC/TTCGT	A1
49	HZS2-C	7	7	-/	-	15	-	-	-	-	-	24	TGCC/TTCAT	A1
50	HGZ8L2	7	7	-/-	-	15	-	-	-	-	-	24	TGCC/TTCAT	A1
51	LC1	1	1	-/-	-	15	-	-	-	-	-	24	TTTT/TTCGG	A1
52	GZ-B	3	3	-/-	-	72	-	39	-	-	-	24	TTTT/TTCGT	A3
53	GZ-C	14	14	-/-	-	52	-	39, 12	-	1x3	-	24	CTTT/TTCGT	A3
54	HSZ2-A	5	5	-/-	-	52	-	-	-	-	-	24	TGCC/TTCAT	A1
55	HZS2-Fb	5	5	-/-	-	42	-	-	-	-	-	24	TGCC/TTCGT	A1
56	HSZ-Bb	14	14	-/-	-	250	29	-	-	-	-	24	CGCC/TCCAT	B2
57	HSZ-Cb	16	16	-/-	-	51	29	-	-	-	-	24	CGCC/TCCAT	B2
58	HSZ-Bc	13	13	-/-	-	15	29	-	-	-	-	24	CGCC/TCCAT	B2
59	HSZ-Cc	19	19	-/-	-	15	29	-	-	-	-	24	CGCC/TCCAT	B2
60,61	ZS-A,ZS-B	38	38	-/-	-	15	-	53	-	-	-	24	CGCC/TCCAT	B3
62	ZS-C	38	38	-/-	-	51	-	53	-	-	-	24	CGCC/TCCAT	B3
63	LC2	4	4	-/-	-	15	-	386	-	-	-	24	TTT-/ TTCGT	A3
64,65	LC3,LC4	3	3	-/-	-	15	-	386	-	-	-	24	TTT-/ TTCGT	A3
66	LC5	4	4	-/-	-	15	-	386	-	-	-	24	TTT-/ TTCGT	A3
67	AS	0	0	-/-	-	16	-	-	-	-	-	-	TTTT/TTCGT	A1
68	SoD	30	10	1/19	-	15	-	-	-	-	-	-	TTTT/TTCGT	A1
69	ShanghaiQXC1	39	39	-/-	-	79	-	-	-	-	56	-	CGTT/TTCGT	B1
70	ShanghaiQXC2	39	39	-/-	-	79	-	579	-	-	56	-	CGTT/TTCGT	B1
71	Sino1-11	6	6	-/-	-	-	-	-	-	1x3	-	17	TTTT/TTCGG	A1
72	Sino3-11	3	3	-/-	-	-	-	-	-	1x2	-	15	TTTT/TTCGG	A1
73,74	TW2,TW1	1	1	-/-	-	-	-	-	-	-	-	2	TTTT/TTCGT	A1
75	TW3	2	2	-/-	-	-	-	-	-	-	-	2	TTTT/TTCGT	A1
76	TW4	2	2	-/-	-	-	-	-	-	-	-	2	TTTT/TTCGT	A1
77	TW5	1	1	-/-	-	-	-	-	-	-	-	2	TTTT/TTCGG	A1
78	TW6	3	3	-/-	-	-	-	-	-	-	-	2	TTTT/CTTGG	A1
79	TW7	4	4	-/-	-	-	-	-	-	-	-	2	TTTT/CTTGG	A1
80	TW8	3	3	-/-	-	-	-	-	-	-	-	2	TTTT/CTTGG	A1
81	TW9	5	4	-/-	1	-	-	-	-	-	-	2	TTTT/CTTGG	A1
82	TW10	6	5	-/-	1	-	-	-	-	-	-	2	TTTT/CTTGG	A1
83	TW11	9	8	-/-	1	-	-	-	-	1x2		2	TTTT/CTTGG	A1
84	Sin842	4	4	-/-	-	13	-	-	1x1	-	-	1	TTTT/TTCGT	A1
85	Sin852	19	9	10/-	-	1	-	57	-	-	-	1	TTT-/TTCGT	A3
86	Sin3765V	9	9	-/-	-	16	-	-	-	-	-	11	TTTT/TTCGT	A1
87	Sin848	11	11	-/-	-	16	-	-	-	-	-	2	TTTT/TTCGT	A1
88	Sin849	4	4	-/-	-	16	-	49	-	-	1	-	TTTT/TTCGT	A3
89	Sin846	7	7	-/-	-	16	-	137	2x1	-	-	1	TTTT/TTCGT	A3
90	Sin3725V	4	4	-/-	-	16	-	-	-	-	-	5	TTTT/TTCGT	A1
91	SinP1	4	4	-/-	-	16	-	-	2x1	-	-	1	TTTT/TTCGT	A1

Table 1 *Continued*

Identification		SNVs					INDELs						Classification	
Label	**ID**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' del**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3' del**	**3' poly-A**	**Type**	**Group**
92	SinP3	9	4	1/4	-	16	-	-	2x2+9x1	-	-	1	TTTT/TTCGT	A4
93	SinP5	4	4	-/-	-	16	-	-	1x2	1x1	-	1	TTTT/TTCGT	A1
94	SinP4	7	4	-/3	-	16	-	-	1x2	2x1	1	-	TTTT/TTCGT	A1
95	Sin845	10	10	-/-	-	16	-	-	-	-	-	1	TTTT/TTCGT	A1
96	Sin847	12	10	2/-	-	10	-	-	-	-	-	2	TTTT/TTCGT	A1
97	Sin850	11	6	5/-	-	8	-	-	-	-	-	1	TTTT/TTCGT	A1
98	LLJ-2004	11	10	-/-	1	21	-	-	1x6	1x1	-	5	CGCC/TTCAT	B1
99	TJF	17	10	2/1	4	1	-	-	-	-	-	19	TGCC/TTCGT	B1
100	CDC#200301157	2	2	-	-	-	-	-	-	-	-	-	TTTT/TTCGT	A1
101	Sin3408L	4	4	-	-	16	-	-	1x1	2x1	-	5	TTTT/TTCGT	A1
102	SinP2	4	4	-	-	16	-	-	13x1, 1x2	5x1+1x6		2	TTTT/TTCGT	A4
103	Sin3408	14	4	10/-	-	-	5' end (34)	-	-	-	-	6	TTTN/TTCGT	A1

Shaded entries correspond to annotated isolates. **Identification** (**Label** and **ID**) is given in accordance with the labels and identifiers from Table S1. The four **SNVs** columns correspond to: the **total** number of SNVs, the number of SNVs in **genes**, in **5′** and **3′ UTRs**, and in **IGR**. The seven columns named **INDELs** include the number of deletions at the 5′ end (**5′ del**), the length of long insertions (**longIns**) and long deletions (**longDel**), the number and length of short insertions (**shortIns**) and short deletions (**shortDel**) in the form $a \times b$ where b denotes the length and a denotes the number of occurrences, the number of deletions at the 3′ end (**3′ del**), and the length of a poly-A sequence at the 3′ end (**3′ poly-A**). **Classification** includes two columns. The **Type** column corresponds to the nine-locus nucleotides that are given in the form NNNN/NNNNN and represent nucleotides at (relative to CLUSTAL X output) positions 9,420, 17,604, 222,274, 27,891 / 3,861, 9,495, 11,514, 21,773, 26,534, respectively (absolute HSR 1 positions 9,404, 17,564, 22,222, 27,827 / 3,852, 9,479, 11,493, 21,721, 26,477). The last column, **Group**, reflects grouping of isolates.

Nucleotide variations: single nucleotide polymorphism

There were 446 SNV sites and 1,006 SNVs in total in the dataset, with the substitution rate 1.49%, which is about three times higher (both the number of SNVs and the substitution rate) than the corresponding findings (*33*) for 17 isolates. An average number of SNVs per isolate was 10.48, giving an error rate of 3.6×10^{-4} substitutions per nucleotide copied.

There was only one site with multiple base substitutions (the original nucleotide base on that position being T): at the relative (CLUSTAL X) position 8,441 (ORF 1a), isolate ZMY 1 has the nucleotide C (absolute position 8,403), and isolates ShanghaiQXC1, ShanghaiQXC2 have the nucleotide A (absolute positions 8,312 and 7,733, respectively).

The smallest distance between the two neighboring SNV sites in the whole dataset was 1; the largest one was 23,988 (in case of TW3 and TW1), while an average distance between the neighboring SNV sites in the whole dataset was 1,987 positions (Figure S3). The distribution of isolates per SNV number (outside 5′, 3′ UTRs) showed regularity for up to 11 SNVs (almost Gaussian distribution) and irregular decrease for number of SNVs >11 (Figure S4). Thus the number of SNVs less than or equal to 11 per isolate was considered as a "small" number of SNVs, and the number of SNVs greater than 11 was considered as a "large" number of SNVs. Most SNVs are clustered within two regions in ORF 1a and one region at the 3′ end of the viral genome that predominantly consists of small ORFs, leaving two small regions within ORF 1a, and a region that corresponds to ORF 1b as the most conservative ones (Figure 1B).

The entropy of each genome nucleotide position was calculated, showing that the most conserved sites are the ones with the smallest entropy and that the least conserved sites are the ones with the highest entropy (*34*; Figure S5). The nine loci used for classification can be found among the sites with the highest entropy.

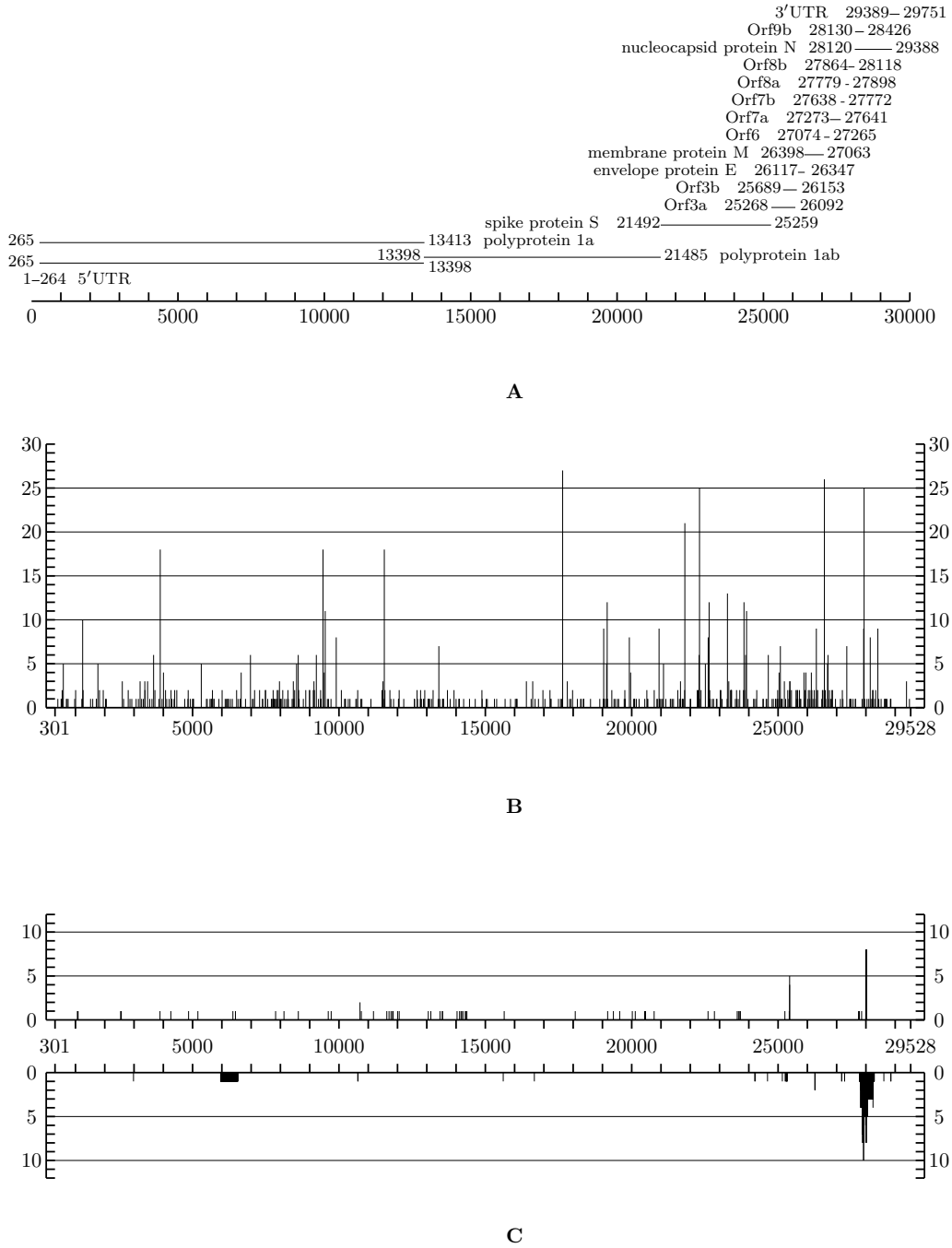


Fig. 1 Density distribution of SNVs (B), INDELs (C), mapped onto the gene map of the HSR 1 isolate, coinciding with the "profile" (A). Central region of the genome is rather conserved (lower density of SNVs is exhibited in the second third of the genome, ORF 1b), while the rest of the genome features high SNVs density. SNV peaks are present at (absolute HSR 1) positions 3,852, 9,404, 9,479, 11,493, 17,564 (ORF 1ab), 21,721, 22,222 (S protein), 26,477 (M protein), and 27,827 (ORF 8a).

Percent of each category of substitution is given in Figure 2. There are 141 transversion sites and 306 transition sites, *i.e.*, 31.54%:68.46%, with 261 transversions (2.72 in average per isolate) and 745 transitions (7.76 in average per isolate).

Length variations, insertions and deletions

Analysis of the SARS-CoV genome showed that long INDELs were concentrated close to the 3′ end (except for the 579-nt deletion in the ShanghaiQXC2 isolate at the position 5,834, located in ORF 1a), while indi-

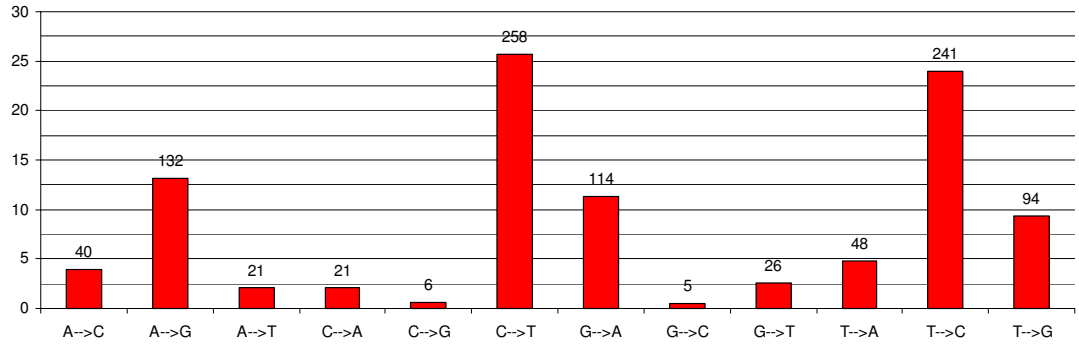


Fig. 2 Distribution of nucleotide substitution categories. The most represented are the substitutions C↔T and the least represented are the substitutions C↔G.

vidual insertions were found along the whole genome, most in the second quarter, and individual deletions were quite rare. Density distribution of INDELs inside the SARS-CoV genome, and 5′ UTR, 3′ UTR length variations, are represented in Figures 1C, S6A and B, respectively. Figure S6C represents the region of the genome between positions 27,700 and 28,300 (ORFs 7b, 8a, 8b, part of N-protein, in HSR 1 annotation), which is especially abundant with INDELs. While individual INDELs are present both in longer and shorter ORFs, longer INDELs are (except for previously mentioned deletions in the ShanghaiQXC2) all located in short ORFs.

Figure 3 represents comparison results of genome primary structure of the analyzed isolates, summarizing the following facts:

Firstly, although the SARS-CoV genome has the established length of 29,727 nt (*12*), most isolates were shorter at the 5′ end (for the first 15 positions, majority of isolates were "empty"), and had various length "poly-A" strings at the 3′ end, or both (Table 1). Several isolates had some short deletions inside the sequence, *e.g.*, Sin2677, Sin2748, TWC, PUMC02, PUMC03, TWJ, WHU, Sino1-11, Sino3-11, TW11, and SinP5.

Secondly, there was a group of isolates that had insertions of length 29 nucleotides (GD01, SZ3, SZ16, GZ02, HSZ-Bb, HSZ-Bc, HSZ-Cb, and HSZ-Cc) at the relative position 27,995 (absolute position 27,869 in SZ3, SZ16, HSZ-Cc, and HSZ-Bc; protein BGI-PUP GZ29-nt-Ins, ORF 8a). Two of them were isolates from palm civet (SZ3 and SZ16) and the other six were isolates from human patients. This specific insertion is also treated as a deletion in all the other isolates, evolved from this early group (*10*).

Thirdly, there were several groups of isolates that had long deletions: GZ-B, GZ-C (length 39 at the relative position 27,882, or absolute position 27,719 in GZ-C, ORF 7b), ZS-A, ZS-C (length 53 at the relative position 27,969, absolute 27,843 in ZS-A, ORF 8a), LC2, LC3, LC5 (length 386 at the relative position 27,829, absolute 27,704 in LC2, ORFs 7b, 8a, 8b), ShanghaiQXC2 (length 579 at the relative position 5959, absolute 5834, ORF 1a), Sin852, Sin849, and Sin846 (of length 57, 49, 137, respectively, at relative positions in region between 27,787 and 27,966, ORFs 7b, 8a) (Tables 1 and S2, Figure 3).

Fourthly, a large number of individual INDELs were identified in ZJ01, ZMY 1, SinP2, and SinP3 (Tables 1 and S2, Figure 3).

Mutation analysis

While the distribution of nucleotides over different distances from SNV sites did not exhibit any regularities, the distribution of different nucleotides on distance 1 left to SNV sites (−1) did exhibit significant difference from their overall percentage in the dataset. The corresponding right (+1) distance distribution of nucleotides is almost uniform (Table 2).

Figure S7 represents differences between the percentage of nucleotides at a given position and in the whole genome, for up to the distance 10 left and right from SNV sites. Figures S8A and B represent distribution of substitutions preceded by different nucleotide bases, and followed by different nucleotide bases, respectively. It can be seen that on the C↔T substitutions, both C→T and T→C, are favored by the preceding A and the following T (almost 40% of all the C↔T substitutions; Figure 2), while the substitution T→C is almost prohibited by the preceding T (only 3%). Clustered substitutions of length 2 are rare (*e.g.*, TC→AA, GA→CA).

♦ = *aaattttgtt tcatctatac aggaaaagcc aacc*
♠ = *ttaatt*
♠ = *caagaatgta*
❣ = ---...--- (66 deletions)
✗ = ---...--- (71 deletions)

A

B

Fig. 3 Comparison of nucleotide structures of SARS-CoV complete genome isolates, represented in parts A and B of the figure according to similarity in their SNVs or INDELs positions.

Table 2 Distribution of Nucleotides on Distance 1 Left and Right to SNV Sites

Nt	(−1)num	(−1)%	(−1)diff%	(+1)num	(+1)%	(+1)diff%
A	358	35.59%	7.17%	283	28.13%	−0.29%
C	179	17.79%	−2.16%	203	20.18%	0.23%
G	230	22.86%	2.05%	215	21.37%	0.56%
T	238	23.66%	−7.13%	302	30.02%	−0.77%

The distribution of nucleotides on distance 1 left to SNV sites (−1) and right to SNV sites (+1) is presented in total number of nucleotides, percentage, and difference from their overall percentage in the dataset.

Codon usage

Analysis of distribution of individual nucleotides over the three codon positions in annotated ORFs of all the annotated isolates showed that, except for short proteins such as E, M, and presumptive ORFs, all the codons exhibit the same tendency of T nucleotide dominating at the third codon position, and the G nucleotide dominating at the first codon position, while A and C appearing more often at the second codon position than elsewhere. Figure S9 represents distribution of nucleotides over the three codon positions in individual ORFs, and in total.

Analysis of codon usage demonstrated the same facts as the distribution of nucleotides over the three codon positions. In total, the third nucleotide favored T (40.10%) over A (24.83%), C (18.90%), and G (16.16%). It was especially true for four-codon families a.a. (Thr, Pro, Ala, Gly, and Val). The same held for four-codon subsets of six-codon families (Arg, Leu, and Ser), differring at the third codon position only. The above was true for the ORF 1ab, S and N proteins, but not for another two structural proteins (E and M). The codon usage for SARS-CoV genome proteins is represented by Table S3, and it is consistent with the results obtained for another human CoV genome, HCoV-NL63 (*32*).

Changes in amino acids

Besides the number of SNVs, isolates differed in positions of SNVs, too. Table S4 represents positions where two or more SNVs occurred, for all the annotated isolates, along with nucleotides and ORFs (based on HSR 1 annotation), type of mutation (transition/transversion), a.a. position in ORF, a.a. change, a.a. property change, and nucleotide position in codon. Positions of multiple SNVs have been chosen in order to reduce the chance of erroneously determined SNV. There were 91 such SNV sites with 288 SNVs. It is interesting to notice that there were no multiple base substitutions (more than two different bases) in any of these positions. There were 227 transitions at 75 sites and 61 transversion at 16 sites, out of which 5 were in structural proteins: 2 in S, 2 in E, and 1 in M proteins. The most common mutation was C↔T mutation (45 sites or 50%), followed by A↔G (30 sites), A↔T and G↔T (7 sites each), and C↔A (2 sites). There was no mutation of the type C↔G.

There were 28 SNV sites corresponding to the first codon position (20 transitions and 8 transversions), 2 of which representing silent mutation sites (C↔T, Leu). There were 33 SNV sites corresponding to the second codon position (31 transitions and 2 transversions), all of which cause a.a. change. There were 30 SNV sites corresponding to the third codon position (25 transitions and 5 transversions), 29 of which representing silent mutation sites (the only non-silent one is G↔T, Leu ↔ Phe).

There were 31 synonymous multiple substitution sites and 60 non-synonymous ones, with substitution rate 0.31% (91/29,228) and non-synonymous substitution rate 0.21%, which is consistent with the corresponding findings for 17 SASR-CoV isolates (*33*). The number of multiple substitutions was for about 30% lower than the number of the overall substitutions, and so were the substitution rate and non-synonymous substitution rate.

Table S5 summarizes the above findings. It represents the number of transition and transversion sites and the number of SNVs (in the form N_1/N_2) per position in codon and per mutation type, as well as the percentage of SNVs, and the number of silent mutation sites and silent SNVs.

Concerning non-synonymous sites, 35 are within pp 1ab, 5 within ORF 3, 1 within E protein, 3 within M protein, 1 within ORF 6, 1 within ORF 8a, 1 within ORF 8b, 1 within N protein, and 11 within S protein

(only for two-or-more substitution sites, and only in annotated isolates).

Mutation analysis of the S protein

The S protein is of particular interest for mutation analysis, being the key for host range determination. Multiple sequence alignment of the S protein in all the 96 SARS-CoV isolates showed that five of them, namely ZMY 1, SinP2, SinP3, SinP4, and Sin3408L, had large discrepancies with all the others due to individual insertions or deletions in them. Since such significant mismatches in the S protein sequence seemed to be the result of erroneous sequencing, we eliminated these five isolates and analyzed the S protein in the remaining 91 isolates.

There were 34 isolates without SNVs in the S protein: TW2-TW11, Sino3-11, AS, LC1, WHU, TWC3, PUMC01-PUMC03, CUHK-AG01, CUHK-AG3, Taiwan TC1-3, TWC, Sin2748, Sin2500, Sin2677, CUHK-Su10, HKU-39849, TWH, TWJ, TWK, TWY, and HSR 1. There were 62 SNV sites with 208 SNVs in total, and no multiple mutations. Table S6 represents SNV sites and all the SNVs in the S protein of the 91 isolates, along with nucleotides, type of mutation (transition/transversion), a.a. position in the protein, a.a. change, a.a. properties change, nucleotide position in codon, and number of SNVs at each SNV site. These findings overlap with the results reported in Song *et al* (Ref. *7*; concerning SNVs with multiple occurrences, in 103 S protein genes, some of which being nucleotide-identical, with 80% in common with our dataset), and are consistent on the intersecting data. Table 3 summarizes the results from Table S6.

Table 3 Mutation Analysis of the S Protein: Categories of Nucleotide Substitutions

			1.pos	2.pos	3.pos	Total No.		1.pos%	2.pos%	3.pos%	Total%	Silent
Transitions	A-G	A→G	6/15	2/8	3/20	11/43	16/73	7.21%	3.85%	9.62%	20.68%	3/20
		G→A	2/7	2/22	1/1	5/30		3.37%	10.58%	0.48%	14.43%	1/1
	C-T	C→T	3/7	6/25	6/19	15/51	24/94	3.37%	12.02%	9.13%	24.52%	6/19
		T→C	2/3	3/29	4/11	9/43		1.44%	13.94%	5.29%	20.67%	4/11
	Total		13/32	13/84	14/51		40/167	15.38%	40.38%	24.52%	80.28%	14/51
Transversions	A-C	A→C	2/2	1/1	2/2	5/5	8/10	0.96%	0.48%	0.96%	2.40%	2/2
		C→A	1/1	1 / 2	1 / 2	3/5		0.48%	0.96%	0.96%	2.40%	0
	A-T	A→T	0	0	0	0	6/7	0	0	0	0	0
		T→A	2/2	1/1	3 / 4	6/7		0.96%	0.48%	1.92%	3.37%	1/1
	G-C	G→C	1/1	0	0	1/1	3/5	0.48%	0	0	0.48%	0
		C→G	0	2/4	0	2/4		0	1.92%	0	1.92%	0
	G-T	G→T	0	1/1	1/1	2/2	5/19	0	0.48%	0.48%	0.96%	1/1
		T→G	2/14	0	1/3	3/17		6.73%	0	1.44%	8.17%	1/3
	Total		8/20	6/9	8/12		22/41	9.62%	4.33%	5.77%	19.72%	5/7
Total			21/52	19/93	22/63		62/208	25.00%	44.71%	30.29%	100%	19/58

S proteins in 91 isolates are considered. The number of transition and transversion sites and the number of SNVs (in the form N_1/N_2) per position in codon and per mutation type, as well as the percentage of SNVs, and the number of silent mutation sites and silent SNVs (in the form N_1/N_2), are presented.

Out of 62 SNV sites, 19 were observed to be synonymous, with 58 synonymous SNVs in total, and 43 were observed to be non-synonymous substitution sites, with 150 non-synonymous SNVs in total (Table S6). Substitution rate was 1.65% (62/3,768) and non-synonymous substitution rate was 1.14% (43/3,768), which is consistent with findings for the whole genome in the enlarged dataset, and is about three times higher than the corresponding findings for 17 isolates in Bi *et al* (Ref. *33*; 22 substitution sites, 13 non-synonymous, substitution rates 0.58, 0.35, respectively). As represented on Figure 4, most non-synonymous a.a. substitutions are located in external domain (ED); 14 of non-synonymous substitution are in RBD, 3 of them in the most narrow intersecting range. As it concerns epitopes, 40 of non-synonymous

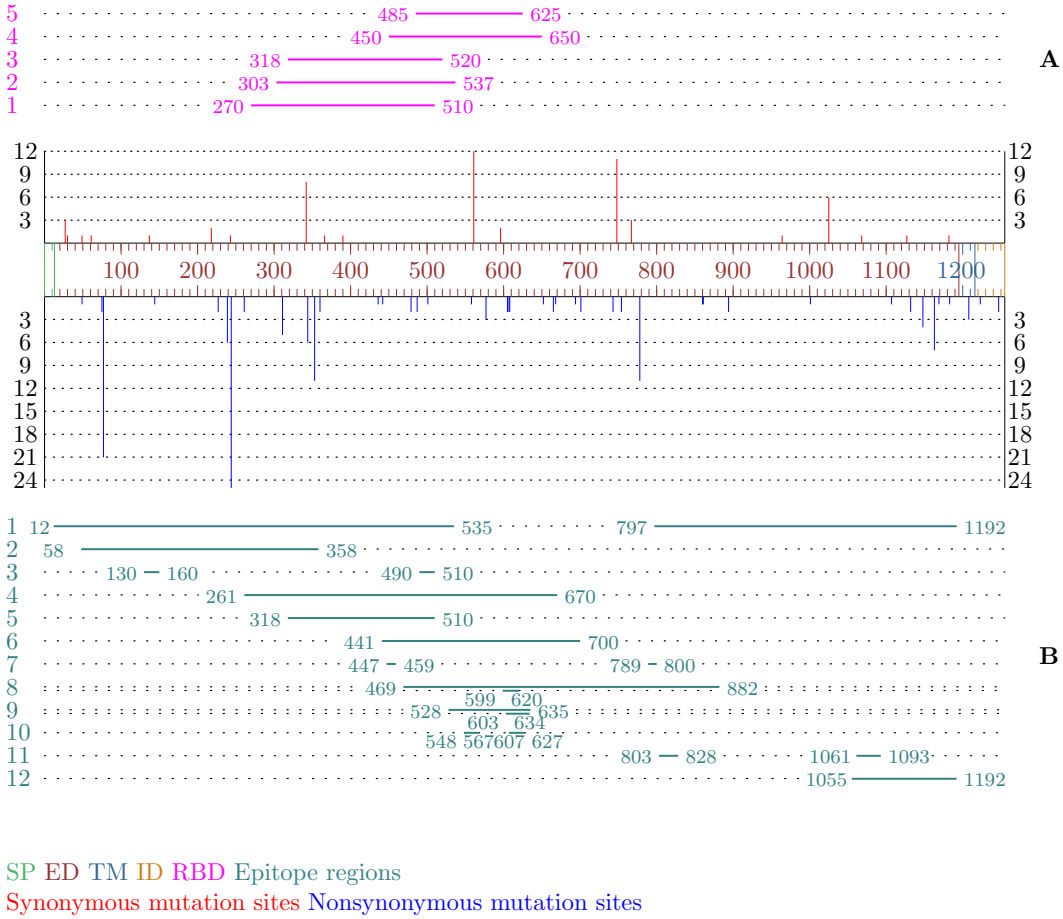


Fig. 4 Positions of synonymous and non-synonymous a.a. substitutions plotted against S protein primary structure. The y-axis represents number of SNVs per positions. SP, signal peptide; ED, external domain; TM, trans-membrane domain; and ID, internal domain (http://expasy.org/). **A**. RBD determined by: 1. Babcock *et al* (*16*), 2. Xiao *et al* (*17*), 3. Wong *et al* (*18*), 4. Zhao *et al* (*19*), and 5. Zhou *et al* (*20*); **B.** epitope regions determined by: 1. Wang *et al* (*31*), 2. Chou *et al* (*29*), 3. Greenough *et al* (*30*), 4. Sui *et al* (*26*), 5. van den Brink *et al* (*28*), 6. Lu *et al* (*24*), 7. Hua *et al* (*23*), 8. Ren *et al* (*21*), 9. He *et al* (*22*), 10. Zhou *et al* (*20*), 11. Zhang *et al* (*27*), and 12. Keng *et al* (*25*).

a.a. substitutions are located in overall epitope domains determined by various researchers. Finally, one non-synonymous a.a. substitution is located in trans-membrane domain (TM) and two in internal domain (ID).

The coefficients Ks (number of synonymous substitutions per synonymous site) and Ka (number of non-synonymous substitutions per non-synonymous site) were calculated for all the 91 S proteins in the dataset to be Ks = 0.00135, Ka = 0.00103, and the ratio Ka/Ks had the value $0.7629 < 1$, which may be interpreted as evidence for the S protein not being subjected to the Darwinian selection. These findings are consistent with the similar analysis performed for 20 SARS-CoV isolates by Hu *et al* (*35*) giving the ratio value of 0.65. Values of the corresponding coefficients and the ratio Ka/Ks for the 89 human patient isolates only, present even stronger evidence of the S protein being subject to negative selection: Ks = 0.00121, Ka = 0.00080, Ka/Ks = 0.661. The coefficients Ka, Ks, and the Ka/Ks ratio for all the human patients' isolates and each of the palm civet isolates as the outgroup, are represented in Table 4. These values indicated that the S gene was subjected to the Darwinian selection during virus evolution (transmission from animals to humans), which is consistent with the analysis performed by Yeh *et al* (*36*), for 28 human isolates and the SZ3 palm civet as the outgroup, giving the corresponding ratio value of 1.657, and with the analysis performed by He *et al* (*8*), indicating that the S gene showed the strongest positive selection pressures initially, with eventual stabilization.

Table 4 Mutation Analysis of the S Protein: Coefficients Ka, Ks, and the Ratio Ka/Ks with An Outgroup

Outgroup	Ka	Ks	Ka/Ks
SZ16 (AY304488)	0.006257	0.004930	1.26935>1
SZ3 (AY304486)	0.005889	0.003803	1.54856>1

Coefficients Ka, Ks are calculated for all the human patients' isolates and one of the palm civet isolates as an outgroup.

Phylogenetic analysis

Phylogenetic tree, drawn using the PhyloDraw program for the CLUSTAL X output of aligning the 96 isolates, is represented in Figure 5. Its close relationship to the classification proposed in the paper suggests that classification of SARS-CoV isolates might be obtained by applying the computational analysis based on genome polymorphism.

All the isolates were classified according to their genome polymorphism—SNVs and INDELs, the procedure being proposed in our previous paper (*37*). Since SNV contents turned out to be a more distinguishable property than the presence of INDELs, as the *first classification criterion* we took the number and positions of SNVs. For the "profile" isolate, as the referent isolate, number of SNVs for different isolates varied from 0 (HSR 1, AS) to 78 (ZMY 1) (Table 1). All the isolates were classified into two groups based on the number of SNVs with the "profile"—those with less than or equal to 11 SNVs, and those having more than 11 SNVs. Thus, the first classification criterion resulted in two groups (Table 1):

Group A—isolates with less than or equal to 11 SNVs (Tables 1 and S2);

Group B—isolates with more than 11 SNVs relative to the "profile" isolate.

Positions of SNVs moved several isolates between the two groups (SoD from B to A, since the most of its SNVs are in 3′ UTR; CUHK-W1 from A to B, since its number of SNVs with the "profile" of the A group is larger than the one of the B group; WHU from B to A; GZ50 from A to B; GD69 from B to A; and GZ-C from B to A).

The *second classification criterion* was presence and position of long INDELs inside the basic A, B groups. We identified the following subgroups:

A2, B2—subgroups of the A, B groups, respectively, with long insertions. A2 remained empty, while B2 contained 8 isolates with 29-nt insertions.

A3, B3—subgroups of the A, B groups, respectively, with long deletions. A3 group consisted of 3-isolate subgroup (LC2, LC3, and LC5) with a deletion of length 386, Sin852 with a deletion of length 57, 2-isolates subgroup (GZ-B and GZ-C) with a deletion of length 39, Sin849 (deletion of length 49, embedded), and Sin846 (137, overlapping). B3 subgroup consisted of the isolates ZS-A (ZS-B) and ZS-C with the deletion of length 53.

A4 and **B4** were the subgroups with many individual INDELs (Table 1). The rests of the A, B groups were denoted as **A1** and **B1**, respectively.

It can be noted that proposed grouping of 96 isolates, based on SNV and INDEL contents, conserved the earlier classification T-T-T-T/C-G-C-C (*38*), and partially coincided with the extension of this classification (*39*, *40*). The four loci (9,404, 17,564, 22,222, and 27,827), as the basis for this classification, fitted nicely into our grouping (basically A1 group coincided with T-T-T-T type, while B1 group coincided with C-G-C-C type), expressing two inter-types: T-G-C-C [isolates GZ50, HZS2-D, HZS2-E, HZS2-C, HGZ8L2, HSZ2-A, NS-1(BJ04), HZS2-Fc, HZS2-Fb, and TJF] and C-G-T-T (isolates ShanghaiQXC1 and ShanghaiQXC2) (Figure 5). We found that another five loci, which are among the most represented SNVs' loci (positions 3,852, 9,479, 11,493, 21,721, and 26,477; Figure 1), further refined our classification providing for sub-classification of the basic types.

There were two basic nine-locus types: TTTT/TTCGG and CGCC/TTCAT, mostly coinciding with the A1, B1 groups, and the two inter-groups: an inter-(A-B)-group had the inter-type TGCC/TTCGT, and a subgroup of the group B1 (two Shanghai isolates) represented another inter-type CGTT/TTCGT (Figure 5). The proposed extension to the two main sequence variants (TTTT, CGCC) for an enlarged set of isolates, is in accordance with the new insights into possible epidemiological spread, both in space and time (*36, 38, 41*). Namely, positions 3,852 and 11,493 differentiated between the two subgroups of the group A1 (all of the TTTT type): Taiwan epidemic (nucleotides C, T) from the other late epidemic isolates (nucleotides T, C) (*41*), *i.e.*, isolates closer to a Hong Kong virus unrelated to Hotel M (nucleotides C, T: isolates TW6-TW10, Taiwan TC1-TC3), and the others from the Hotel M lineage [nucleotides T, C: isolates from Canada (Tor2), Singapore (all Sin's), Frankfurt (FRA Fr 1), Taiwan (TW1-TW5), Hong Kong (HKU 39849), Italy

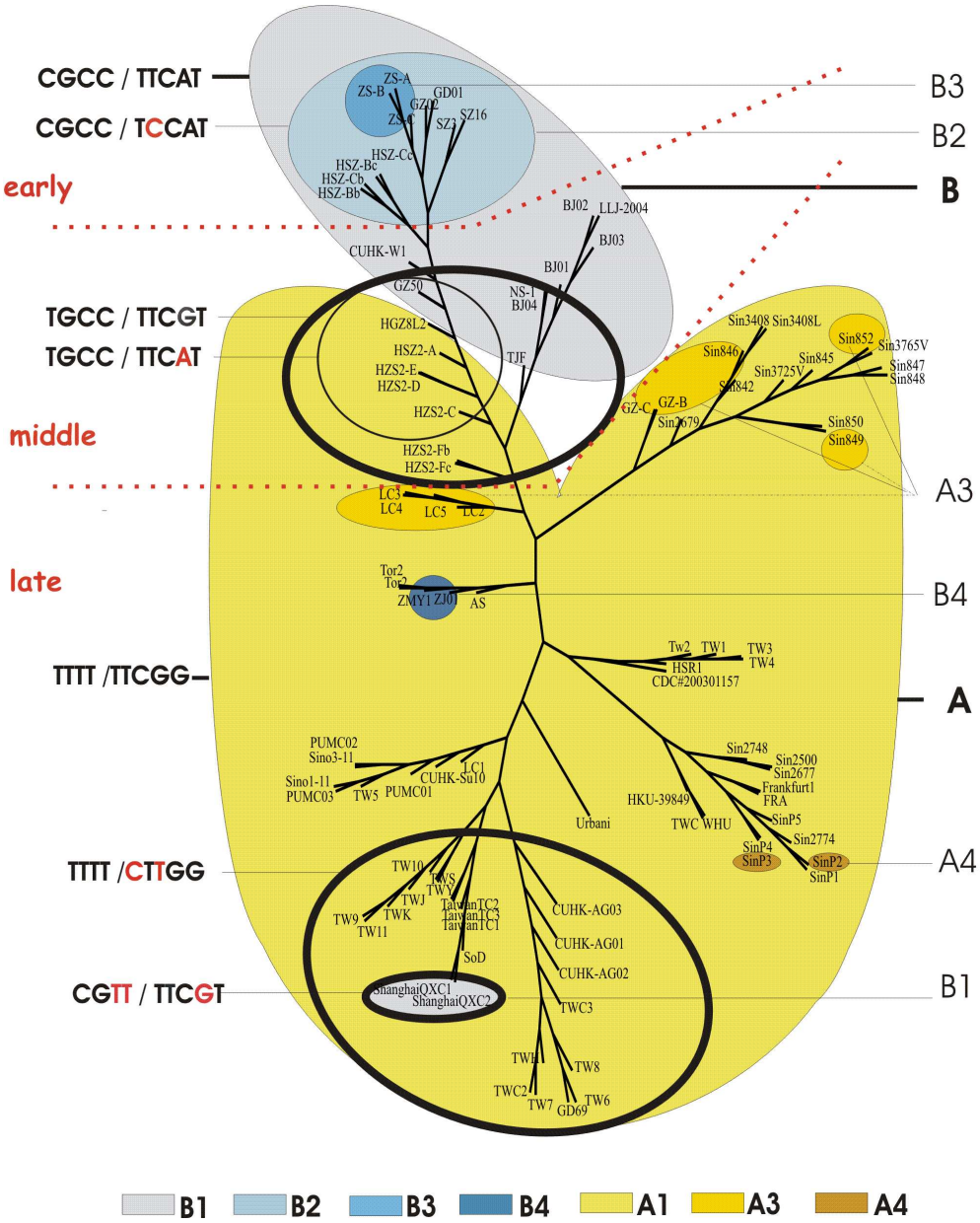


Fig. 5 Three-level classification of 103 SARS-CoV genome isolates. Grouping of isolates is based on genome polymorphism, and classification is based on nine distinguished loci, mapped onto the bootstrapped phylogenetic tree obtained using CLUSTAL X and Neighbor Joining method, and drawn using PhyloDraw programs. Bootstrapping is performed with random number generated seed 111 and number of trials in bootstrap 1000. The two basic groups, A and B, are represented in yellow and blue, respectively. Types obtained according to the nine genome loci (9,404, 17,564, 22,222, 27,827 / 3,852, 9,479, 11,493, 21,721, 26,477) are labeled along the left edge of the figure and have the form NNNN / NNNNN, where N represents any nucleotide. Different subtypes are denoted by the corresponding substituted nucleotides in red. Dotted lines distinguish between the three epidemiological phases.

(HSR 1), China (ZJ01), *etc.*] (*36*). Position 9,479 decomposed the B group [differentiated subgroup B1 (T) from the subgroups B2, B3 (C)], position 21,721 distinguished the group A from the group B. Precise characterization based on the nine loci, for all the isolates, is given in Table 1 and Figure 5.

As compared to genotype clustering of SARS-CoV covering the epidemics from 2002 to 2004 (*7*, *8*), it can be noticed that the grouping we proposed was at most in accordance with it. Namely, the following correspondence between the two grouping schemes may be established:

Firstly, genotype class CGCC/TCCAT (covering B2 and B3 subgroups), corresponded to human patients' isolates from the early phase 2002–2003 (ZS, HSZ, GD01, GZ02—Guangzhou, China), and palm civet isolates (SZ3, SZ16—Hong Kong).

Secondly, genotype class TGCC/TTCGT, TGCC/TTCAT (small part of A1 group), as well as CGCC/TTCAT (B1 group), corresponded to human patients middle phase 2002–2003 (positions 3,852, 9,479, 11,493, 26,477 determined this subclass); Beijing (BJ01–BJ04), and Hong Kong (CUHK W1).

Thirdly, genotype classes TTTT/NNNNN, CGTT/NNNNN (almost the entire A group and Shanghai part of the B1 group) corresponded to human patients late phase 2002–2003 (Figure 5)—Singapore (Sin s), Taiwan (TW1-11), Shanghai (QX1, QX2), Italy (HSR 1), Canada (Tor2), Hanoi (Urbani), Hong Kong (HKU39849, CUHK-AG0x), China (ZJ01, WHU, PUMC0x), Frankfurt (FRA, Frankfurt 1), *etc.*

The two basic groups, A and B, were rather contiguously mapped onto the phylogenetic tree, showing a high degree of accordance among the proposed grouping and the phylogenetic relationships. Exceptions represented the two isolates of the B4 group, with large number of SNVs and individual insertions (ZMY 1, ZJ01), as well as the two isolates of the B1 group (Shanghai QXC1 and QXC2), all of which being at large root-distances (Figure S10).

Materials and Methods

Dataset

Nucleotide sequences of 103 SARS-CoV complete genomes were taken from the PubMed NCBI Entrez database (http://www.ncbi.nlm.nih.gov/entrez) in GenBank and FASTA formats. Since there were 7 pairs of nucleotide-identical isolates, we considered the dataset to consist of 96 isolates (Tables 1 and S1). All the sequences are between 29,013 (ShanghaiQXC2) and 29,767 (Sin3408) in length. Out of all, 19 sequences have ambiguous nucleotide codes, *i.e.*, N, M, R, Y, W, K, S (Table S1). Out of 96 isolates, 42 are fully or partially annotated. All the annotated isolates have the S protein annotated of length 3,768 and the N protein of length 1,269 nucleotides, 4 isolates do not have the E protein (CUHK-Su10, PUMC01, PUMC02, PUMC03) of length 231 (except for Sino1-11 of length 228 nucleotides), 1 isolate does not have the M protein (CUHK-Su10) of length 666 nucleotides. Out of 42 annotated isolates, 13 have 5′ UTR and 12 have 3′ UTR determined. All the isolates are human sourced except for two isolated from palm civet (*Paguma larvata*), SZ3 and SZ16.

Genome polymorphism

The CLUSTAL X program, version 1.83 (*42*) has been applied to all the isolates from the dataset. The overall CLUSTAL X output had length of 29,903 nt. Then 5′ UTR and 3′ UTR were identified based on positions in annotated isolates. Coding region encompassed the interval (301, 29,528), and had the length of 29,228 nt.

We developed a program in Perl language for analysis of a CLUSTAL X output. The program first calculated an "average" isolate, the so called "profile", by counting, for each position in the CLUSTAL X output, the number of occurrences of each different letter (including dash), and by choosing the most represented one; positions containing dashes in the "profile" are called "empty positions", all the others being "non-empty" ones. The program then counted SNVs, INDELs, and calculated their absolute and relative positions, for every isolate with respect to the "profile", and for different genome regions (ORFs, 5′ UTR, 3′ UTR, and IGRs).

Substitution rate for the SARS-CoV genome and for the S protein for all the sequences in the dataset was calculated by dividing the total number of SNV sites by the length of the corresponding nucleotide sequence; non-synonymous substitution rate for the S protein was calculated by dividing the total number of non-synonymous SNV sites by the length of the S protein.

Entropy of sites

The entropy of each site has been calculated based on number of SNVs at that site, in order to estimate the sites' conserveness. If we denote by p(b)—probability of occurrence of the nucleotide b (b being A, C, G, or T), and under assumption of sites being independent, we calculated the entropy of positions by the following formula (*43*): E = − sum p(b)* log[p(b)] (sum over b). In this definition, p(b)* log[p(b)] is taken to be zero if p(b) = 0.

Phylogenetic investigations

The first type of classification was performed the same way as in Pavlovic-Lazetic *et al* (*37*). It is based on genome polymorphism (SNVs and INDELs). The distribution of isolates per SNV numbers (outside 5′, 3′

UTRs) was analyzed and the isolates were primarily classified into two groups—isolates with "small" number of SNVs and isolates with "large" number of SNVs. The isolates "close to border" were further tested (on the number of SNVs) against the profile isolates of each of the two groups, resulting in some isolates changing the group. A sub-classification was then performed on the presence of long or short INDELs inside each of the two groups.

The second type of classification was performed based on contents of the most represented SNV sites. Except for earlier identified positions (9,404, 17,564, 22,222, 27,827) classifying isolates into TTTT/CGCC genotypes (*38*, *39*), some other positions (genotypes) were identified as potential bases for subclassification.

In order to compare the two classification schemes developed, with the existing programming systems for phylogenetic analysis, phylogenetic bootstrapped tree was produced using CLUSTAL X program and the Neighbor Joining (NJ) method. The NJ method, as well as parsimony and the probabilistic models, produces unrooted trees. In order to produce the consensus tree, bootstrapping is performed with random number generated seed 111 and number of trials in bootstrap 1000. The tree is drawn using the PhyloDraw program (*44*) and the proposed classification schemes were mapped onto it.

Annotation and analysis of the S protein

All the S protein sequences (those extracted from annotated isolates and the others we annotated using the publicly available program from PubMed tools for data mining; http://www.ncbi.nlm.nih.gov/gorf/gorf.html) have been aligned using CLUSTAL X program. Then the S protein was analyzed using the same methods as for the complete isolates.

Mutation analysis of the S protein

Non-synonymous nucleotide substitution per non-synonymous site (Ka) and synonymous nucleotide substitution per synonymous site (Ks) were calculated using the DnaSP 4.0 program (*45*). It is based on a method defined by Nei and Gojovori (*46*) that estimates the numbers of synonymous and non-synonymous nucleotide substitutions between two DNA sequences by counting the number of such substitutions in the corresponding pairs of codons. It also takes into account different evolutionary pathways between pairs of codons. The DnaSP program may run with or without an outgroup. The ratio Ka/Ks is considered as a selection parameter (Ka/Ks $>$ 1 is usually interpreted as an indicator of positive selection). The coefficients Ka, Ks, as well as the ratio Ka/Ks were calculated first for the S protein in all the isolates in the dataset, without an outgroup. Since among the 91 isolates there were 89 human patients' isolates and 2 palm civet isolates (SZ3, SZ16), we then calculated the Ka and Ks coefficients and the ratio Ka/Ks for the 89 human patients' isolates only, without an outgroup, too. Eventually, we ran the program for all the human patients' isolates and each of the palm civet isolates as the outgroup, in order to test the hypothesis that the S gene was subjected to positive selection during virus transmission from animals to humans.

Acknowledgements

This work was supported by the Ministry of Science and Technology, Republic of Serbia, Project No. 1858.

References

1. Peiris, J.S., *et al.* 2004. Severe acute respiratory syndrome. *Nat. Med.* 10: S88-97.
2. Fouchier, R.A., *et al.* 2003. Aetiology: Koch's postulates fulfilled for SARS virus. *Nature* 423: 240.
3. Rota, P.A., *et al.* 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394-1399.
4. Marra, M.A., *et al.* 2003. The genome sequence of the SARS-associated coronavirus. *Science* 300: 1399-1404.
5. Guan, Y., *et al.* 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 302: 276-278.
6. Stavrinides, J. and Guttman, D.S. 2004. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J. Virol.* 78: 76-82.
7. Song, H.D., *et al.* 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc. Natl. Acad. Sci. USA* 102: 2430-2435.
8. He, J.F., *et al* (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.
9. Stadler, K., *et al.* 2003. SARS—beginning to understand a new virus. *Nat. Rev. Microbiol.* 1: 209-218.

10. Chiu, R.W., *et al.* 2005. Tracing SARS-coronavirus variant with large genomic deletion. *Emerg. Infect. Dis.* 11: 168-170.
11. Vega, V.B., *et al.* 2004. Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. *BMC Infect. Dis.* 4: 32.
12. Ziebuhr, J. 2004. Molecular biology of severe acute respiratory syndrome coronavirus. *Curr. Opin. Microbiol.* 7: 412-419.
13. Groneberg, D.A., *et al.* 2005. Molecular mechanisms of severe acute respiratory syndrome (SARS). *Respir. Res.* 6: 8.
14. Tan, Y.J., *et al.* 2005. Characterization of viral proteins encoded by the SARS-coronavirus genome. *Antiviral Res.* 65: 69-78.
15. Li, W., *et al.* 2003. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 426: 450-454.
16. Babcock, G.J., *et al.* 2004. Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with receptor. *J. Virol.* 78: 4552-4560.
17. Xiao, X., *et al.* 2003. The SARS-CoV S glycoprotein: expression and functional characterization. *Biochem. Biophys. Res. Commun.* 312: 1159-1164.
18. Wong, S.K., *et al.* 2004. A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2. *J. Biol. Chem.* 279: 3197-3201.
19. Zhao, J.C., *et al.* 2005. Prokaryotic expression, refolding, and purification of fragment 450-650 of the spike protein of SARS-coronavirus. *Protein Expr. Purif.* 39: 169-174.
20. Zhou, T., *et al.* 2004. An exposed domain in the severe acute respiratory syndrome coronavirus spike protein induces neutralizing antibodies. *J. Virol.* 78: 7217-7226.
21. Ren, Y., *et al.* 2003. A strategy for searching antigenic regions in the SARS-CoV spike protein. *Geno. Prot. Bioinfo.* 1: 207-215.
22. He, Y., *et al.* 2004. Identification of immunodominant sites on the spike protein of severe acute respiratory syndrome (SARS) coronavirus: implication for developing SARS diagnostics and vaccines. *J. Immunol.* 173: 4050-4057.
23. Hua, R., *et al.* 2004. Identification of two antigenic epitopes on SARS-CoV spike protein. *Biochem. Biophys. Res. Commun.* 319: 929-935.
24. Lu, L., *et al.* 2004. Immunological characterization of the spike protein of the severe acute respiratory syndrome coronavirus. *J. Clin. Microbiol.* 42: 1570-1576.
25. Keng, C.T., *et al.* 2005. Amino acids 1055 to 1192 in the S2 region of severe acute respiratory syndrome coronavirus S Protein induce neutralizing antibodies: implications for the development of vaccines and antiviral agents. *J. Virol.* 79: 3289-3296.
26. Sui, J., *et al.* 2004. Potent neutralization of severe acute respiratory syndrome (SARS) coronavirus by a human mAb to S1 protein that blocks receptor association. *Proc. Natl. Acad. Sci. USA* 101: 2536-2541.
27. Zhang, H., *et al.* 2004. Identification of an antigenic determinant on the S2 domain of the severe acute respiratory syndrome coronavirus spike glycoprotein capable of inducing neutralizing antibodies. *J. Virol.* 78: 6938-6945.
28. van den Brink, E.N., *et al.* 2005. Molecular and biological characterization of human monoclonal antibodies binding to the spike and nucleocapsid proteins of severe acute respiratory syndrome coronavirus. *J. Virol.* 79: 1635-1644.
29. Chou, C.F., *et al.* 2005. A novel cell-based binding assay system reconstituting interaction between SARS-CoV S protein and its cellular receptor. *J. Virol. Methods* 123: 41-48.
30. Greenough, T.C., *et al.* 2005. Development and characterization of a severe acute respiratory syndrome-associated coronavirus-neutralizing human monoclonal antibody that provides effective immunoprophylaxis in mice. *J. Infect. Dis.* 191: 507-514.
31. Wang, S., *et al.* 2005. Identification of two neutralizing regions on the severe acute respiratory syndrome coronavirus spike glycoprotein produced from the mammalian expression system. *J. Virol.* 79: 1906-1910.
32. Pyrc, K., *et al.* 2004. Genome structure and transcriptional regulation of human coronavirus NL63. *Virol. J.* 1: 7.
33. Bi, S., *et al.* 2003. Complete genome sequences of the SARS-CoV: the BJ group (Isolates BJ01–BJ04). *Geno. Prot. Bioinfo.* 1: 180-192.
34. Mooney, S.D. and Klein, T.E. 2002. The functional importance of disease-associated mutation. *BMC Bioinformatics* 3: 24.
35. Hu, L.D., *et al.* 2003. Mutation analysis of 20 SARS virus genome sequences: evidence for negative selection in replicase ORF1b and spike gene. *Acta Pharmacol. Sin.* 24: 741-745.
36. Yeh, S.H., *et al.* 2004. Characterization of severe respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution. *Proc. Natl. Acad. Sci. USA* 101: 2542-2547.
37. Pavlovic-Lazetic, G.M., *et al.* 2004. Bioinformatics analysis of SARS coronavirus genome polymorphism. *BMC Bioinformatics* 5: 65.
38. Ruan, Y.J., *et al.* 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* 361: 1779-1785.
39. Chim, S.S., *et al.* 2004. Genomic sequencing of a

SARS coronavirus isolate that predated the Metropole Hotel case cluster in Hong Kong. *Clin. Chem.* 50: 231-233.

40. Wang, Z.G., *et al.* 2004. Molecular biological analysis of genotyping and phylogeny of severe acute respiratory syndrome associated coronavirus. *Chin. Med. J* (*Engl.*) 117: 42-48.
41. Lan, Y.C., *et al.* 2005. Phylogenetic analysis and sequence comparison of structural and non-structural SARS coronavirus proteins in Taiwan. *Infect. Genet. Evol.* 5: 261-269.
42. Thompson, J.D., *et al.* 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24: 4876-4882.
43. Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory.* John Wiley & Sons, Inc., New York, USA.
44. Choi, J.H., *et al.* 2000. PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics* 16: 1056-1058.
45. Rozas, J., *et al.* 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496-2497.
46. Nei, M. and Gojovori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotde substitutions. *Mol. Biol. Evol.* 3: 418-426.

Supporting Online Material

http://www.gpbjournal.org/journal/pdf/GPB3(1)-04.pdf
Figures S1–S10, Tables S1-S6

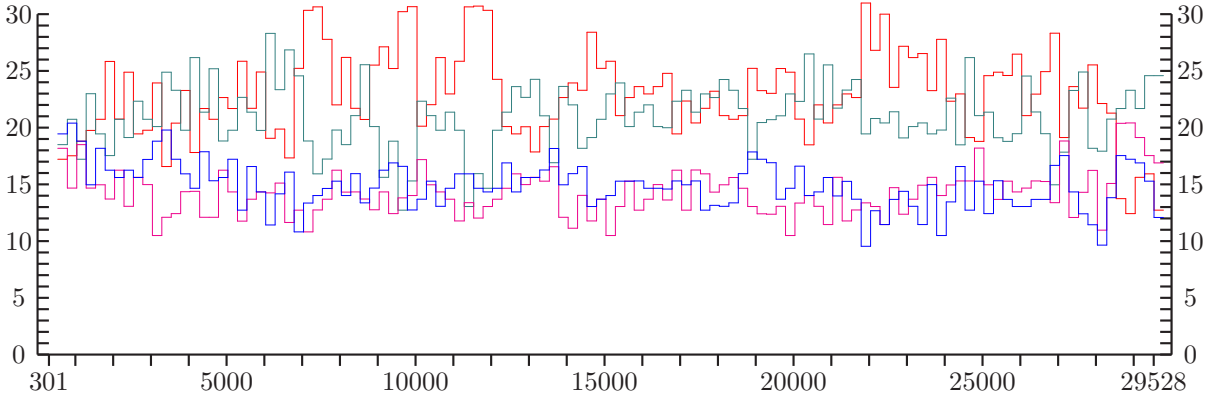


Figure S1. Percentage of nucleotides in 250-nt segments of the SARS-CoV genome

nucleotide G nucleotide T nucleotide C nucleotide A

Non-UTR regions of all the isolates are considered; positions (x-axis) are on the ClustalX output scale; percentage of nucleotides in the corresponding blocks are represented on y-axis

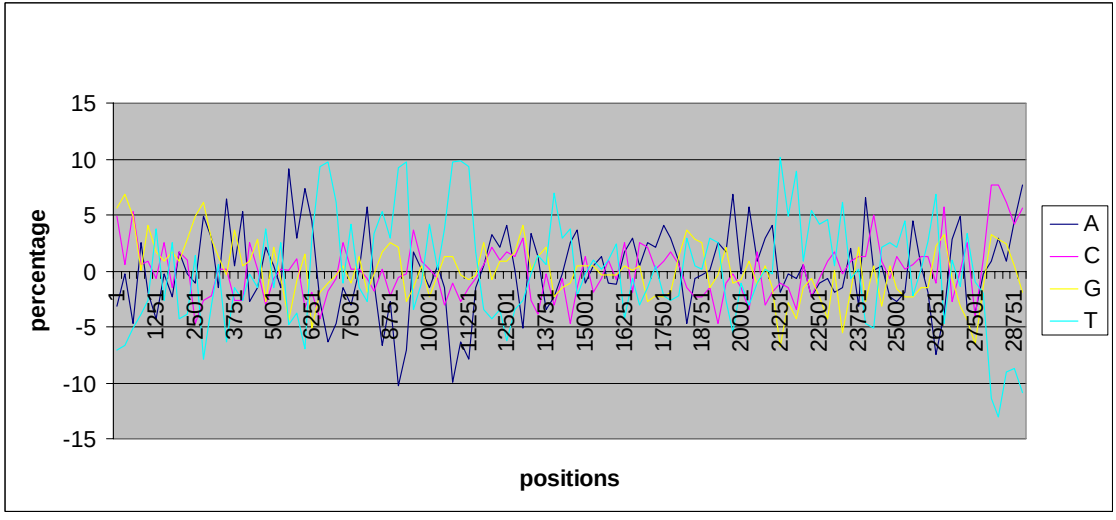


Figure S2. Deviation of percentages of nucleotides over 250-nt blocks from the corresponding percentages in the whole dataset

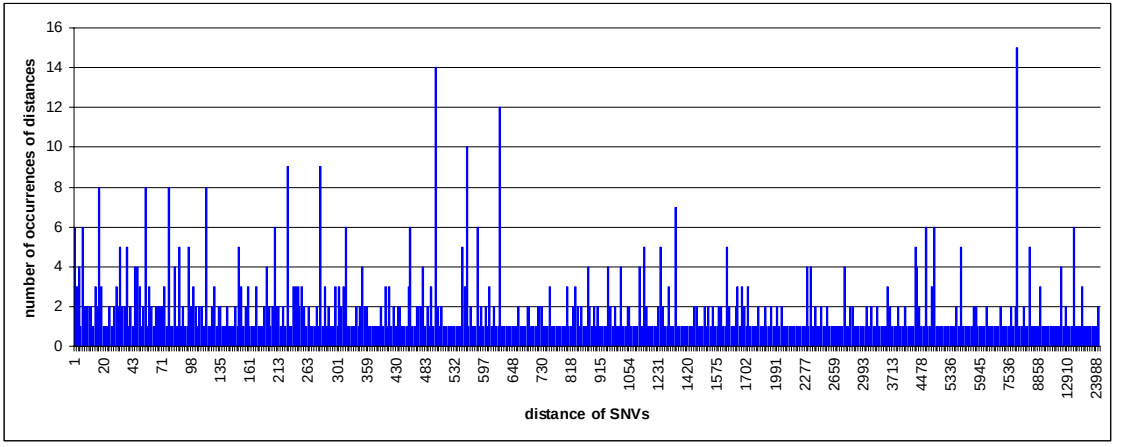


Figure S3. Distribution of distances between the neighboring SNVs, and numbers of their occurrences

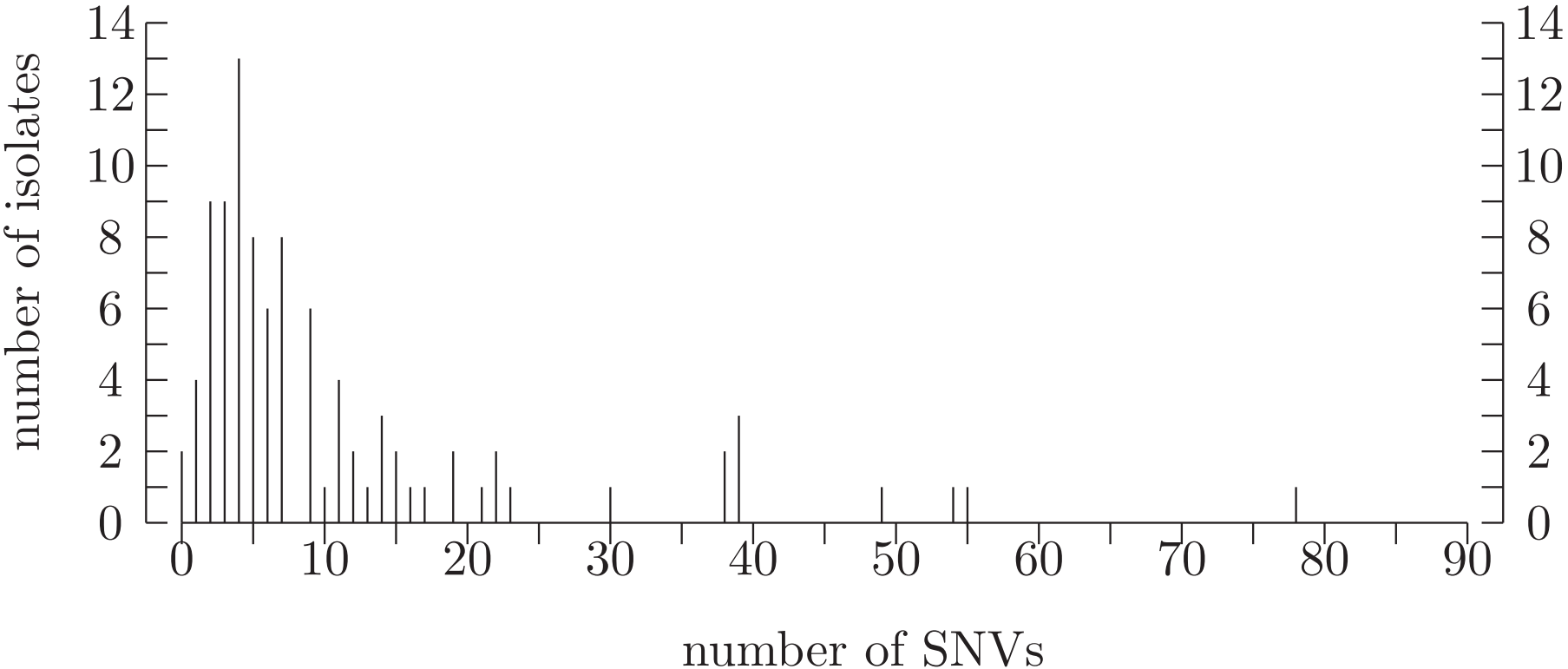


Figure S4. Distrubution of isolates per number of SNVs

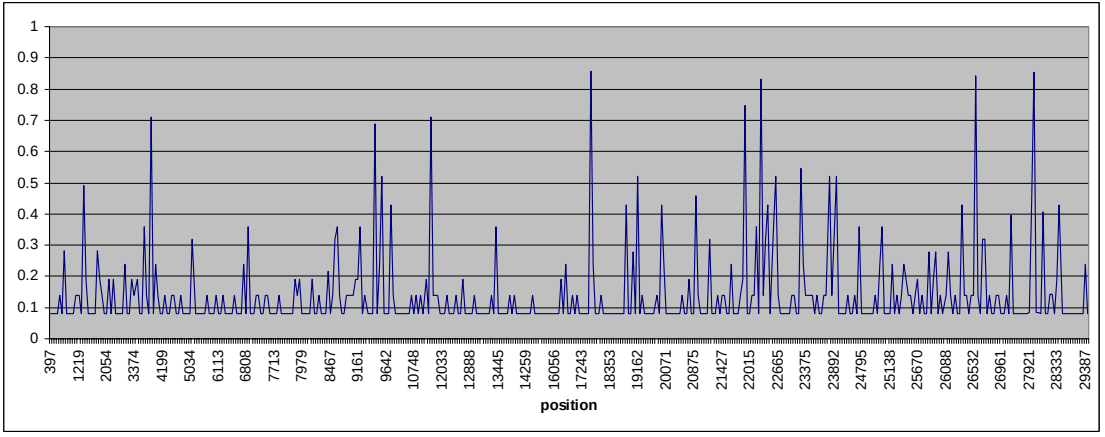


Figure S5. Entropy of genome nucleotide positions

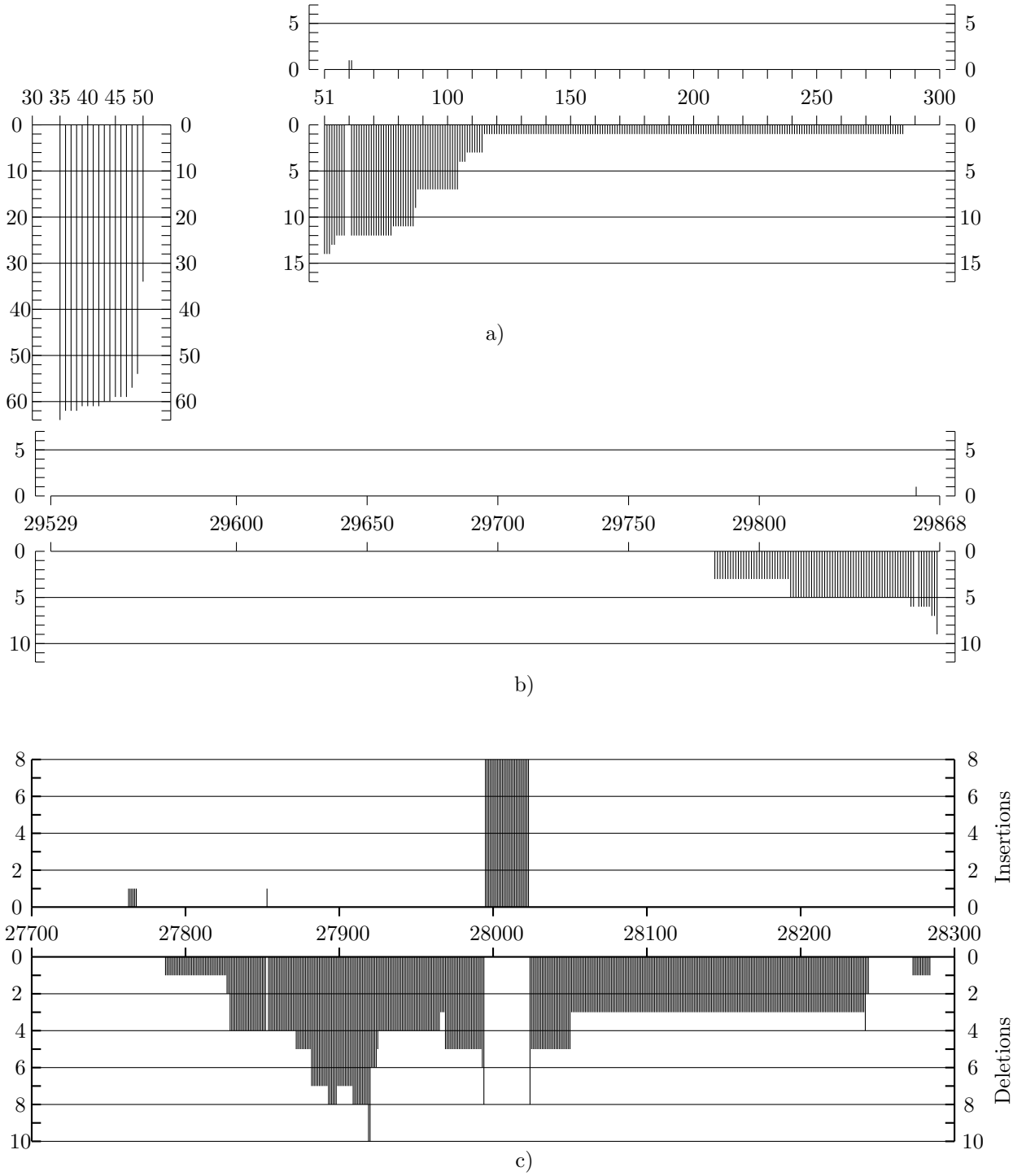
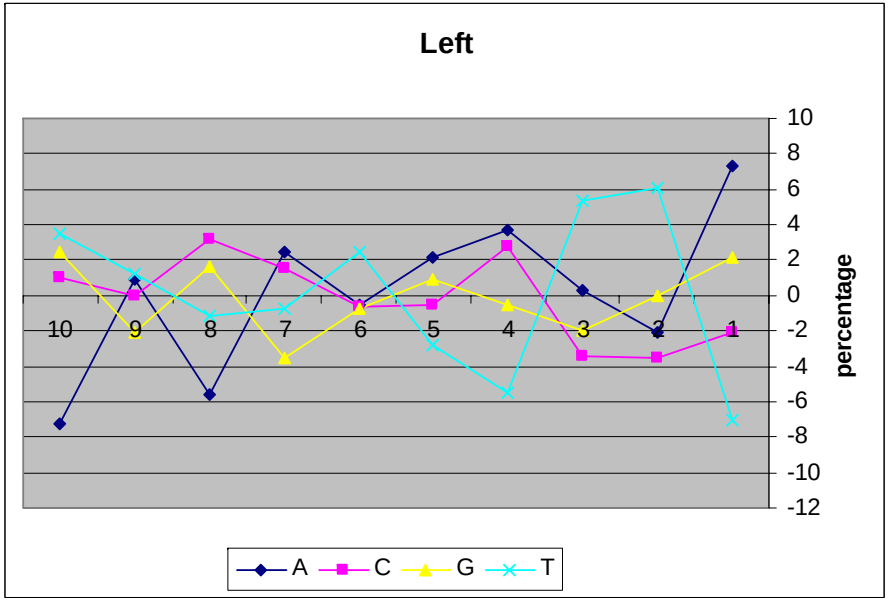


Figure S6. Distribution of INDELs in 5'UTR (a), 3'UTR (b) and in interval 27700-28300 (c)



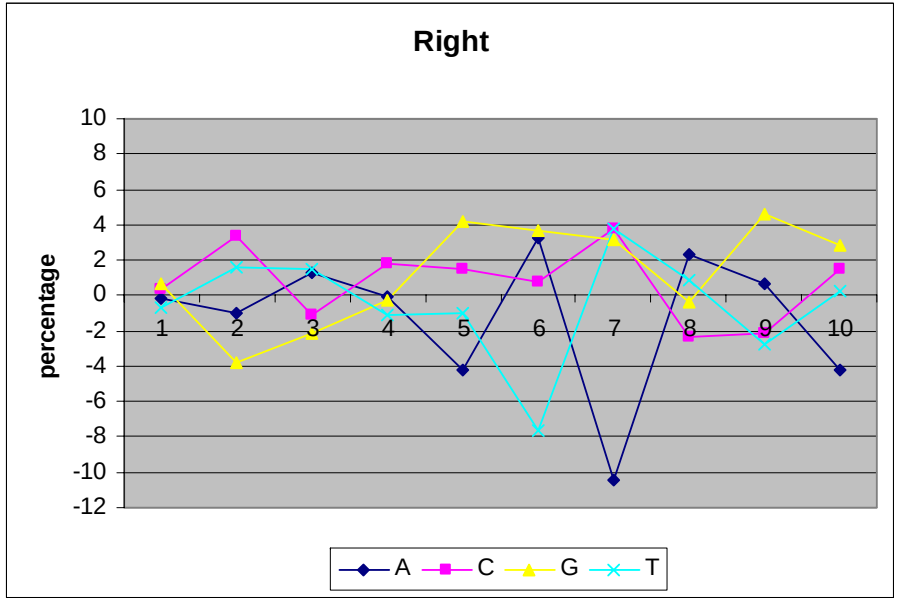
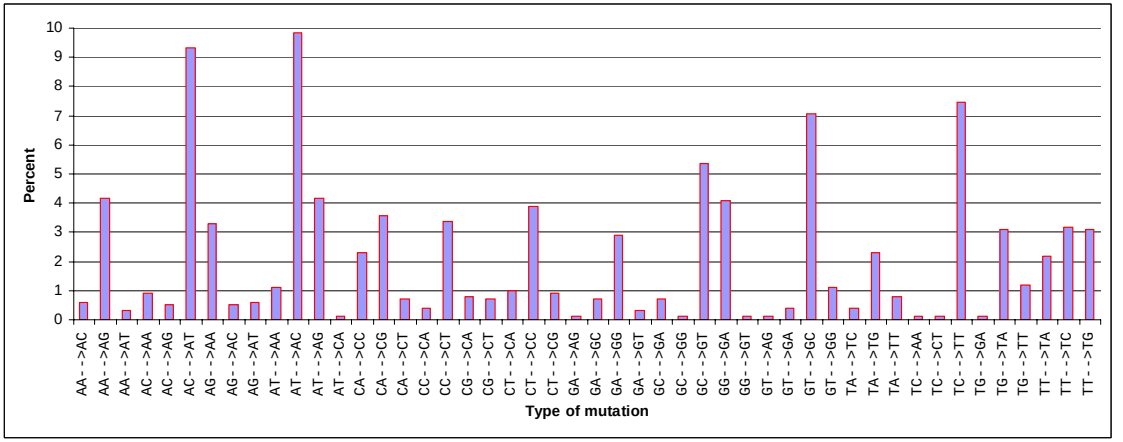
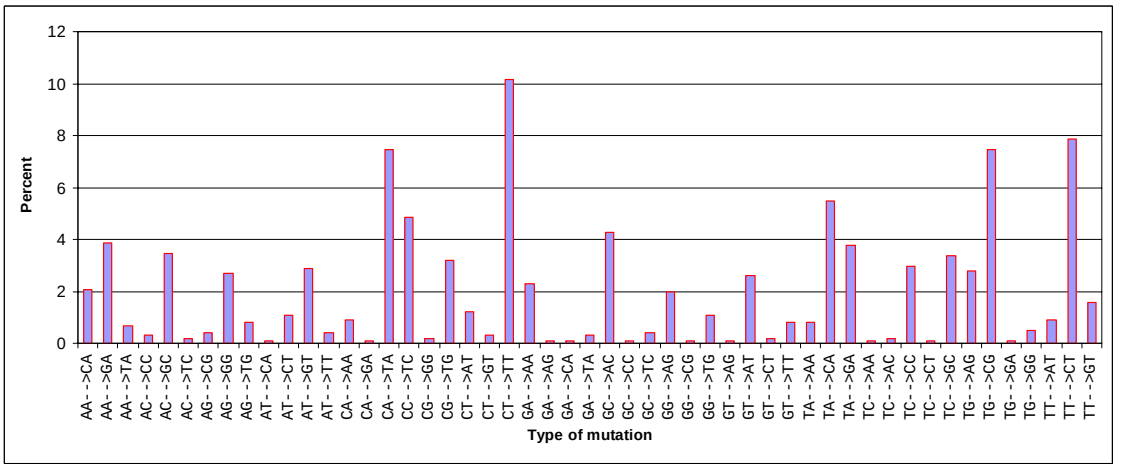


Figure S7. Differences between the percentage of nucleotides at a given position and in the whole genome, for up to the distance 10 left and right from SNV sites

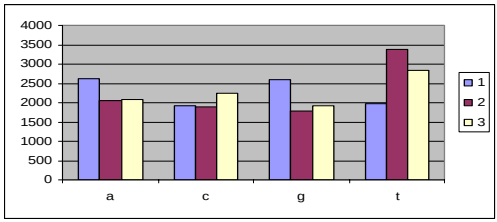


a)

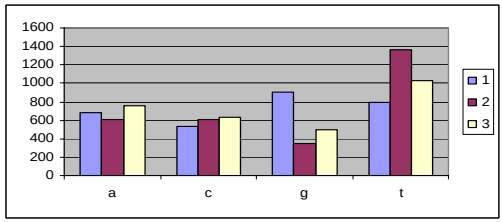


b)

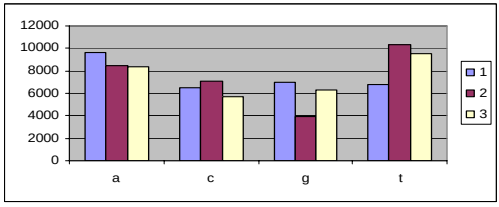
Figure S8. Distribution of substitutions preceded by different nucleotide bases (a) and followed by different nucleotide bases (b)



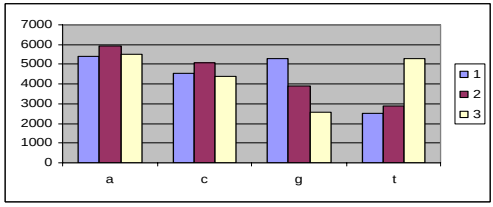
Envelope protein



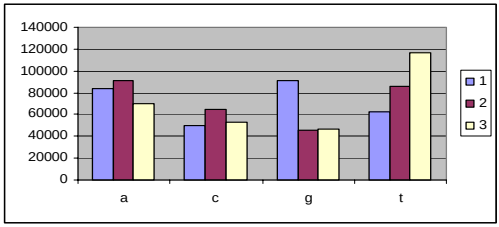
Membrane protein



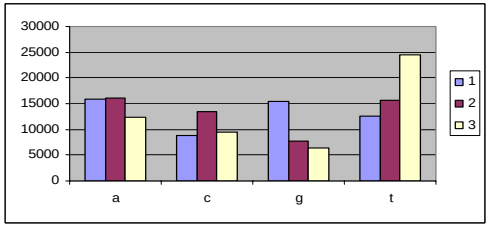
Unknown (putative) orfs



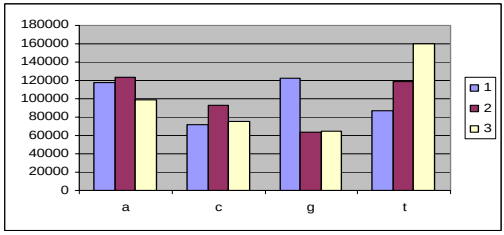
Nucleocapside protein



orf1ab



Spike protein



Total in proteins

Figure S9. Distribution of nucleotides over the three codon positions in specific proteins and in total

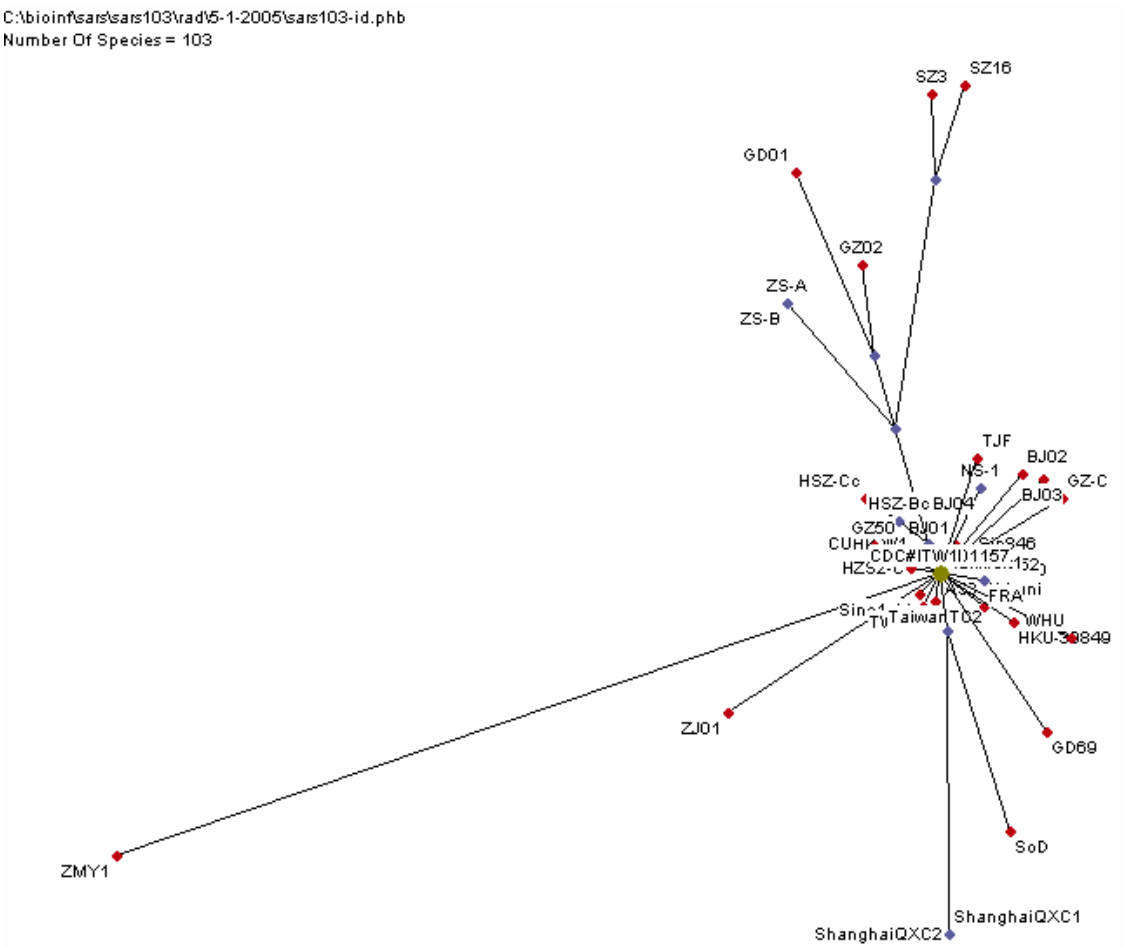


Figure S10. Root distances for isolates from the dataset

The isolates ZMY1, ZJ01, and the two Shanghai isolates have the largest root distance in the A group (the south side of the picture), and thus may not belong to it. (The isolates SoD and GD69 also have large root distances, but thay have been moved into the A group on the basis of their SNV sites)

Table S1. List of all the isolates considered. It includes identifiers, accession numbers, revision dates, country and source of the isolates considered, labels to be referred to in this paper, as well as the length of isolates and the ambiguous nucleotide codes. The labels are assigned in the (approximate) order of the first submission of isolates

Label	ID	Accession No.	Length	Revision date	Country/Source	GI	FirstSubmitted	No. ambigBases
1	Tor2	NC_004718.3	29751	30-SEP-2004	Canada: Toronto, patient #2	30271926	13-APR-2003	
2		AY274119.3		24-MAR-2004	Canada: Toronto, patient #2	30248028	23-APR-2003	
3	Urbani	AY278741.1	29727	12-AUG-2003	USA: Atlanta cell_line="Vero"	30027617	17-APR-2003	
4	CUHK-W1	AY278554.2	29736	31-JUL-2003	China: Hong Kong	30027610	17-APR-2003	
5	BJ01	AY278488.2	29725	01-MAY-2003	China: Beijing	30275666	17-APR-2003	
6	BJ02	AY278487.3	29745	05-JUN-2003	China: Beijing	31416292	17-APR-2003	
7	BJ03	AY278490.3	29740	05-JUN-2003	China: Beijing	31416305	17-APR-2003	
8	BJ04	AY279354.2	29732	05-JUN-2003	China: Beijing	31416306	19-APR-2003	
9	NS-1	AY508724.1		17-JAN-2004		40795744	18-DEC-2003	
10	GD01	AY278489.2	29757	18-AUG-2003	China: Beijing	31416290	17-APR-2003	
11	HKU-39849	AY278491.2	29742	29-AUG-2003	China: Hong Kong	30023963	18-APR-2003	
12	CUHK-Su10	AY282752.2	29736	17-NOV-2003	China: Hong Kong	30421451	24-APR-2003	
13	Sin2500	AY283794.1	29711	12-AUG-2003	Singapore	30468042	27-APR-2003	
14	Sin2679	AY283796.1	29711	12-AUG-2003	Singapore	30468044	27-APR-2003	
15	Sin2774	AY283798.2	29711	02-OCT-2003	Singapore	37361915	27-APR-2003	
16	Sin2677	AY283795.1	29705	12-AUG-2003	Singapore	30468043	27-APR-2003	
17	Sin2748	AY283797.1	29706	12-AUG-2003	Singapore	30468045	27-APR-2003	N=1
18	Frankfurt 1	AY291315.1	29727	16-MAR-2004	Germany: Frankfurt	31581502	06-MAY-2003	
19	FRA	AY310120.1	29740	12-DEC-2003	Germany: patient from Frankfurt	33578015	29-MAY-2003	
20	ZJ01	AY297028.1	29715	19-MAY-2003	China: Beijing	30910859	12-MAY-2003	M=1
21	SZ3	AY304486.1	29741	05-NOV-2003	China: Hong Kong	34482137	26-MAY-2003	
22	SZ16	AY304488.1	29731	05-NOV-2003	China: Hong Kong	34482139	27-MAY-2003	
23	GZ50	AY304495.1	29720	05-NOV-2003	China: Hong Kong	34482146	27-MAY-2003	
24	GD69	AY313906.1	29754	15-DEC-2003	China: Jiangmen, Guangdong	37960831	03-JUN-2003 isolated in May 2003	
25	TWC	AY321118.1	29725	26-JUN-2003	Taiwan, first fatal case	31873092	11-JUN-2003	
26	HSR 1	AY323977.2	29751	15-OCT-2003	Italy	33115118	16-JUN-2003	R=1
27	Taiwan TC1	AY338174.1	29573	28-JUL-2003	Taiwan	32493129	08-JUL-2003	
28	Taiwan TC2	AY338175.1	29573	28-JUL-2003	Taiwan	32493130	09-JUL-2003	
29	Taiwan TC3	AY348314.1	29573	29-JUL-2003	Taiwan	33188324	23-JUL-2003	
30	CUHK-AG01	AY345986.1	29736	29-NOV-2003	Hong Kong: patient #1 of the Amoy Gardens cohort	33114190	18-JUL-2003	
31	CUHK-AG02	AY345987.1		29-NOV-2003	patient #2 of the Amoy Gardens cohort	33114202	18-JUL-2003	
32	CUHK-AG03	AY345988.1	29736	29-NOV-2003	Hong Kong: patient #3 of the Amoy Gardens cohort	33114214	18-JUL-2003	
33	PUMC01	AY350750.1	29738	17-NOV-2003	China: Beijing	38231927	24-JUL-2003	
34	PUMC02	AY357075.1	29738	17-NOV-2003	China: Beijing	38231932	31-JUL-2003	
35	PUMC03	AY357076.1	29745	17-NOV-2003	China: Beijing	38231937	31-JUL-2003	
36	ZMY 1	AY351680.1	29749	03-AUG-2003	China: Guangdong	33304219	28-JUL-2003	
37	TWH	AP006557.1	29727	02-AUG-2003	Taiwan: patient #01 – cultured	33411399	30-JUL-2003	
38	WC2	AY362698.1		13-AUG-2003	Taiwan: Hoping Hospital	33518724	05-AUG-2003	
39	TWK	AP006559.1	29727	02-AUG-2003	Taiwan: patient #06 - primary	33411429	30-JUL-2003	
40	TWS	AP006560.1	29727	02-AUG-2003	Taiwan: patient #04 - primary	33411444	30-JUL-2003	
41	TWY	AP006561.1	29727	02-AUG-2003	Taiwan: patient #02 - primary	33411459	30-JUL-2003	
42	TWC3	AY362699.1	29727	13-AUG-2003	Taiwan: Hoping Hospital – throat swab	33518725	05-AUG-2003	
43	TWJ	AP006558.1	29725	02-AUG-2003	Taiwan: patient #043 Primary	33411414	30-JUL-2003	
44	GZ02	AY390556.1	29760	31-JAN-2004	China: Guangzhou	41323719	15-SEP-2003	
45	WHU	AY394850.2	29728	07-JUN-2004	China: Wuhan	40795428	18-SEP-2003	
46	HZS2-D	AY394989.1	29736	29-JAN-2004	China: Guangzhou	37624332	19-SEP-2003	
47	HZS2-E	AY394990.1	29736	29-JAN-2004	China: Guangzhou	37624333	19-SEP-2003	
48	HZS2-Fc	AY394991.1	29736	29-JAN-2004	China: Guangzhou	37624334	19-SEP-2003	
49	HZS2-C	AY394992.1	29736	29-JAN-2004	China: Guangzhou	37624335	19-SEP-2003	
50	HGZ8L2	AY394993.1	29736	29-JAN-2004	China: Guangzhou	37624336	19-SEP-2003	
51	LC1	AY394998.1	29736	29-JAN-2004	China: Guangzhou	37624341	19-SEP-2003	
52	GZ-B	AY394978.1	29640	29-JAN-2004	China: Guangzhou	37624321	19-SEP-2003	
53	GZ-C	AY394979.1	29645	29-JAN-2004	China: Guangzhou	37624322	19-SEP-2003	
54	HSZ2-A	AY394983.1	29699	29-JAN-2004	China: Guangzhou	37624326	19-SEP-2003	
55	HZS2-Fb	AY394987.1	29709	29-JAN-2004	China: Guangzhou	37624330	19-SEP-2003	
56	HSZ-Bb	AY394985.1	29530	29-JAN-2004	China: Guangzhou	37624328	19-SEP-2003	
57	HSZ-Cb	AY394986.1	29729	29-JAN-2004	China: Guangzhou	37624329	19-SEP-2003	
58	HSZ-Bc	AY394994.1	29765	29-JAN-2004	China: Guangzhou	37624337	19-SEP-2003	
59	HSZ-Cc	AY394995.1	29765	29-JAN-2004	China: Guangzhou	37624338	19-SEP-2003	
60	ZS-B	AY394996.1	29683	29-JAN-2004	China: Guangzhou	37624339	19-SEP-2003	
61	ZS-A	AY394997.1	29683	29-JAN-2004		37624340	19-SEP-2003	
62	ZS-C	AY395003.1	29647	29-JAN-2004	China: Guangzhou	37624346	19-SEP-2003	
63	LC2	AY394999.1	29350	29-JAN-2004	China: Guangzhou	37624342	19-SEP-2003	
64	LC3	AY395000.1	29350	29-JAN-2004	China: Guangzhou	37624343	19-SEP-2003	
65	LC4	AY395001.1	29350	29-JAN-2004		37624344	19-SEP-2003	
66	LC5	AY395002.1	29350	29-JAN-2004	China: Guangzhou	37624345	19-SEP-2003	
67	AS	AY427439.1	29711	21-OCT-2003	Italy: Milan	37576845	02-OCT-2003	
68	SoD	AY461660.1	29715	23-NOV-2003	Russia	38385714	31-OCT-2003	
69	ShanghaiQXC1	AY463059.1	29592	05-JAN-2004	Shanghai	40457433	11-NOV-2003	M=1, R=1, Y=1
70	ShanghaiQXC2	AY463060.1	29013	05-JAN-2004	Shanghai	40457448	11-NOV-2003	R=1, Y=1
71	Sino1-11	AY485277.1	29741	30-NOV-2003	China: Beijing	38505482	21-NOV-2003	
72	Sino3-11	AY485278.1	29740	30-NOV-2003	China: Beijing	38505491	21-NOV-2003	
73	TW1	AY291451.1	29729	25-FEB-2004	Taiwan	30698326	06-MAY-2003	
74	TW2	AY502925.1		25-FEB-2004	Taiwan	40548897	15-DEC-2003	
75	TW3	AY502926.1	29729	25-FEB-2004	Taiwan	40548909	15-DEC-2003	M=1, Y=1
76	TW4	AY502927.1	29729	25-FEB-2004	Taiwan	40548921	15-DEC-2003	R=1, W=1
77	TW5	AY502928.1	29729	25-FEB-2004	Taiwan	40548933	15-DEC-2003	
78	TW6	AY502929.1	29729	25-FEB-2004	Taiwan	40548945	15-DEC-2003	
79	TW7	AY502930.1	29729	25-FEB-2004	Taiwan	40548957	15-DEC-2003	
80	TW8	AY502931.1	29729	25-FEB-2004	Taiwan	40548969	15-DEC-2003	M=1
81	TW9	AY502932.1	29729	25-FEB-2004	Taiwan	40548981	15-DEC-2003	R=1
82	TW10	AY502923.1	29729	25-FEB-2004	Taiwan	40548873	15-DEC-2003	

Label	ID	Accession No.	Length	Revision date	Country/Source	GI	FirstSubmitted	No. ambigBases
83	TW11	AY502924.1	29727	25-FEB-2004	Taiwan	40548885	15-DEC-2003	W=1, Y=2
84	Sin842	AY559081.1	29716	24-SEP-2004	Singapore	45644994	24-FEB-2004	
85	Sin852	AY559082.1	29670	24-SEP-2004	Singapore	45644996	24-FEB-2004	
86	Sin3765V	AY559084.1	29722	24-SEP-2004	Singapore	45645000	24-FEB-2004	
87	Sin848	AY559085.1	29713	24-SEP-2004	Singapore	45645001	24-FEB-2004	
88	Sin849	AY559086.1	29661	24-SEP-2004	Singapore	45645003	24-FEB-2004	
89	Sin846	AY559094.1	29577	24-SEP-2004	Singapore	45645021	24-FEB-2004	
90	Sin3725V	AY559087.1	29716	24-SEP-2004	Singapore	45645004	24-FEB-2004	Y=5
91	SinP1	AY559088.1	29714	24-SEP-2004	Singapore	45645007	24-FEB-2004	R=1
92	SinP3	AY559090.1	29725	24-SEP-2004	Singapore	45645013	24-FEB-2004	R=1
93	SinP5	AY559092.1	29713	24-SEP-2004	Singapore	45645017	24-FEB-2004	R=1
94	SinP4,	AY559091.1	29710	24-SEP-2004	Singapore	45645016	24-FEB-2004	R=1
95	Sin845	AY559093.1	29712	24-SEP-2004	Singapore	45645019	24-FEB-2004	
96	Sin847	AY559095.1	29719	24-SEP-2004	Singapore	45645022	24-FEB-2004	
97	Sin850	AY559096.1	29720	24-SEP-2004	Singapore	45645023	24-FEB-2004	
98	LLJ-2004	AY595412.1	29716	29-JUN-2004	"aerosol sample" China: Beijing	49176846	08-APR-2004	K=3, M=1, R=5, S=1, W=1
99	TJF	AY654624.1	29745	24-JUL-2004	China	50365700	16-JUN-2004	
100	CDC#200301157	AY714217.1	29727	28-SEP-2004	USA	52546959	10-AUG-2004	
101	Sin3408	AY559083.1	29767	24-SEP-2004	Singapore	45644998	24-FEB-2004	N=263, Y=1
102	SinP2	AY559089.1	29717	24-SEP-2004	Singapore	45645010	24-FEB-2004	N=30, R=1, S=1
103	Sin3408L	AY559097.1	29715	24-SEP-2004	Singapore	45645024	24-FEB-2004	N=105, W=3, Y=1

Table S2. SARS-CoV genome polymorphism (extended table). All the 103 isolates are represented. Shaded entries correspond to annotated isolates. Last three isolates contain large number of ambiguous nucleotide codes (N).
IDENTIFICATION of isolates is given in accordance to the Labels, IDs and Accession numbers from the table S1. The four **SNVs** columns correspond to the total number of SNVs, number of SNVs in genes, number of SNVs in 5' and 3' UTRs, and the number of SNVs in intergenic regions (IGR). The eight columns named **INDELs** include number of deletions at the 5' end (at **first 15** genome positions, where most of the isolates are empty, and at the **following 5'UTR** positions, where most of the isolates are non-empty); the length and position of long insertions and long deletions (**longIns**, **longDel**), both absolute for the isolate in question and relative to the CLUSTAL output; number and length of short insertions and short deletions (**shortIns**, **shortDel**) in the form *axb* where *b* denotes the length and *a* the number of occurrences, along with their absolute and relative positions; number of deletions at the 3'end (**3'del)** and the length of a poly-a sequence at the 3' end (**3'poly-a**). **CLASSIFICATION** includes two columns. The **Type** column corresponds to the nine loci nucleotides that are given in the form NNNN/NNNNN and represent nucleotides at (relative to CLUSTAL output) positions 9420, 17604, 222274, 27891 and 3861, 9495, 11514, 21773, 26534, respectively (absolute Tor2 positions 9404, 17564, 22222, 27827 / 3852, 9479, 11493, 21721, 26477). The last column, **Group,** reflects the proposed grouping of isolates.

IDENTIFICATION			SNVs				INDELs								CLASSIFICATION	
Label	**ID**	**Accession No.**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' Del -first 15**	**5' Del -follow.**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3'del**	**3'poly-a**	**Type**	**Group**
1,2	Tor2	Nc_004718.3	2	2	-/-	-	-	-	-	-	-	-	-	24	tttt/ttcgt	A1
3	Urbani	Ay278741.1	5	5	-/-	-	-	-	-	-	-	-	-	-	tttt/ttcgt	A1
4	CUHK-W1	Ay278554.2	9	8	-/-	1	15	-	-	-	-	-	-	24	cgcc/ttcat	B1
5	BJ01	Ay278488.2	12	11	-/-	1	15	4	-	-	-	-	-	17	cgcc/ttcat	B1
6	BJ02	Ay278487.3	22	22	-/-	-	-	-	-	-	-	-	-	18	cgcc/ttcat	B1
7	BJ03	Ay278490.3	22	22	-/-	-	4	-	-	-	-	-	-	17	cgcc/ttcat	B1
8,9	NS-1(BJ04)	AY508724.1	15	14	-/-	1	15	1	-	-	-	-	-	21	tgcc/ttcgt	B1
10	GD01	Ay278489.2	49	49	-/-	-	15	1	29 (27868/ 27995)	-	-	-	-	17	cgcc/tccat	B2
11	HKU-39849	Ay278491.2	9	9	-/-	-	-	-	-	-	-	-	-	15	tttt/ttcgt	A1
12	CUHK-Su10	Ay282752.2	2	1	-/-	1	15	-	-	-	-	-	-	24	tttt/ttcgg	A1
13	Sin2500	Ay283794.1	2	2	-/-	-	15	1	-	-	-	-	-	-	tttt/ttcgt	A1
14	Sin2679	Ay283796.1	2	2	-/-	-	15	1	-	-	-	-	-	-	tttt/ttcgt	A1
15	Sin2774	Ay283798.2	4	4	-/-	-	15	1	-	-	-	-	-	-	tttt/ttcgt	A1
16	Sin2677	Ay283795.1	3	3	-/-	-	15	1	-	-	-	1x6 (27766 / 27893)	-	-	tttt/ttcgt	A1
17	Sin2748	Ay283797.1	1	1	-/-	-	15	1	-	-	-	1x5 (27794 / 27921)	-	-	tttt/ttcgt	A1
18	Frankfurt 1	Ay291315.1	7	7	-/-	-	-	-	-	-	-	-	-	-	tttt/ttcgt	A1
19	FRA	Ay310120.1	7	7	-/-	-	-	-	-	-	-	-	-	13	tttt/ttcgt	A1
20	ZJ01	Ay297028.1	23	23	-/-	-	14	-	-	-	7x1 (8548,12975, 13374,13450, 13462,15558, 27734 / 8612,13051, 13451,13527, 13539,15644, 27853)	2x1 (15526, 28465 / 15611, 28613 /)	3	-	tttt/ttcgt	B4
21	SZ3	Ay304486.1	54	53	-/1	-	15	-	29 (27869 / 27995)	-	-	-	-	-	cgcc/tccat	B2
22	SZ16	Ay304488.1	55	55	-/-	-	15	-	29 (27869 / 27995)	-	-	-	10	-	cgcc/tccat	B2
23	GZ50	Ay304495.1	11	10	-/1	-	15	-	-	-	-	-	-	8	tgcc/ttcat	B1
24	GD69	Ay313906.1	21	21	-/-	-	-	-	-	-	1x1 (14295 / 14370), 1x10 (20374/ 20456)	-	-	16	tttt/cttgg	A1
25	TWC	Ay321118.1	2	2	-/-	-	-	-	-	-	-	1x2 (27806 / 27919)	-	-	tttt/ttcgt	A1
26	HSR 1	Ay323977.2	0	0	-/-	-	-	-	-	-	-	-	-	24	tttt/ttcgt	A1
27	Taiwan TC1	Ay338174.1	4	4	-/-	-	15	54	-	-	-	-	85	-	tttt/cttgg	A1
28	Taiwan TC2	Ay338175.1	9	9	-/-	-	15	54	-	-	-	-	85	-	tttt/cttgg	A1
29	Taiwan TC3	Ay348314.1	7	6	-/-	1	15	54	-	-	-	-	85	-	tttt/cttgg	A1

IDENTIFICATION			SNVs				INDELs								CLASSIFICATION	
Label	**ID**	**Accession No.**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' Del -first 15**	**5' Del -follow.**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3'del**	**3'poly-a**	**Type**	**Group**
30,31	CUHK-AG01(02)	AY345986.1	3	3	-/-	-	15	-	-	-	-	-	-	24	tttt/cttgg	A1
32	CUHK-AG03	AY345988.1	5	4	-/-	1	15	-	-	-	-	-	-	24	tttt/cttgg	A1
33	PUMC01	AY350750.1	3	2	-/-	1	13	-	-	-	-	-	-	24	tttt/ttcgg	A1
34	PUMC02	AY357075.1	2	1	-/-	1	14	-	-	-	-	1x2 (27869 / 27994 (1), 28024 (1))	-	27	tttt/ttcgg	A1
35	PUMC03	AY357076.1	4	3	-/-	1	14	-	-	-	-	1x3 (26142 / 26260)	-	35	tttt/ttcgg	A1
36	ZMY 1	Ay351680.1	78	77	1/-	-	-	-	-	-	24x1 (1031, 1041, 1053, 2521, 2528, 3850, 4220, 4826, 6339, 6426, 7798, 8084, 11729 11782, 14063, 14113, 14134,14177, 14261, 14271, 19118, 20692, 22750, 25301 / 1067, 1077, 1089, 2558, 2565, 3887, 4257, 4863, 6377, 6464, 7836, 8122, 11777,11830, 14120, 14170, 14191,14234, 14318, 14328, 19178,20766, 22825,25381)	2x1 (10601, 10609 / 10642, 10651) 1x2 (28734 / 28852)	-	2	tttt/ttcgt	B4
37,38	TWH (TWC2)	Ap006557.1	4	4	-/-	-	-	-	-	-	-	-	-	-	tttt/cttgg	A1
39	TWK	Ap006559.1	7	7	-/-	-	-	-	-	-	-	-	-	-	tttt/cttgg	A1
40	TWS	Ap006560.1	6	6	-/-	-	-	-	-	-	-	-	-	-	tttt/cttgg	A1
41	TWY	Ap006561.1	6	6	-/-	-	-	-	-	-	-	-	-	-	tttt/cttgg	A1
42	TWC3	Ay362699.1	3	3	-/-	-	-	-	-	-	-	-	-	-	tttt/cttgg	A1
43	TWJ	Ap006558.1	6	6	-/-	-	-	-	-	-	-	1x2 (27167 / 27271)	-	-	tttt/cttgg	A1
44	GZ02	AY390556.1	39	39	-/-	-	-	-	29 (27884 / 27995)	-	-	-	-	4	cgcc/tccat	B2
45	WHU	Ay394850.2	15	15	-/-	-	-	-	-	-	-	1x2 (27808 / 27919)	-	3	tttt/ttcgt	A1
46	HZS2-D	AY394989.1	5	5	-/-	-	15	-	-	-	-	-	-	24	tgcc/ttcat	A1
47	HZS2-E	AY394990.1	5	5	-/-	-	15	-	-	-	-	-	-	24	tgcc/ttcat	A1
48	HZS2-Fc	AY394991.1	6	6	-/-	-	15	-	-	-	-	-	-	24	tgcc/ttcgt	A1
49	HZS2-C	AY394992.1	7	7	-/	-	15	-	-	-	-	-	-	24	tgcc/ttcat	A1
50	HGZ8L2	AY394993.1	7	7	-/-	-	15	-	-	-	-	-	-	24	tgcc/ttcat	A1
51	LC1	AY394998.1	1	1	-/-	-	15	-	-	-	-	-	-	24	tttt/ttcgg	A1
52	GZ-B	AY394978.1	3	3	-/-	-	15	57	-	39 (27699 / 27882)	-	-	-	24	tttt/ttcgt	A3

IDENTIFICATION			SNVs				INDELs								CLASSIFICATION	
Label	**ID**	**Accession No.**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' Del -first 15**	**5' Del -follow.**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3'del**	**3'poly-a**	**Type**	**Group**
53	GZ-C	AY394979.1	14	14	-/-	-	15	37	-	39 (27719 / 27882), 12 (28039 / 28273)	-	1x3 (27791 / 27993 (2), 28024 (1))	-	24	cttt/ttcgt	A3
54	HSZ2-A	AY394983.1	5	5	-/-	-	15	37	-	-	-	-	-	24	tgcc/ttcat	A1
55	HZS2-Fb	AY394987.1	5	5	-/-	-	15	27	-	-	-	-	-	24	tgcc/ttcgt	A1
56	HSZ-Bb	AY394985.1	14	14	-/-	-	15	235	29 (27634 / 27995)	-	-	-	-	24	cgcc/tccat	B2
57	HSZ-Cb	AY394986.1	16	16	-/-	-	15	36	29 (27833 / 27995)	-	-	-	-	24	cgcc/tccat	B2
58	HSZ-Bc	AY394994.1	13	13	-/-	-	15	-	29 (27869 / 27995)	-	-	-	-	24	cgcc/tccat	B2
59	HSZ-Cc	AY394995.1	19	19	-/-	-	15	-	29 (27869 / 27995)	-	-	-	-	24	cgcc/tccat	B2
60,61	ZS-A (ZS-B)	AY394997.1	38	38	-/-	-	15	-	-	53 (27843 / 27969 (26), 28024 (27))	-	-	-	24	cgcc/tccat	B3
62	ZS-C	AY395003.1	38	38	-/-	-	15	36	-	53 (27807 / 27969 (26), 28024 (27))	-	-	-	24	cgcc/tccat	B3
63	LC2	AY394999.1	4	4	-/-	-	15	-	-	386 (27704 / 27829)	-	-	-	24	ttt-/ttcgt	A3
64,65	LC4 (LC3)	AY395001.1	3	3	-/-	-	15	-	-	386 (27704 / 27829)	-	-	-	24	ttt-/ttcgt	A3
66	LC5	AY395002.1	4	4	-/-	-	15	-	-	386 (27704 / 27829)	-	-	-	24	ttt-/ttcgt	A3
67	AS	Ay427439.1	0	0	-/-	-	15	1	-	-	-	-	-	-	tttt/ttcgt	A1
68	SoD	AY461660.1	30	10	1/19	-	15	-	-	-	-	-	-	-	tttt/ttcgt	A1
69	ShanghaiQXC1	AY463059.1	39	39	-/-	-	15	64	-	-	-	-	56	-	cgtt/ttcgt	B1
70	ShanghaiQXC2	AY463060.1	39	39	-/-	-	15	64	-	579 (5834 / 5959 (418), 6378 (86), 6465 (75))	-	-	56	-	cgtt/ttcgt	B1
71	Sino1-11	AY485277.1	6	6	-/-	-	-	-	-	-	-	1x3 (26156 / 26260)	-	17	tttt/ttcgg	A1
72	Sino3-11	AY485278.1	3	3	-/-	-	-	-	-	-	-	1x2 (27883 / 27994 (1), 28024 (1))	-	15	tttt/ttcgg	A1
73,74	TW2 (TW1)	Ay502925.1	1	1	-/-	-	-	-	-	-	-	-	-	2	tttt/ttcgt	A1
75	TW3	AY502926.1	2	2	-/-	-	-	-	-	-	-	-	-	2	tttt/ttcgt	A1
76	TW4	AY502927.1	2	2	-/-	-	-	-	-	-	-	-	-	2	tttt/ttcgt	A1
77	TW5	AY502928.1	1	1	-/-	-	-	-	-	-	-	-	-	2	tttt/ttcgg	A1
78	TW6	AY502929.1	3	3	-/-	-	-	-	-	-	-	-	-	2	tttt/cttgg	A1
79	TW7	AY502930.1	4	4	-/-	-	-	-	-	-	-	-	-	2	tttt/cttgg	A1
80	TW8	AY502931.1	3	3	-/-	-	-	-	-	-	-	-	-	2	tttt/cttgg	A1
81	TW9	AY502932.1	5	4	-/-	1	-	-	-	-	-	-	-	2	tttt/cttgg	A1
82	TW10	AY502923.1	6	5	-/-	1	-	-	-	-	-	-	-	2	tttt/cttgg	A1
83	TW11	AY502924.1	9	8	-/-	1	-	-	-	-	-	1x2 (27068 / 21172)	-	2	tttt/cttgg	A1
84	Sin842	AY559081.1	4	4	-/-	-	13	-	-	-	1x1 (13953 / 14034)	-	-	1	tttt/ttcgt	A1
85	Sin852	AY559082.1	19	9	10/-	-	1	-	-	57 (27797 /	-	-	-	1	ttt-/ttcgt	A3

IDENTIFICATION			SNVs				INDELs								CLASSIFICATION	
Label	**ID**	**Accession No.**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' Del -first 15**	**5' Del -follow.**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3'del**	**3'poly-a**	**Type**	**Group**
										27909)						
86	Sin3765V	AY559084.1	9	9	-/-	-	15	1	-	-	-	-	-	11	tttt/ttcgt	A1
87	Sin848	AY559085.1	11	11	-/-	-	15	1	-	-	-	-	-	2	tttt/ttcgt	A1
88	Sin849	AY559086.1	4	4	-/-	-	15	1	-	49 (27745 / 27872)	-	-	1	-	tttt/ttcgt	A3
89	Sin846	AY559094.1	7	7	-/-	-	15	1	-	137 (27663 / 27787 (66), 27854 (71)	2x1 (9567, 9667 / 9634, 9734)	-	-	1	tttt/ttcgt	A3
90	Sin3725V	AY559087.1	4	4	-/-	-	15	1	-	-	-	-	-	5	tttt/ttcgt	A1
91	SinP1	AY559088.1	4	4	-/-	-	15	1	-	-	2x1 (2488, 25277 / 2543, 25394)	-	-	1	tttt/ttcgt	A1
92	SinP3	AY559090.1	9	4	1/4	-	15	1	-	-	2x2, (10, 25286 / 60, 25394) 9x1 (10651,10697,11112,11567, 11782,11933,11988,22506, 29716 / 10718,10764,11179,11634, 11852,12003,12058,22607, 29860)	-	-	1	tttt/ttcgt	A4
93	SinP5	AY559092.1	4	4	-/-	-	15	1	-	-	1x2 (25276 / 25394)	1x1(28088 / 28242)	-	1	tttt/ttcgt	A1
94	SinP4	AY559091.1	7	4	-/3	-	15	1	-	-	1x2 (25274 / 25394)	2x1 (24085, 24106 / 24201, 24223)	1	-	tttt/ttcgt	A1
95	Sin845	AY559093.1	10	10	-/-	-	15	1	-	-	-	-	-	1	tttt/ttcgt	A1
96	Sin847	AY559095.1	12	10	2/-	-	10	-	-	-	-	-	-	2	tttt/ttcgt	A1
97	Sin850	AY559096.1	11	6	5/-	-	8	-	-	-	-	-	-	1	tttt/ttcgt	A1
98	LLJ-2004	AY595412.1	11	10	-/-	1	15	6	-	-	1x6 (27637 / 27763)	1x1 (2919 / 2982)	-	5	cgcc/ttcat	B1
99	TJF	AY654624.1	17	10	2/1	4	1	-	-	-	-	-	-	19	tgcc/ttcgt	B1
100	CDC#200301157	AY714217.1	2	2	-	-	-	-	-	-	-	-	-	-	tttt/ttcgt	A1
101	Sin3408L	AY559097.1	4	4	-	-	15	1	-	-	1x1 (5120 / 5181)	2x1 (24524, 25029 / 24639, 25145)	-	5	tttt/ttcgt	A1
102	SinP2	AY559089.1	4	4	-	-	15	1	-	-	13x1 (10649,11638,13060,17980, 19282,19497,19928,20029, 23499,23543,23589,23607, 25127 / 10718,11710,13138,18071, 19374,19589,200	5x1 (16584, 25141, 25144, 25169, 25171 / 16674, 25246, 25250, 25276, 25279), 1x6 (25189 / 25298)	-	2	tttt/ttcgt	A4

IDENTIFICATION			SNVs				INDELs								CLASSIFICATION	
Label	**ID**	**Accession No.**	**Total**	**Genes**	**5'/3' UTR**	**IGR**	**5' Del -first 15**	**5' Del -follow.**	**longIns**	**longDel**	**shortIns**	**shortDel**	**3'del**	**3'poly-a**	**Type**	**Group**
											20,20121, 23604,23648,23694,23712, 25232), 1x2 (25278 / 25394)					
103	Sin3408	AY559083.1	14	4	10/-	-	-	-	5'end (34)	-	-	-	-	6	tttn/ttcgt	A1

Table S3. RSCU for all the annotated proteins and in total, for all the annotated isolates.
The highest RSCU synonumous codons are in bold

a.a.	codon	TOTAL	Spike	Envel.	Membr.	Nucleo.	1AB	Other
Ala	gca	1.11	0.86	0.00	0.63	1.06	1.16	1.32
	gcc	0.59	0.57	1.00	0.84	0.94	0.55	0.54
	gcg	0.24	0.24	**2.00**	0.42	0.35	0.2	0.40
	gct	**2.07**	**2.33**	1.00	**2.10**	**1.65**	**2.09**	**1.74**
Arg	aga	**2.09**	**1.85**	0.00	0.80	**1.94**	**2.22**	**2.20**
	agg	0.97	**1.85**	0.00	0.80	0.39	0.85	1.68
	cga	0.47	0.61	**3.00**	0.80	1.16	0.26	1.13
	cgc	0.76	0.46	0.00	1.20	1.55	0.76	0.15
	cgg	0.12	0.15	0.00	0.80	0.00	0.07	0.24
	cgt	1.59	1.07	**3.00**	**1.60**	0.97	1.85	0.60
Asn	aac	0.75	0.69	0.80	**1.23**	0.72	0.74	0.93
	aat	**1.25**	**1.31**	**1.20**	0.77	**1.28**	**1.26**	**1.07**
Asp	gac	0.75	0.63	0.00	**1.67**	**1.09**	0.72	**1.21**
	gat	**1.25**	**1.37**	**2.00**	0.33	0.91	**1.28**	0.79
Cys	tgc	0.75	0.93	**1.33**	0.32	0.00	0.69	**1.06**
	tgt	**1.25**	**1.07**	0.67	**1.68**	**2.00**	**1.31**	0.94
Gln	caa	**1.21**	**1.60**	0.00	**1.20**	**1.41**	**1.31**	**1.13**
	cag	0.79	0.40	0.00	0.80	0.59	1	0.87
Glu	gaa	**1.05**	1.00	**2.00**	**1.15**	1.00	0.69	0.79
	gag	0.95	1.00	0.00	0.85	1.00	**1.1**	**1.21**
Gly	gga	0.91	1.01	2.00	0.80	1.42	0.9	0.71
	ggc	1.01	1.26	0.00	0.80	1.42	1.07	1.38
	ggg	0.18	0.20	0.00	0.80	0.27	0.14	0.18
	ggt	**1.9**	**1.52**	2.00	**1.60**	0.89	2.11	**1.73**
His	cac	0.71	0.27	0.00	**2.00**	0.80	0.68	**1.05**
	cat	**1.29**	**1.73**	0.00	0.00	**1.20**	1.32	0.95
Ile	ata	0.66	0.54	1.00	0.50	0.28	0.65	1.05
	atc	0.63	0.35	1.00	0.33	0.81	0.72	0.51
	att	**1.7**	**2.11**	1.00	**2.17**	**1.91**	1.62	**1.44**
Leu	cta	0.69	0.61	0.86	1.16	1.15	0.6	1.15
	ctc	0.82	1.15	0.00	1.15	0.69	0.8	0.69
	ctg	0.59	0.24	0.86	0.58	0.92	0.62	0.60
	ctt	**1.76**	**2.06**	**2.57**	**1.55**	**1.62**	1.76	**1.44**
	tta	1.06	1.15	0.86	0.58	0.23	1.11	0.98
	ttg	1.08	0.79	0.86	0.97	1.38	1.11	1.13
Lys	aaa	**1.06**	**1.13**	**2.00**	**1.33**	**1.38**	1.03	**1.03**
	aag	0.94	0.87	0.00	0.67	0.62	0.97	0. 97
Met	atg	1	1.00	1.00	1.00	1.00	1	1.00
Phe	ttc	0.76	0.58	**1.49**	**1.04**	**1.39**	0.78	0.68
	ttt	**1.24**	**1.42**	0.51	0.96	0.61	1.22	**1.32**
Pro	cca	1.7	1.33	2.00	**2.32**	1.29	1.79	**1.91**
	ccc	0.41	0.22	0.00	0.06	1.16	0.37	0.51
	ccg	0.17	0.14	0.00	0.83	0.26	0.12	0.49
	cct	**1.72**	**2.31**	2.00	0.79	1.29	1.72	1.08
Ser	agc	0.53	0.38	0.86	1.00	0.86	0.47	1.06
	agt	1.13	0.69	0.86	0.50	1.20	1.28	0.69
	tca	1.76	1.75	0.86	**2.00**	1.54	1.75	**2.24**
	tcc	0.41	0.44	0.00	1.00	0.51	0.4	0.25
	tcg	0.24	0.19	**1.71**	0.50	0.17	0.18	0.76
	tct	**1.92**	**2.56**	**1.71**	1.00	**1.71**	1.92	1.01
Thr	aca	1.56	1.49	1.60	**1.54**	1.34	1.63	1.30
	acc	0.55	0.48	0.00	0.92	0.61	0.54	0.65
	acg	0.19	0.16	1.60	0.31	0.00	0.16	0.44
	act	**1.69**	**1.86**	0.80	1.24	**2.06**	1.67	**1.60**
Trp	tgg	1	1.00	0.00	1.00	1.00	1	1.00
Tyr	tac	0.88	0.67	**1.99**	**1.33**	**1.64**	0.86	0.88
	tat	**1.12**	**1.33**	0.01	0.67	0.36	1.14	**1.12**
Val	gta	0.85	0.53	0.86	1.25	0.36	0.9	0.83
	gtc	0.69	0.83	0.79	0.50	1.46	0.67	0.40
	gtg	0.79	0.57	0.57	**1.50**	0.73	0.79	1.01
	gtt	**1.68**	**2.07**	**1.78**	0.76	1.46	1.64	**1.75**
End	taa	**1.98**	**3.00**	**3.00**	**3.00**	**3.00**	0.93	**1.85**
	tag	0.42	0.00	0.00	0.00	0.00	0.85	0.22
	tga	0.6	0.00	0.00	0.00	0.00	0.9	0.94

Table S4. Positions of multiple SNVs in annotated isolates.
Positions of two or more SNVs are presented along with nucleotides and ORFs (based on HSR 1 annotation),
type of mutation (transition / transversion), a.a. position in ORF, a.a. change, a.a. properties change,
nucleotide position in codon and the number of SNVs. SNVs are in red.

| | 1ab | 1ab | 1ab

Table S5. Categories of nucleotide substitutions in the whole genome (annotated isolates)
Only sites where two or more isolates have an SNV are considered

			1.pos	2.pos	3.pos	TotalNo		1.pos %	2.pos %	3.pos %	Total %	silent
Transitions	A-G	A→G	5/10	6/12	6/20	17/42	30/71	3.47%	4.17%	6.94%	14.58%	6/20
		G→A	4/14	6/9	3/6	13/29		4.86%	3.12%	2.08%	10.07%	3/6
	C-T	C→T	6/19	8/24	7/35	21/78	45/156	6.60%	8.33%	12.15%	27.08%	9/46
		T→C	2/9	14/39	8/30	24/78		3.12%	13.54%	10.42%	27.08%	8/30
	Total		17/52	34/84	24/91		75/227	18.06%	29.17%	31.60%	78.82%	26/102
Transversions	A-C	A→C	0	1 / 2	1 / 2	2/4	2/4	0	0.69%	0.69%	1.39%	1/2
		C→A	0	0	0	0		0	0	0	0	0
	A-T	A→T	3/6	0	0	3/6	7/14	2.08%	0	0	2.08%	0
		T→A	0	0	4/8	4/8		0	0	2.78%	2.78%	4/8
	G-C	G→C	0	0	0	0	0	0	0	0	0	0
		C→G	0	0	0	0		0	0	0	0	0
	G-T	G→T	2/4	0	1/2	3/6	7/43	1.39%	0	0.69%	2.08%	0
		T→G	3/14	1/23	0	4/37		4.86%	7.99%	0	12.85%	0
	Total		8/24	2/25	6/12		16/61	8.33%	8.68%	4.17%	21.18%	5/10
Total			25/76	36/109	30/103		91/288	26.39%	37.85%	35.77%	100%	31/112

Table S6. Mutation analysis of S-protein.
Nonsynonumous mutation sites are in gray.

Relative nucleotide position	81	90	146	147	183	224	230	411	430	654	681	716	729	731	782	931	1026	1031	1058	1079	1098	1170	1306	1325	1437	1460	1502	1672	1683	1729	1788	1813	1819	1822	1954	1994	2003	2102	2227	2244	2261	2301	2332	2380	2578	2581	2680	2892	3002	3075	3204	3321	3381	3395	3442	3487	3505	3546	3547	3623	3667	3739
PROFILE	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	t	t	c
AY394999	g	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	t	t	c
AY395000	g	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	t	t	c
AY395002	g	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	t	t	c
AY323977	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY338175	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY345988	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY350750	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY357076	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394998	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502923	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502926	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502929	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502932	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY463059	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	g	t	a	t	g	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	g	c	a	a	t	c	c	t	c
AY463060	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	g	t	a	t	g	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	g	c	a	a	t	c	c	t	c
NC_004718	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	g	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394979	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	c	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	a	a	c	t	c
AP006560	t	t	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278741	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	c	a	c	a	a	t	c	c	t	c
AY297028	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	t	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY313906	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	g	t	t	a	c	a	a	t	c	c	t	c
AY394978	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	g	t	c	c	t	c
AY485277	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	c	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY714217	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	g	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278487	t	c	c	c	a	c	a	c	a	t	c	c	c	c	c	a	g	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394983	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278489	t	c	c	a	a	c	a	c	a	c	c	t	c	c	c	a	g	g	c	t	c	g	t	a	t	c	a	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	g	t	g	a	a	g	g	c	a	t	t	a	c	a	a	t	c	c	t	c
AY390556	t	c	c	a	a	c	a	c	a	c	c	t	c	c	c	a	g	g	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	c	a	t	t	a	c	g	a	t	c	c	t	c
AY394996	t	c	c	a	a	g	a	c	a	t	c	t	c	c	c	a	g	g	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	c	a	t	t	a	c	a	a	t	c	c	t	c
AY395003	t	c	c	a	a	g	a	c	a	t	c	t	c	c	c	a	g	g	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	c	a	t	t	a	c	a	a	t	c	c	t	c
AY304486	t	c	c	a	a	c	a	c	a	t	a	t	c	c	a	g	g	g	c	c	c	a	t	a	a	g	t	t	c	t	t	g	c	a	a	c	g	t	g	a	t	c	g	c	g	a	g	g	g	c	a	t	t	a	c	g	a	t	c	c	t	c
AY304488	t	c	c	a	a	c	a	c	a	t	a	t	t	c	a	g	g	g	c	c	c	g	t	a	a	g	t	a	c	t	t	g	c	a	a	c	g	t	g	a	t	c	g	c	g	a	g	t	g	c	a	t	t	a	c	g	a	t	c	c	t	c
AY394985	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	t	a	t	t	a	c	g	a	t	c	c	t	c
AY394994	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	c	a	t	t	a	c	g	a	t	c	c	t	c
AY394986	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	t	a	t	t	a	c	g	a	t	c	c	t	t
AY394995	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	t	a	t	t	a	c	g	a	t	c	c	t	t
AY304495	t	c	t	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	g	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278488	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278554	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394989	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394990	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394993	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY595412	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278490	t	c	c	a	a	c	a	c	c	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	c	c	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394992	t	c	c	a	a	c	a	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	t	a	a	t	c	c	t	c
AY508724	t	c	c	a	c	c	g	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	t	t	a	t	t	a	c	a	a	t	c	c	t	c
AY654624	t	c	c	a	a	c	g	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	a	t	a	c	a	a	t	c	c	t	c
AY394987	t	c	c	a	a	c	g	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394991	t	c	c	a	a	c	g	c	a	t	c	c	c	c	c	g	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY291315	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	t	a	a	t	c	c	t	c
AY310120	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	t	a	a	t	c	c	t	c
AY461660	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	t	a	a	t	c	c	t	c
AY559081	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559083	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559085	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559086	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559087	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559093	t	c	c	a	a	c	g	t	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	a	c
AY559094	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559095	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559096	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559082	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	c	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559084	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	t	t	c	g	t	a	t	c	t	t	t	t	c	g	t	a	a	t	g	c	a	g	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY283796	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	t	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY394850	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY485278	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502925	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502928	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502931	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502930	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502927	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY502924	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY427439	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY362699	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY357075	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY348314	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY345986	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY338174	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY321118	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY283797	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY283795	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY283794	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY282752	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY278491	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AP006561	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AP006559	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AP006558	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AP006557	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	c	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559092	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	t	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY559088	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	t	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
AY283798	t	c	c	a	a	c	g	c	a	t	c	c	c	t	c	g	a	a	c	t	c	g	t	a	t	c	t	t	c	t	t	g	t	a	a	t	g	c	a	a	c	t	t	c	g	a	a	g	g	t	a	t	t	a	c	a	a	t	c	c	t	c
nucleotide change	t→g	c→t	c→t	a→c	a→c	c→g	g→a	c→t	a→c	t→c	c→a	c→t	c→t	t→c	c→a	g→a	a→g	a→g	c→t	t→c	c→t	g→a	t→c	a→c	t→a	c→g	t→a	t→a	c→t	t→g	t→c	g→a	t→c	a→g	a→g	t→c	g→a	c→t	a→g	a→g	c→t	c→t	t→g	c→t	g→c	a→c	a→g	g→t	g→t	t→c	a→g	t→a	t→c	a→g	c→t	a→g	a→g	t→a	c→a	a→g	t→a	c→t
transition(Ts) / transversion(Tv)	Tv	Ts	Ts	Tv	Tv	Tv	Ts	Ts	Tv	Ts	Tv	Ts	Ts	Ts	Tv	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Tv	Tv	Tv	Tv	Tv	Tv	Tv	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Ts	Tv	Ts	Tv	Tv	Ts	Tv	Ts	Ts	Ts	Tv	Ts	Ts	Ts	Ts	Ts	Tv	Tv	Ts	Tv	Ts
a.a. changes	Ala	Tyr	Ser→Leu	Ser	Pro	Thr→Arg	Gly→Asp	Phe	Met→Leu	Pro	Asn→Lys	Ser→Leu	Asp	Ile→Thr	Thr→Lys	Gly→Arg	Arg	Lys→Arg	Ser→Phe	Phe→Ser	Cys	Lys	Tyr→His	Tyr→Ser	Asn→Lys	Thr→Ser	Phe→Tyr	Phe→Ile	Ser	Ser→Ala	Val	Asp→Asn	Ser→Pro	Thr→Ala	Ile→Val	Leu→Ser	Ser→Asn	Ser→Leu	Thr→Ala	Ala	Ala→Val	Val	Tyr→Asp	Pro→Ser	Val→Leu	Ser→Arg	Thr→Ala	Ser	Arg→Met	Cys	Lys	Asn→Lys	Leu	Glu→Gly	Leu→Phe	Lys→Glu	Asn→Asp	Leu	Gln→Lys	Ala→Val	Cys→Ser	Leu→Phe
a.a. position	27	30	49	49	61	75	77	137	144	218	227	239	243	244	261	311	342	344	353	360	366	390	436	442	479	487	501	558	561	577	596	605	607	608	652	665	668	701	743	748	754	767	778	794	860	861	894	964	1001	1025	1068	1107	1127	1132	1148	1163	1169	1182	1183	1208	1223	1247
a.a. properties	silent	silent	P+S+T→Hp+Ap	silent	silent	Hp+P+S→P+PCh	Hp+S+T→P+NCh+S	silent	Hp→Hp+Ap	silent	P+S→Hp+P+PCh	P+S+T→Hp+Ap	silent	Hp+Ap→Hp+P+S	Hp+P+S→Hp+P+PCh	Hp+S+T→P+PCh	silent	Hp+P+PCh→P+PCh	P+S+T→Hp+Ar	Hp+Ar→P+S+T	silent	silent	Hp+P+Ar→Hp+P+PCh	Hp+P+Ar→P+S+T	P+S→Hp+P+PCh	Hp+P+S→P+S+T	Hp+Ar→Hp+P+Ar	Hp+Ar→Hp+Ap	silent	P+S+T→Hp+S+T	silent	P+NCh+S→P+S	P+S+T→S	Hp+P+S→Hp+S+T	Hp+Ap→Hp+S+Ap	Hp+Ap→P+S+T	P+S+T→P+S	P+S+T→Hp+Ap	Hp+P+S→Hp+S+T	silent	Hp+S+T→Hp+S+Ap	silent	Hp+P+Ar→P+NCh+S	S→P+S+T	Hp+S+Ap→Hp+Ap	P+S+T→P+PCh	Hp+P+S→Hp+S+T	silent	P+PCh→Hp	silent	silent	P+S→Hp+P+PCh	silent	P+NCh→Hp+S+T	Hp+Ap→Hp+Ar	Hp+P+PCh→P+NCh	P+S→P+NCh+S	silent	→Hp+P+PCh	Hp+S+T→Hp+S+Ap	Hp+P+S+T→P+S+T	Hp+Ap→Hp+Ar
position in codon	3	3	2	3	3	2	2	3	1	3	3	2	3	2	2	1	3	2	2	2	3	3	1	2	3	2	2	1	3	1	3	1	1	1	1	2	2	2	1	3	2	3	1	1	1	1	1	3	2	3	3	3	3	2	1	1	1	3	1	2	1	1
number of SNVs	3	1	1	1	1	2	21	1	1	2	2	6	1	25	2	5	8	6	11	2	1	1	1	1	2	2	1	1	12	3	2	2	2	2	1	2	1	2	2	11	2	3	11	1	1	1	2	1	1	6	1	1	1	2	4	7	1	1	1	3	1	2