academicJournals

Vol. 8(18), pp. 1863-1898, 30 April, 2014
DOI: 10.5897/AJMR12.128
Article Number: 4A0C82244407
ISSN 1996-0808

http://www.academicjournals.org/AJMR

African Journal of Microbiology Research

Full Length Research Paper

Genetic variation analysis of hemagglutinin and neuraminidase of human influenza A/H3N2 virus in Hong Kong (1997-2006)

Shubo Zhang[1], Jake Yue Chen[2,3,4], Jianan Rao[1], Shanghong Zhang[1], Daniel Wai Tin Chan[5], Guanghua Liu[6], Yang Xu[7] and Miao He[1]*

[1]School of Life Sciences, Sun Yat-sen University, China.
[2]School of Informatics, Indiana University, USA.
[3]Department of Computer and Information Science, School of Science, Purdue University, USA.
[4]Indiana Center for Systems Biology and Personalized Medicine, USA.
[5]Research Centre of Built Environment and Energy Conservation, Institute for Advanced Study, Nanchang University, China.
[6]Guangdong AIB Polytechnic College, China.
[7]Department of Biotechnology, Southern Medical University, China.

Received 17 January, 2012; Accepted 7 April, 2014

We analyzed the phylogeny, amino acid variations, positive selection, and glycosylation patterns of hemagglutinin (HA) and neuraminidase (NA) of A/H3N2 in Hong Kong from 1997 to 2006. The results suggest that continuity and latency of influenza viruses might be the reasons why different influenza viruses co-circulate within the same season. Many amino acid mutations were retained for two or more successive years. The preferred antigenic sites of mutation are sites A and B in HAs, and site B in NAs. An influenza pandemic may be caused by higher-than-threshold level of amino acid variations of the virus.

Key words: Influenza virus, A/H3N2, phylogenetic analysis, selection pressure, N-glycosylation.

INTRODUCTION

As members of the family Orthomyxoviridae, influenza A viruses are negative-strand RNA viruses that can be categorized into several subtypes according to the antigenic properties of their surface glycoprotein such as hemagglutinin (HA) and neuraminidase (NA) (Webster et al., 1992). Epidemics of influenza are estimated to affect 3-5 million people per year across the world. In the last century, four influenza pandemics alone claimed more than 50million lives (Stöhr, 2002). A/H3N2 and A/H1N1 are the major influenza A subtypes in human populations.

*Corresponding author. E-mail: lsshem@mail.sysu.edu.cn. Tel: +86 20 84110036.

Influenza A virus causes epidemics and pandemics through antigenic drift or antigenic shift (Yewdell, 2011). Antigenic drift results from an accumulation of point mutations leading to minor and gradual antigenic changes, while antigenic shift involves major antigenic changes by the introduction of new HA and/or NA subtype into human populations (Webster et al., 1992).

How influenza A viruses evolve is a major research topic (Kryazhimskiy et al., 2008). In the past decade, there has been rapid accumulation of sequence data for influenza viruses. Phylogenetic and evolutionary analytical techniques have been developed with advanced in high-throughput molecular biology and computational biology. Detecting positive selection sites of amino acid substitution may help track the evolution of influenza A virus. Positive selection is defined as a nucleotide replacement event. If the mutation of a nucleotide is supported by positive selection, its corresponding virus strain may grow to be the principal influenza virus strain and hence cause influenza prevalence (Bush et al., 1999). Phylogenetic analysis is an important way to understand molecular evolution of influenza A virus. By studying 413 complete genomes of human H3N2 influenza A viruses collected during 1997-2005 in the United States, Nelson et al. (2006) identified stochastic processes as the key determinant of influenza A virus evolution. Bragstad et al. (2008) further discovered that the evolution of influenza A virus was stochastically influenced by small "jumps" in genetic distance rather than constant drift, after studying 234 complete genomes of influenza A viruses during 1999-2006 among the Danish.

Due to the absence of complete genomic sequences of influenza A virus in most countries, many researches usually use partial gene fragments to perform evolutionary and mutation analysis (Mehle et al., 2012). The influenza A virus invades human immunologic system either by point mutation accumulation (drift) of principal surface glycoprotein, hemagglutinin (HA) and neuralminidase (NA), or by gene fragment rearrangement of different influenza viruses in a infected cell. Generally, HA1 area of HA proteins contains concentrated epitope, and should experience the strongest positive selection pressure (Nelson et al., 2007). However, it was detected that no influenza gene had been strongly impacted by positive selection pressure. Kryazhimskiy et al. (2008) discovered that human influenza A virus with positive selection evolution was strongly impacted by the long-term sites and specific preferences of individual amino acids.

Human H3N2 subtype of influenza A viruses that led to the third largest human influenza pandemic in the 20th century originated in 1968 in Hong Kong, which, along with surrounding regions in Southern China, has been referred to as the “epicenter” of pandemic influenza A virus outbreaks (Kryazhimskiy et al., 2008). Hong Kong is located in the subtropical region of the Northern Hemisphere, where warm temperature (averaging 24°C) and high humidity (averaging 79%) are major contributing factors to the spread of influenza virus (Feng et al., 2012; Shaman et al., 2011). A/H3N2 virus emerges from February to March every year in Eastern and Southeast Asia, where the virus mutated continually through cross diffusion, eventually reaching North America and Europe in six to nine months (Russell et al., 2008). The epidemic could also take place from June to August (Khor et al., 2012; Viboud et al., 2006).

Current studies on influenza A virus, particularly at the sequence level, focus on the evolutionary mechanisms and diffusion paths. It is very significant to analyze the mutation patterns of influenza genomic sequences and their relationships with the evolution, diffusion and pathogenesis of A/H3N2. Comparisons between antigenic differences and phylogeny are essential to the understanding how multiple lineages of influenza A virus variants emerge. However, researcher studies on influenza viruses in Southeast Asia are quite limited. There has been no report of the relationship between molecular evolution and the epidemic of viruses in the region.

In this study, we analyzed and compared the genetic diversity and mechanisms underlying the evolutionary dynamics of HA and NA of influenza A/H3N2 viruses isolated in Hong Kong from 1997 to 2006. We find that continuity and latency of influenza viruses might be the reasons why different influenza viruses co-circulate within the same season. Many amino acid mutations are retained for two or more successive years. The preferred antigenic sites of mutation are sites A and B in HAs, and site B in NAs. We discovered several positively selected sites in HAs and NAs. We hypothesize that an influenza pandemic may be caused by higher-than-threshold level of amino acid variations of the virus.

MATERIALS AND METHODS

Epidemiological data and A/H3N2 sequences

The epidemiological data of Hong Kong during 1997-2006 were downloaded from the Department of Health of Hong Kong web site, at http://www.dh.gov.hk/eindex.html. We inferred that the influenza prevalence in Hong Kong during the study period each month, using the ratio of influenza viruses to influenza-like samples recorded.

The HA and NA sequences of A/H3N2 were obtained from the NCBI’s influenza virus resource: http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html. In order to ensure sequential comparability, we downloaded 335 full-length HA’s sequences and 334 full-length NA’s sequences of A/H3N2 of Hong Kong from 1997 to 2006. The sequences of the reference vaccine strains used in this study were collected from the NCBI by searching individual strain’s names. Sequences in other areas of the world were randomly selected from the NCBI’s influenza virus resource, with three sequences per each region per year. We selected

USA, Taiwan, Australia, Denmark, Mainland China, and Japan as the main region, because records about strains in these regions have been kept for a long time.

Sequence alignment and phylogenetic analysis

The HA and NA sequences of A/H3N2 were aligned using BIOEDIT v.7.0.9.0 (Tippmann, 2004) and edited using Mega4 (Tamura et al., 2007). Phylogenetic analysis was performed by using region of 1650 bp for HAs and 1407 bp for NAs. Phylogenetic trees for both HAs and NAs were inferred with the maximum likelihood (ML) method available in the PHYML V3.0 package and SPR branch-swapping (Guindon et al., 2003). In all cases, the simple HKY85 model of nucleotide substitution was used, because the sequences in question are so similar that multiple substitutions can be effectively ignored. The images of the trees were generated using FigTree v1.3.1 (available at http://tree.bio.ed.ac.uk/software/figtree/).

We calculated Pepitope values, the specific measure of antigenic distance between two strains of influenza, by methods suggested by Muñoz et al. (2005). The Pepitope value could also be calculated as the total number of mutations within antibody antigenic sites divided by the length of the antigenic sites. It is assumed that an antigenic epitope which has the greatest ratio of mutations is dominant, because the epitope is influenced by the greatest selective pressure from the immune system. Pepitope distance is defined as the fractional change in the dominant antigenic epitopes of one strain over another strain. Residues in antigenic epitopes were from references (Muñoz et al., 2005）. We grouped the sequences according to their sublineage in the trees, and computed amino acids distances between groups using the MEGA software version 4 (Tamura et al., 2007).

Selection pressure analysis

Selection analysis was carried out by developing probabilistic models of codon substitution with the CodeML program, which is included in the PAML package version 4.0 (Yang, 1997）. This program uses likelihood models that consider heterogeneous substitution ratios (ω = non-synonymous/ synonymous substitution or dN/dS) among sites. We implemented models M0, M1a, M2a, M3, M7, and M8, which are previously diagnosis for their robustness in testing for positive selection (Yang, et al., 2000; Yang, et al., 2005). The parameter ω>1 is considered as an indication of positive selection, whereas ω<1 implies absence of positively selected sites. Three comparisons were conducted. The likelihood ratio test (LRT) compares the likelihood difference ($2\Delta\imath$) twice with $\chi 2$ distribution, if the degrees of freedom (*df*) are equal to the difference in the number of free parameters between the two models (detailed explanation of the models and the parameters can be seen in refs (Wilson et al., 1981; Winter et al., 1981; Wolf et al., 2006). LRT was performed using PAML (Yang, 1997; Yang et al., 2000).

Prediction of glycosylation sites

Potential N-glycosylation sites (amino acids Asn-X-Ser/Thr, where X is not Pro) were predicted by using nine artificial neural networks at the NetNGlyc 1.0 Server (available at www.cbs.dtu.dk) (Gupta, 2004). A threshold value >0.5 average potential score was set to predict glycosylated sites.

RESULTS

In the last 13 years except for 2000, 2001, and 2006, A/H3N2 virus was the dominating subtype in Hong Kong. In 2000 and 2001, higher prevalence of A/H1N1 viruses co-circulating with A/H3N2 viruses was observed. The monthly occurrences of both A/H3N2 and A/H1N1 had similar trends, while their annual dynamics showed inverse trends.

Genetic evolution of influenza A/H3N2 virus

Figures 1 and 2 show the relationships between 335 HA and 334 NA sequences of influenza A/H3N2 samples in Hong Kong and World Health Organization (WHO) (marked in yellow) influenza A/H3N2 vaccine strains. Generally, HA and NA genes formed seasonal phylogenetic clusters. The phylogenetic trees show highly-branched evolution following a major linear trunk route. We observed strains of different lineages and clusters co-circulating within the same season. Interestingly, HA and NA sequences from Hong Kong clustered closely with the WHO vaccine strains A/Sydney/5/97 and A/Moscow/ 10/99 before 2001; however, they were scattered after 2001. On one hand, fewer HA and NA sequences from Hong Kong clustered with the WHO vaccine sequences, A/Fujian/ 411/2002 and A/California/7/2004. On the other hand, A/Wyoming/3/2003 and A/Kumamoto/102 /2002 clustered with HA and NA sequences from Hong Kong very well in the 2003/2004 season; so, did A/Wisonsin/67/2005 for the 2005/2006 season.

The HA sequences of the 2002/2003 season evolved from the 2000/2001 season with the appearance of A/Fujian/411/2002-like strains (Figure 1). In the 2003/2004 season, HAs formed a subclade (A/Wyoming/3/2003 and A/Kumamoto/102/2002-like lineage) evolving from the A/Fujian/411/02 (H3N2)-like lineage from the 2002/2003 season. In the 2005/ 2006 season, A/California/7/2004 (H3N2)-like lineages in the 2004/2005 season continued to circulate together with a slightly different A/Wisconsin/67/2005(H3N2)-like viruses. In 1997/1998 season, the trunk split into three branches. Thereafter in the 2003/2004 and 2005/2006 seasons, the trunk further split into two branches. Although these viruses may come from different sources via latency or migration, they were co-circulating in the same season. In the case of NAs, a multi-furcated tree was formed (Figure 2). The tree was interrupted in the 2001/2002 season and the 2003/2004 season.

Generally, the occurrence of rearrangement can be deduced by the fact that different gene fragments isolated from the same ancestor strain are located at different positions of the phylogenetic tree. We found several possible recombinant events: A/Hong

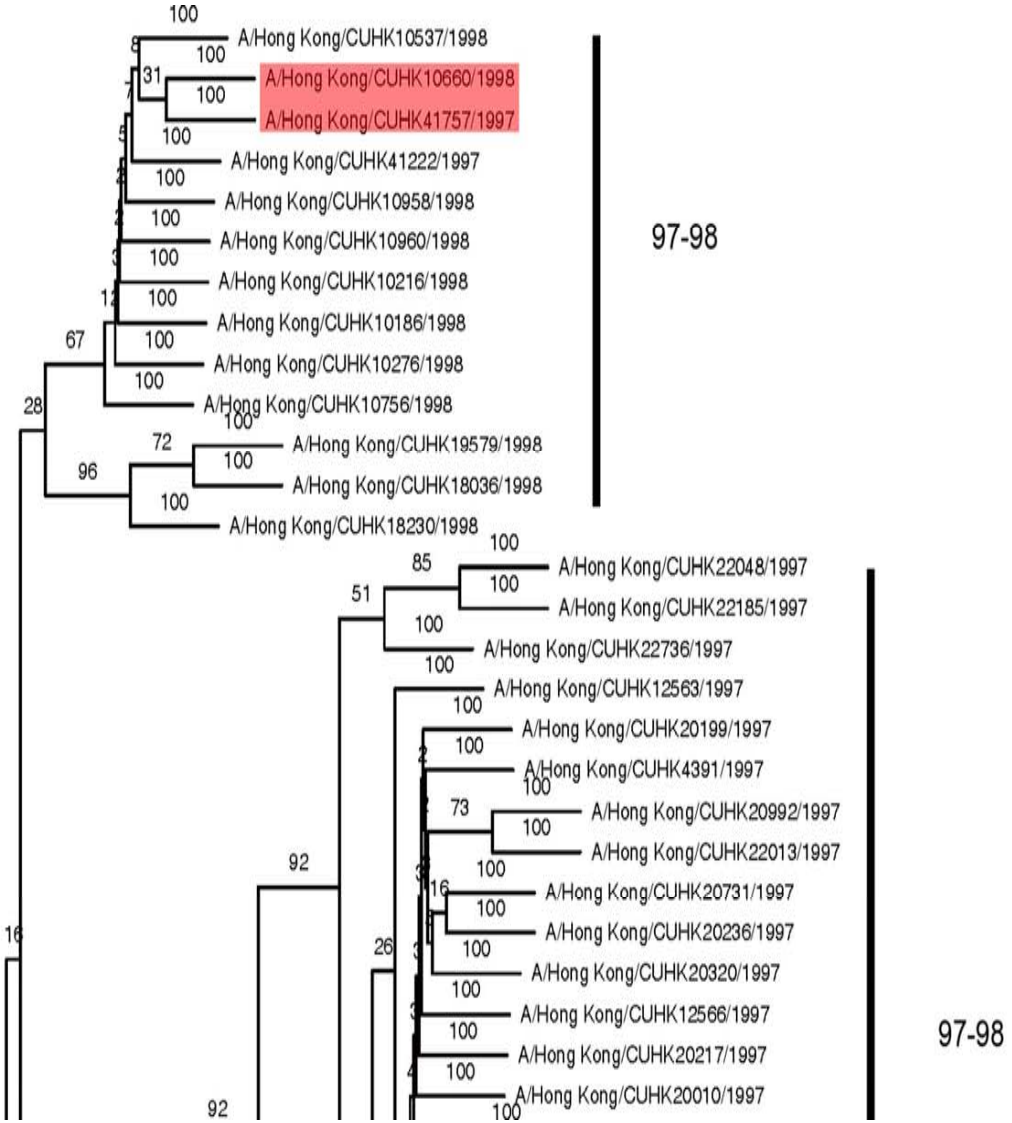


Figure 1. A maximum likelihood tree of 335 HA sequences from Hong Kong (1997-2006) with the HA sequences of 7 WHO seasonal influenza vaccine strains (A/California/7/2004, A/Fujian/411/2002, A/KUMAMOTO/102/2002, A/Moscow/10/99, A/Sydney/5/97, A/Wyoming/03/2003 and A/Wisconsin/67/2005, all of them are yellow highlighted in yellow boxes). The sequences were aligned by BioEdit and edited by MEGA4. The tree was constructed by PHYML3.0 package and SPR branch-swapping with the HKY85 model of nucleotide substitution, and displayed using FigTree, rooted at A/Sydney/5/97. The red boxes highlight sequences from different years in the same branch. The green boxes highlight sequences from different year to the season sequences located. The blue boxes highlight strains which may be generated from rearrangement events.

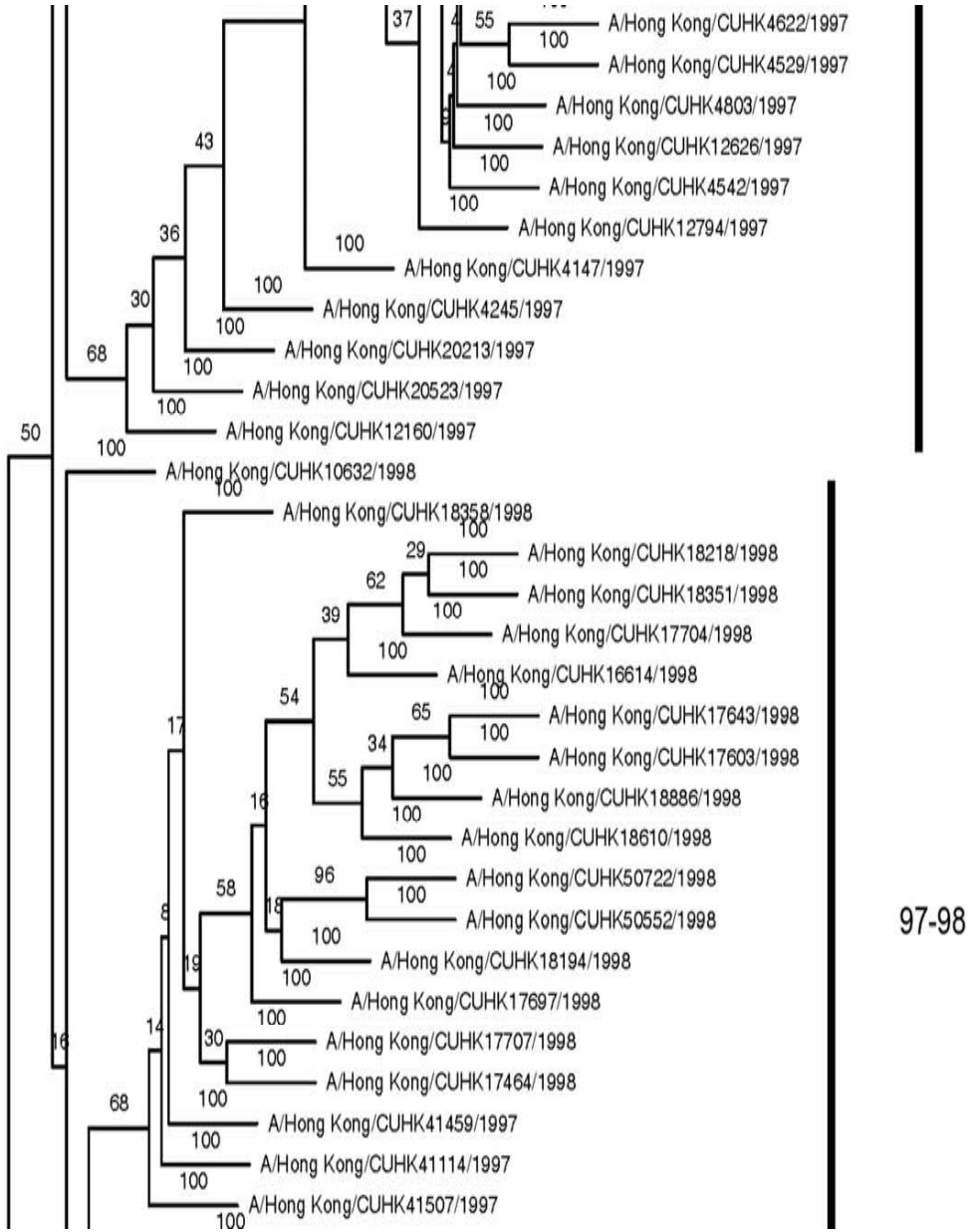


Figure 1. Contd.

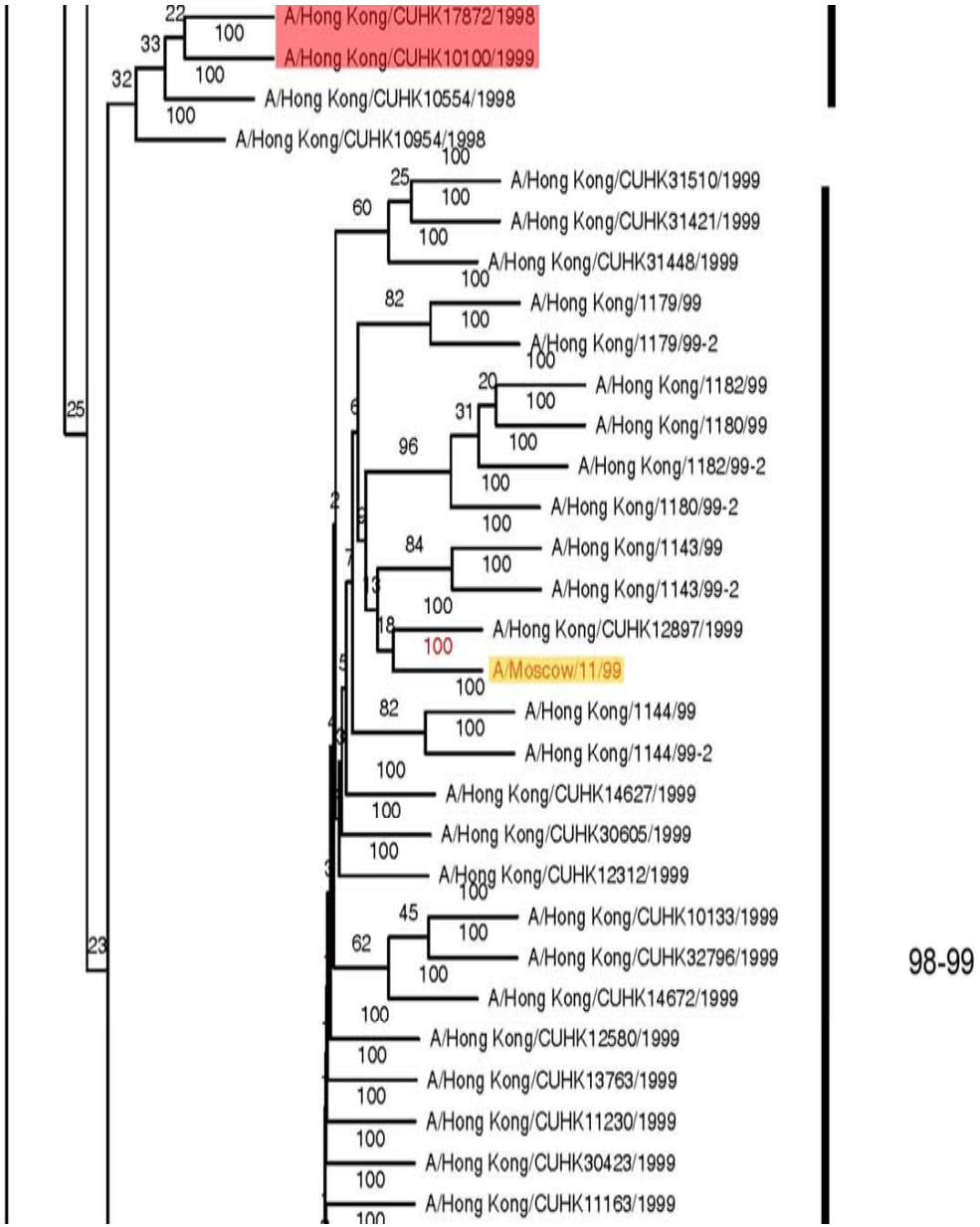


Figure 1. Contd.

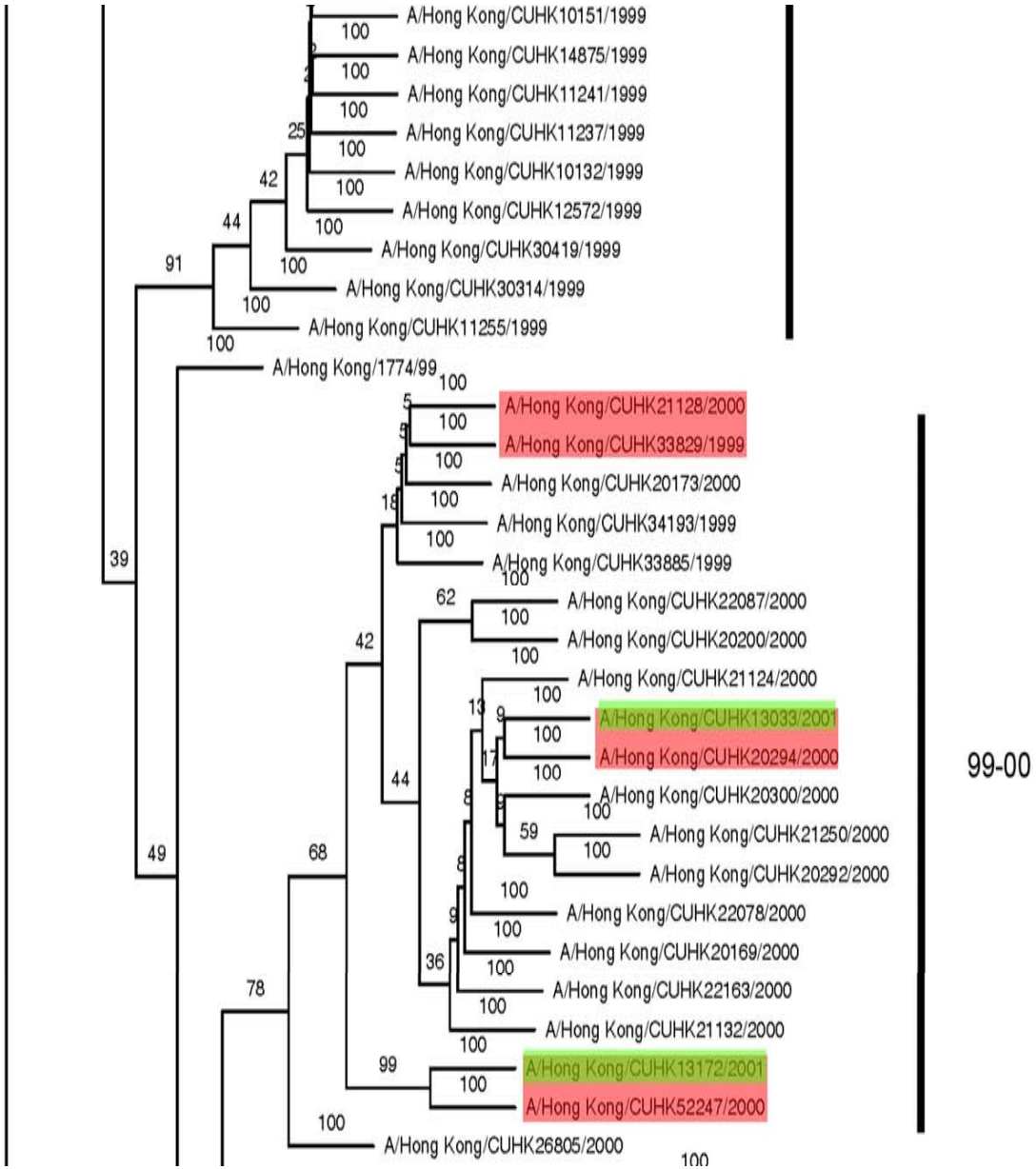


Figure 1. Contd.

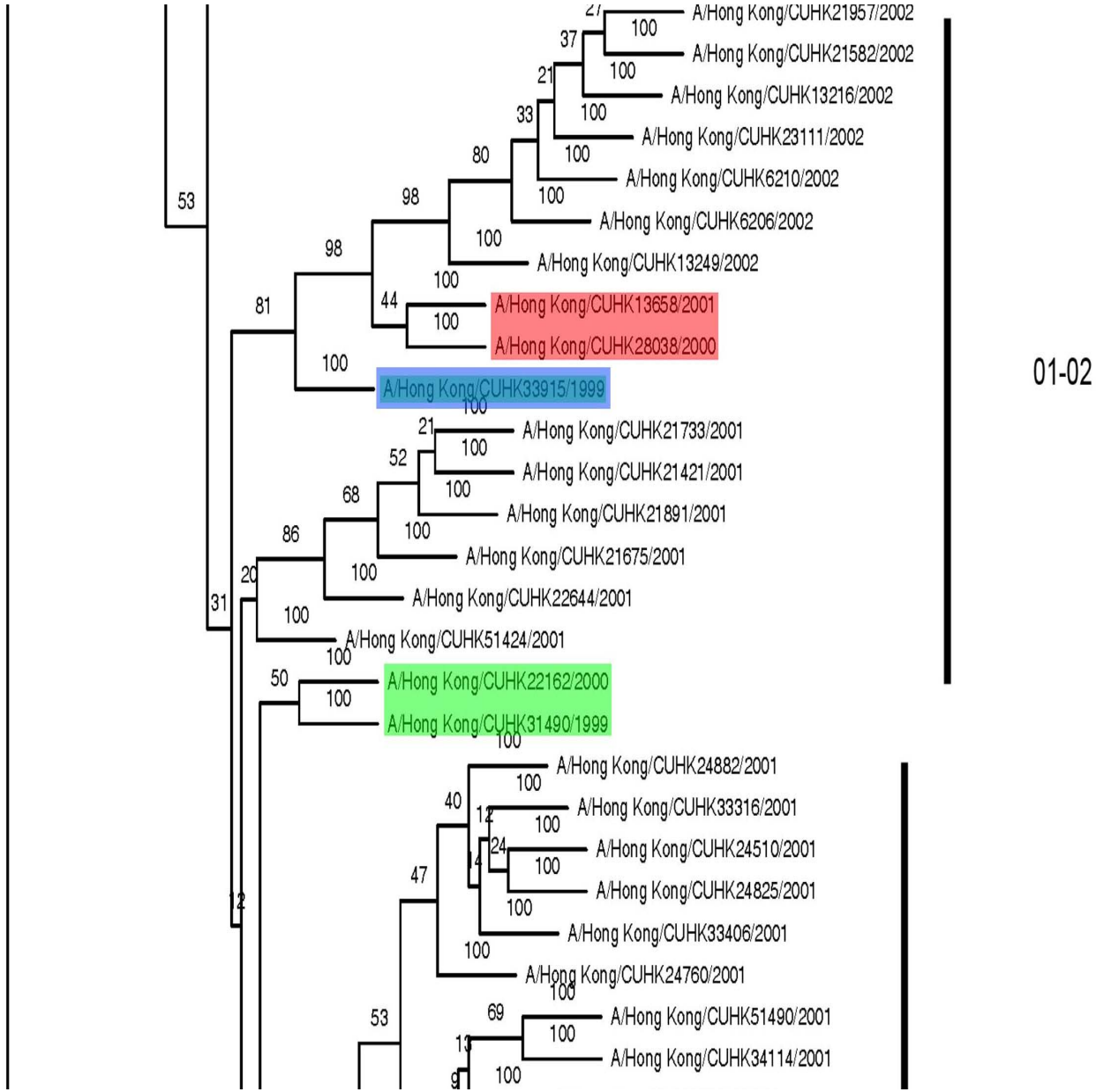


Figure 1. Contd.

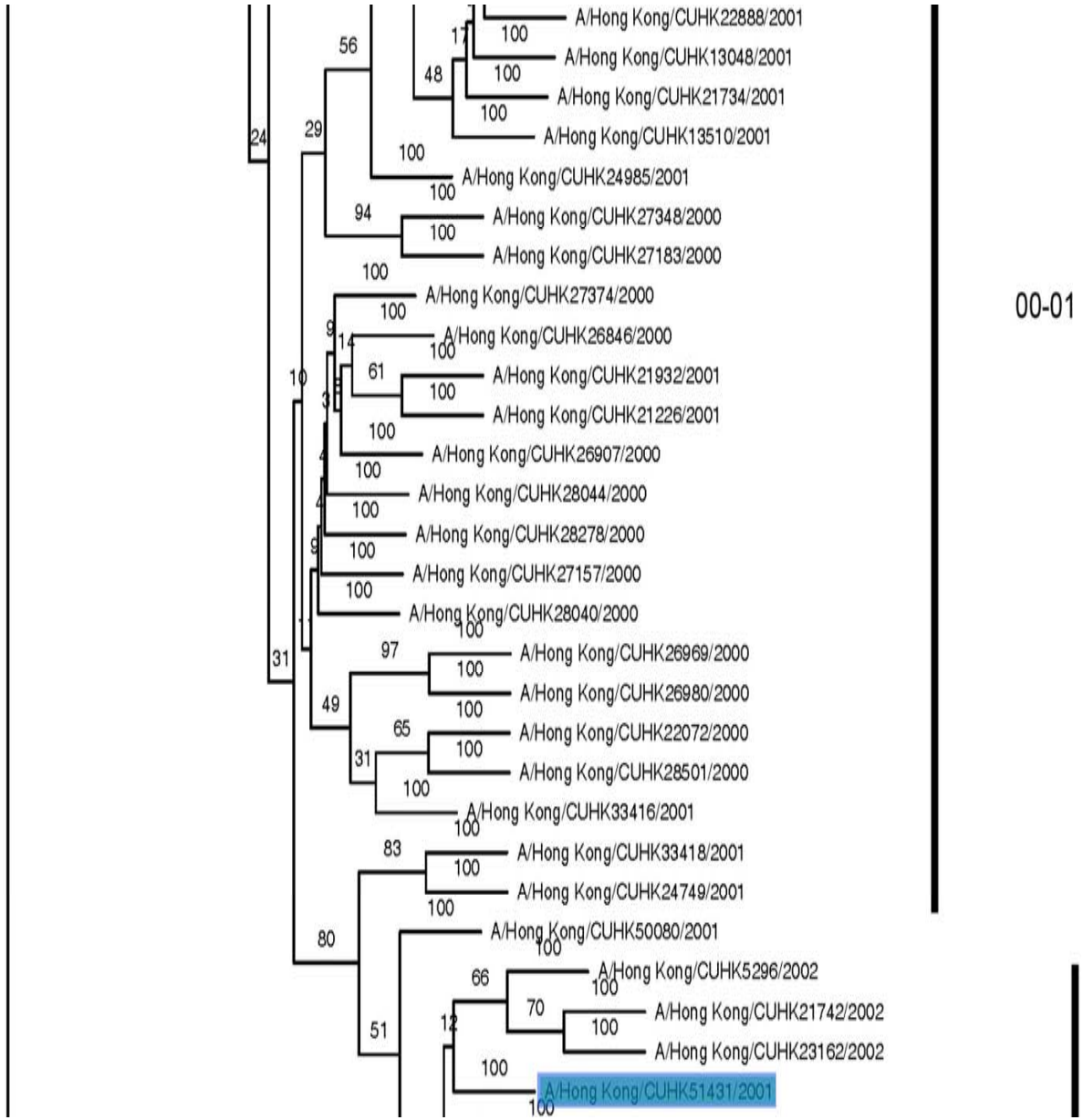


Figure 1. Contd.

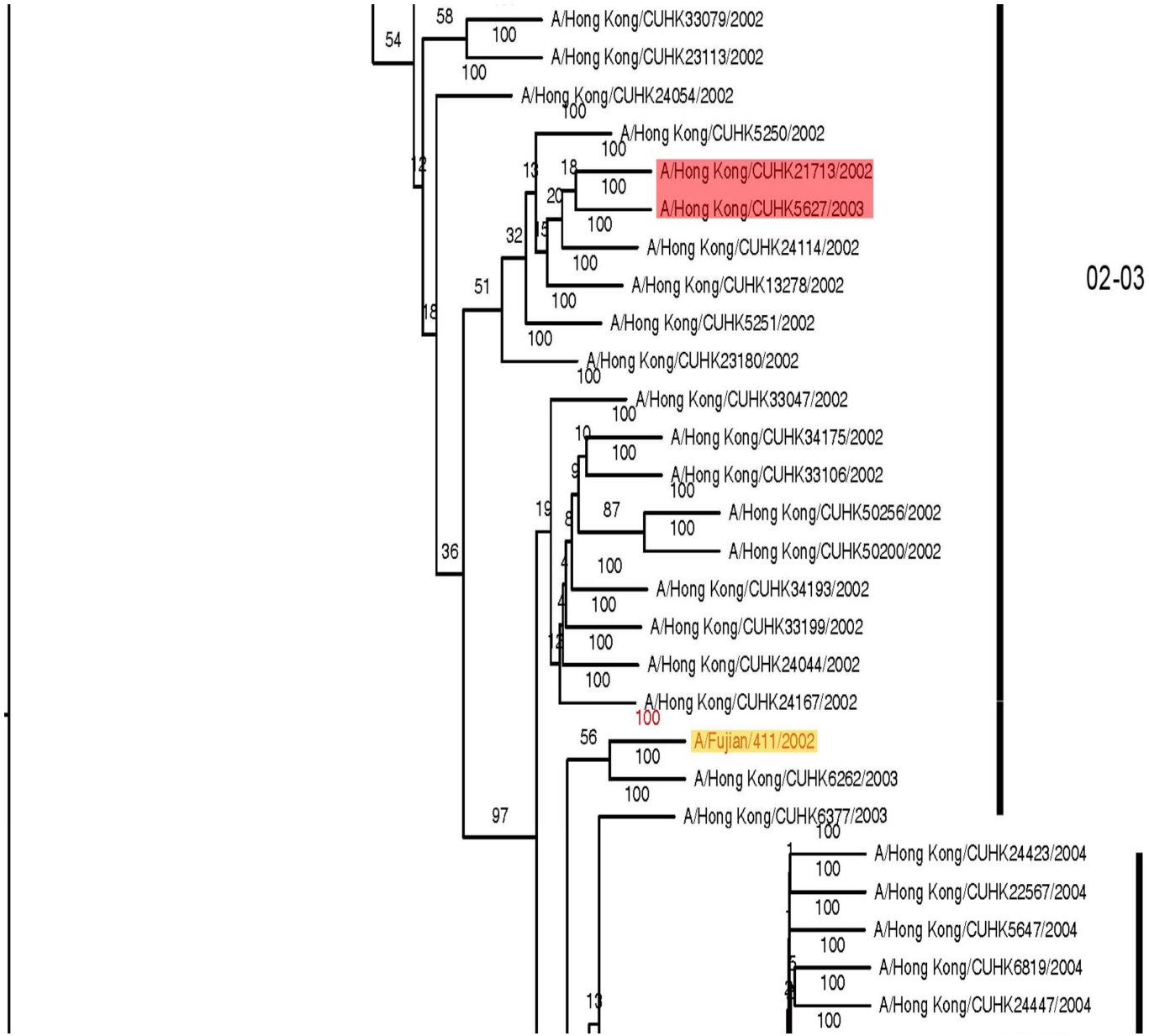


Figure 1. Contd.

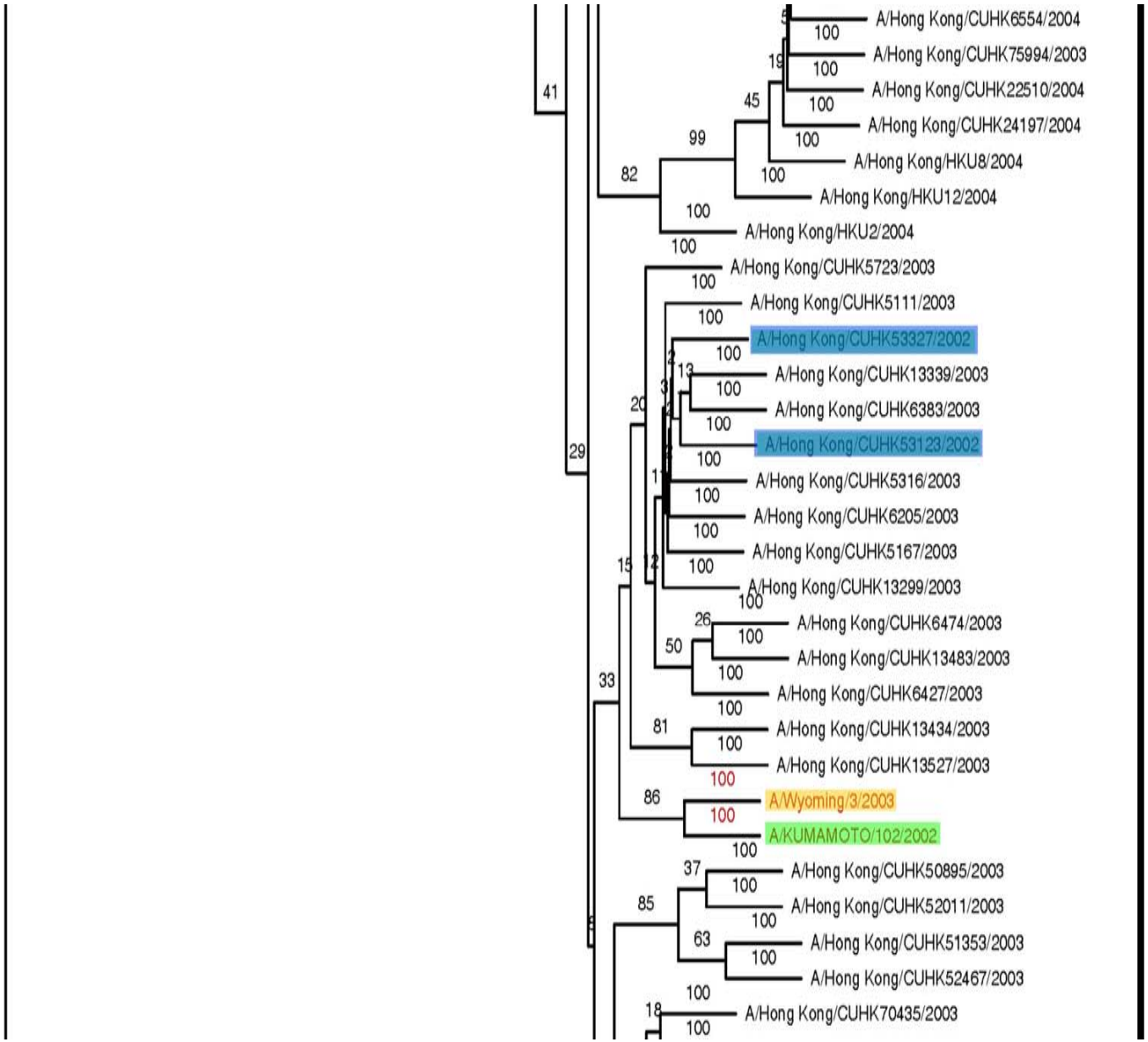


Figure 1. Contd.

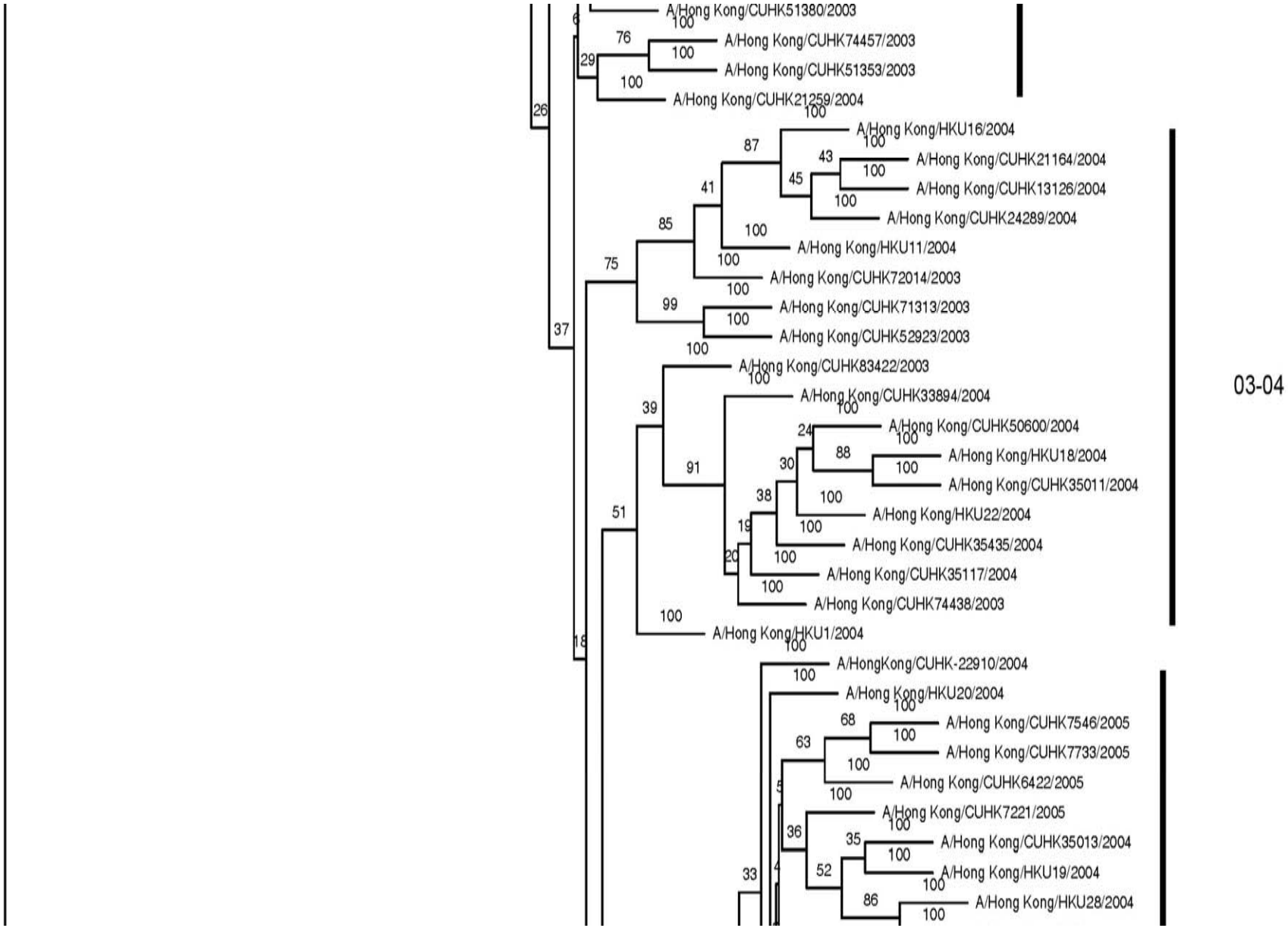


Figure 1. Contd.

Figure 1. Contd.

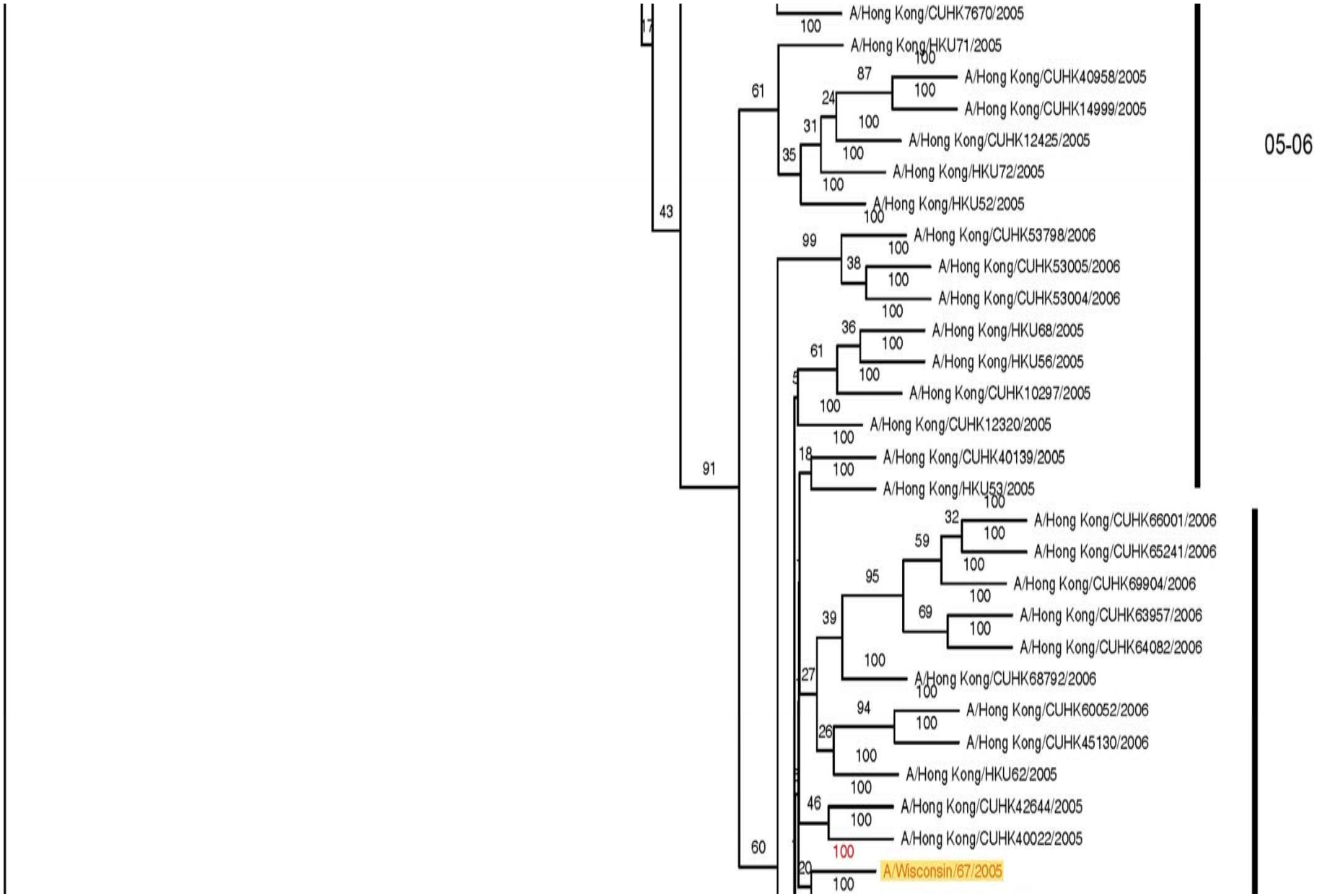


Figure 1. Contd.

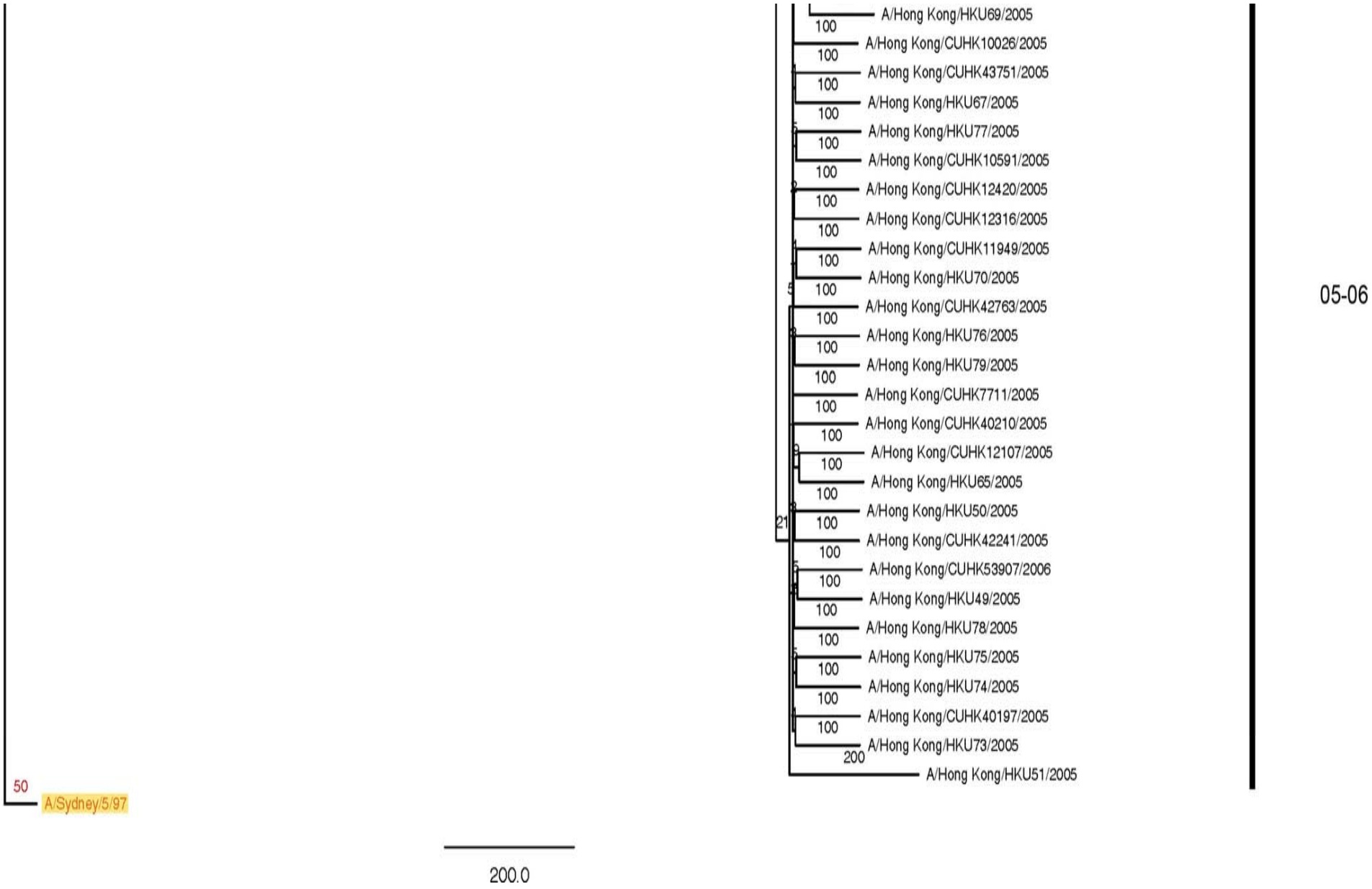


Figure 1. Contd.

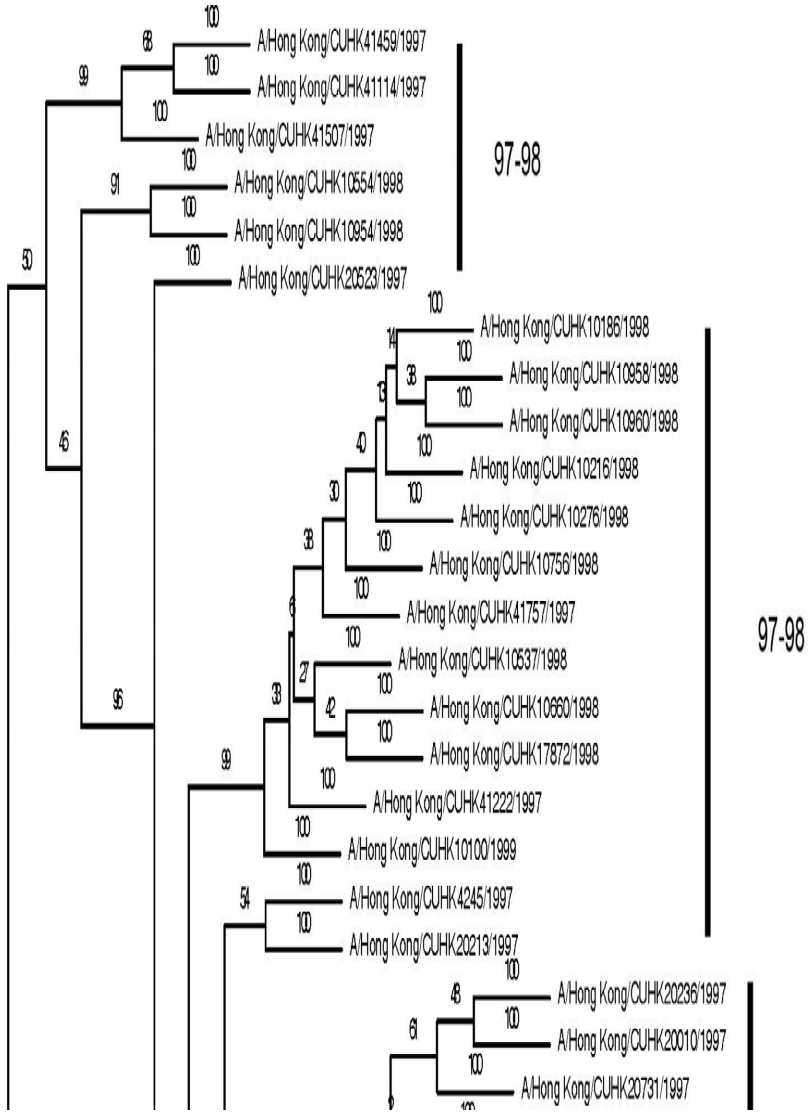


Figure 2. A maximum likelihood tree of 334 NA sequences from Hong Kong (1997-2006) with the HA sequences of 4 WHO seasonal influenza vaccine strains (A/California/7/2004, A/Moscow/10/99, A/Sydney/5/97 and A/Wisconsin/67/2005, all of them are yellow highlighted in yellow boxes), The sequences were aligned by BioEdit and edited in MEGA4. The tree was constructed using PHYML3.0 package and SPR branch-swapping with the HKY85 model of nucleotide substitution, and displayed using FigTree, rooted at A/Sydney/5/97. The red boxes highlight sequences from different years in the same branch. The green boxes highlight sequences from different year to the season sequences located. The blue boxes highlight strains which may be generated from rearrangement events.

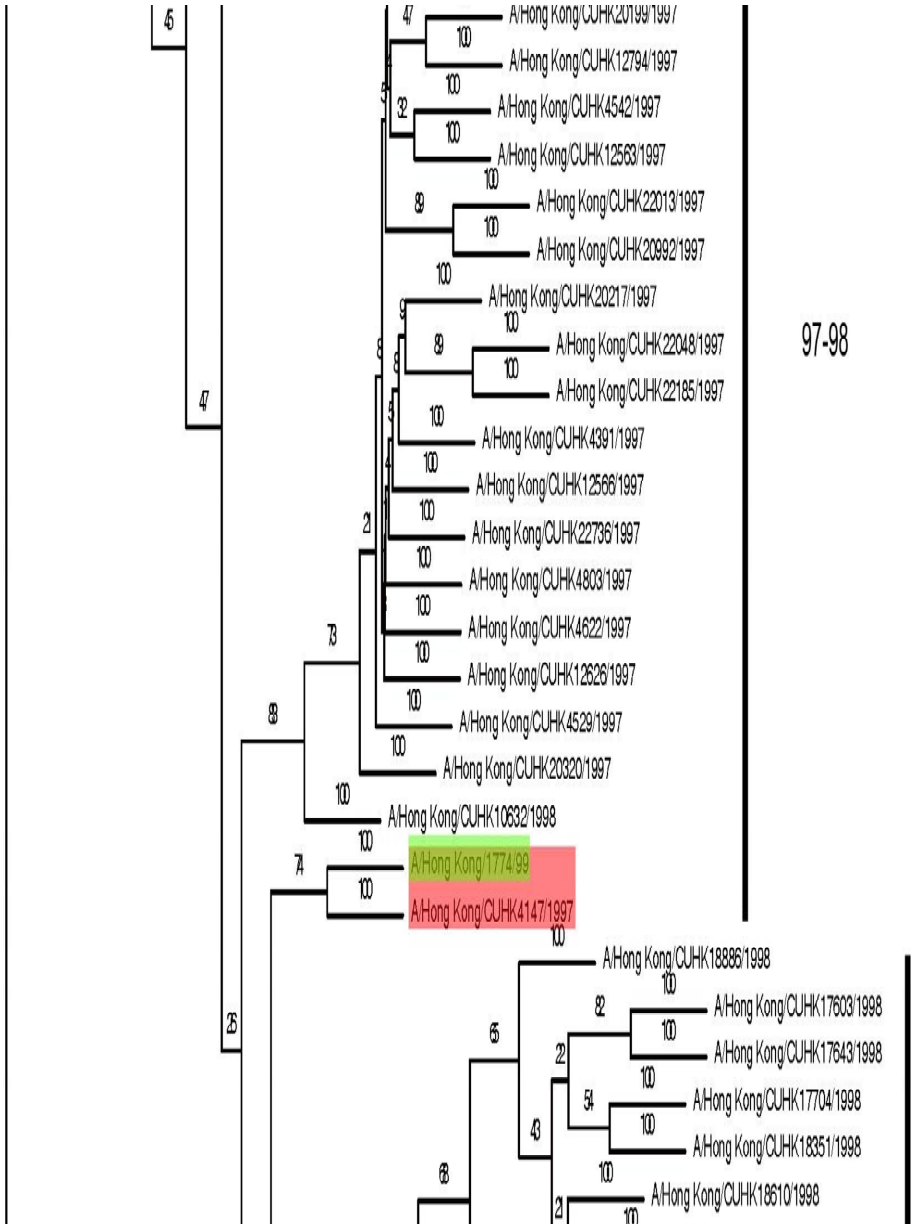


Figure 2. Contd.

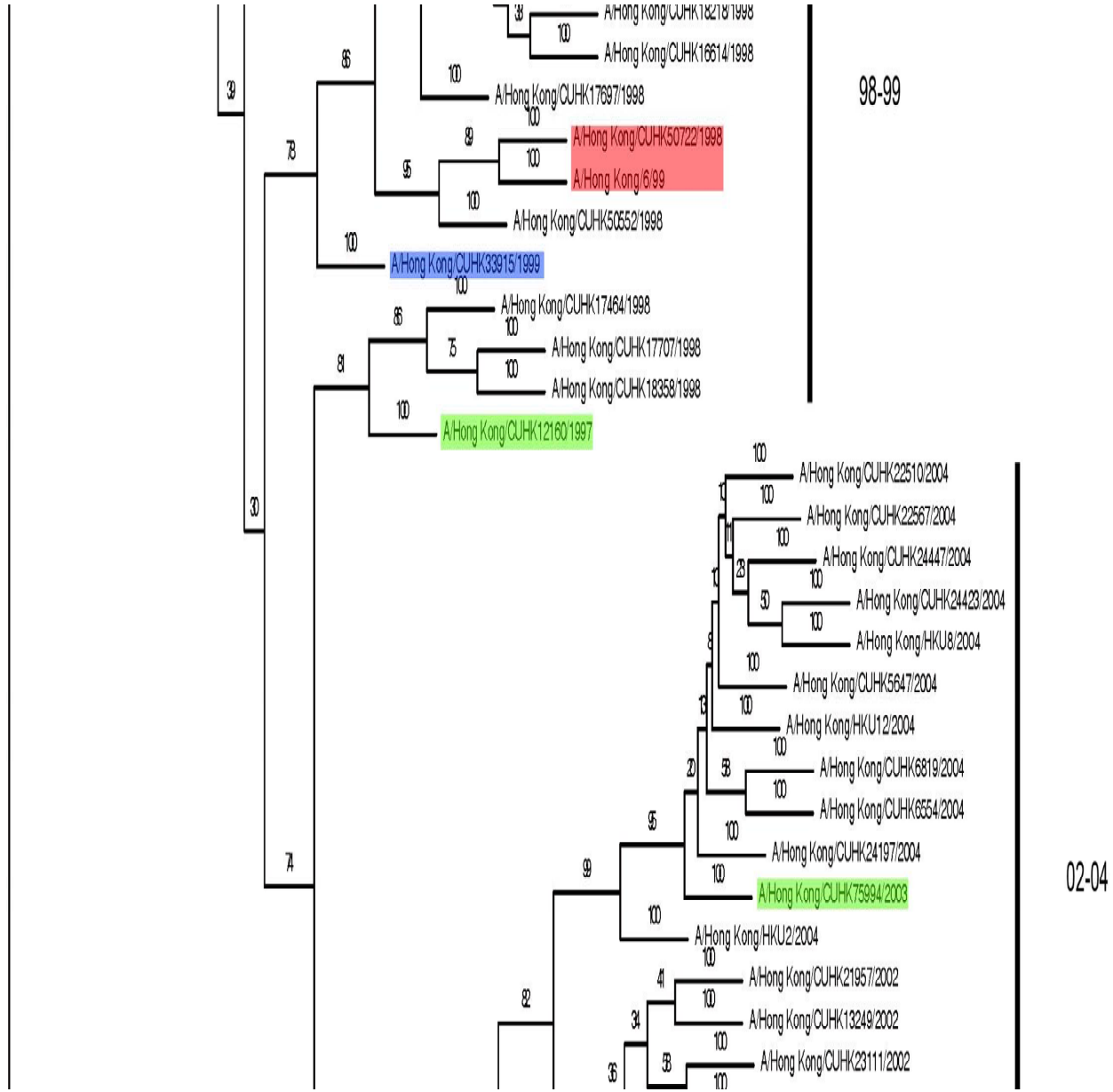


Figure 2. Contd.

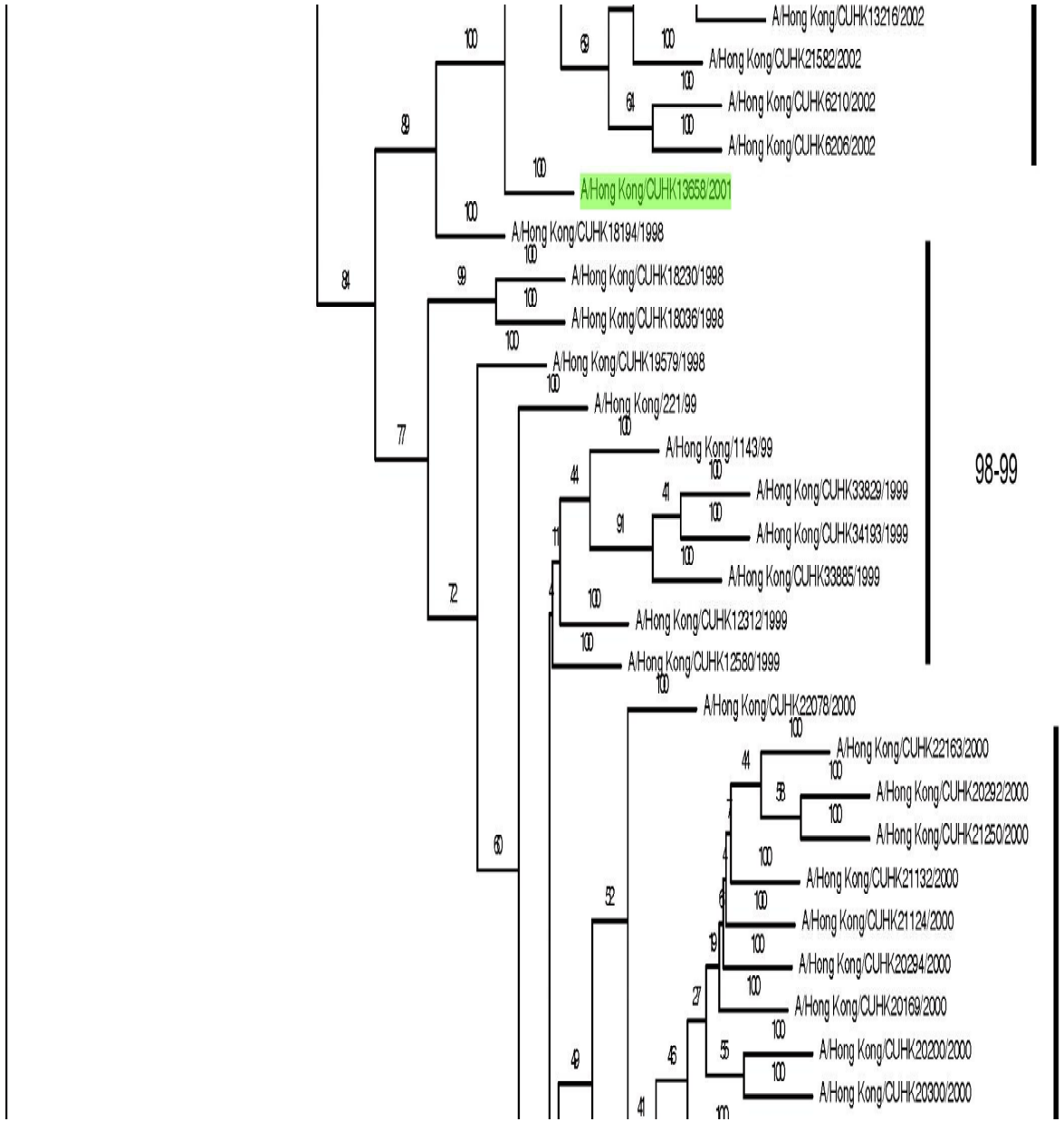


Figure 2. Contd.

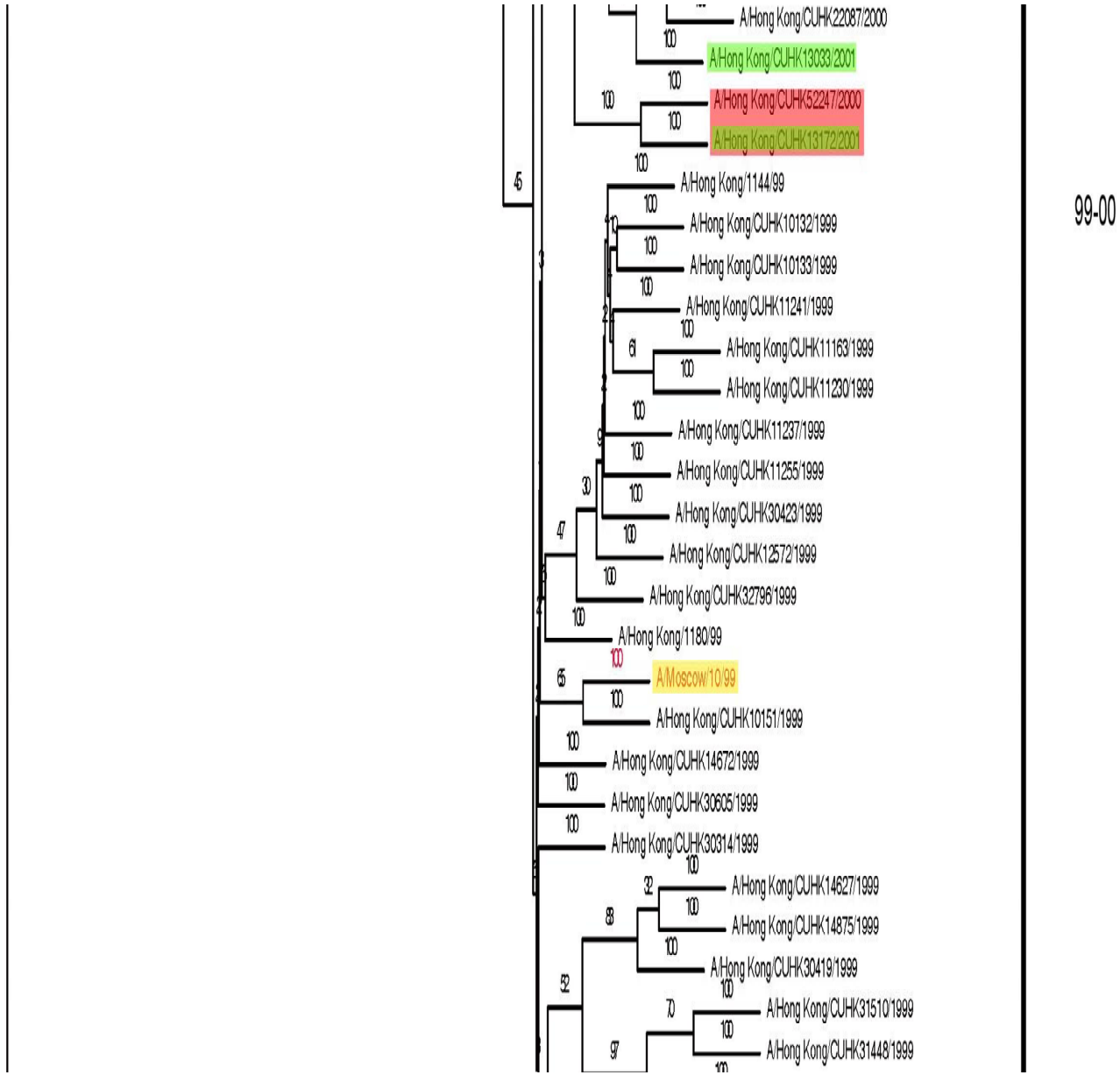


Figure 2. Contd.

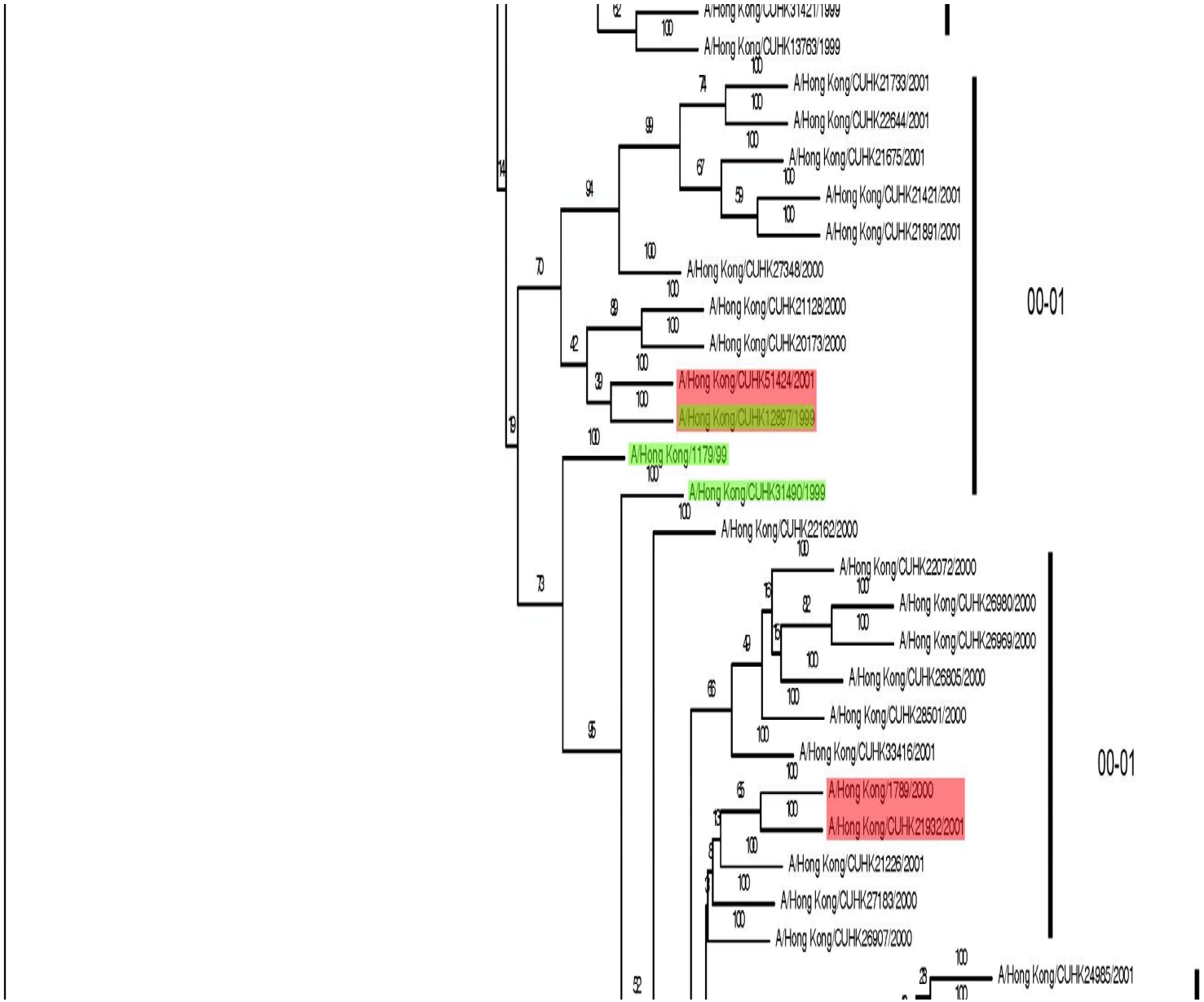


Figure 2. Contd.

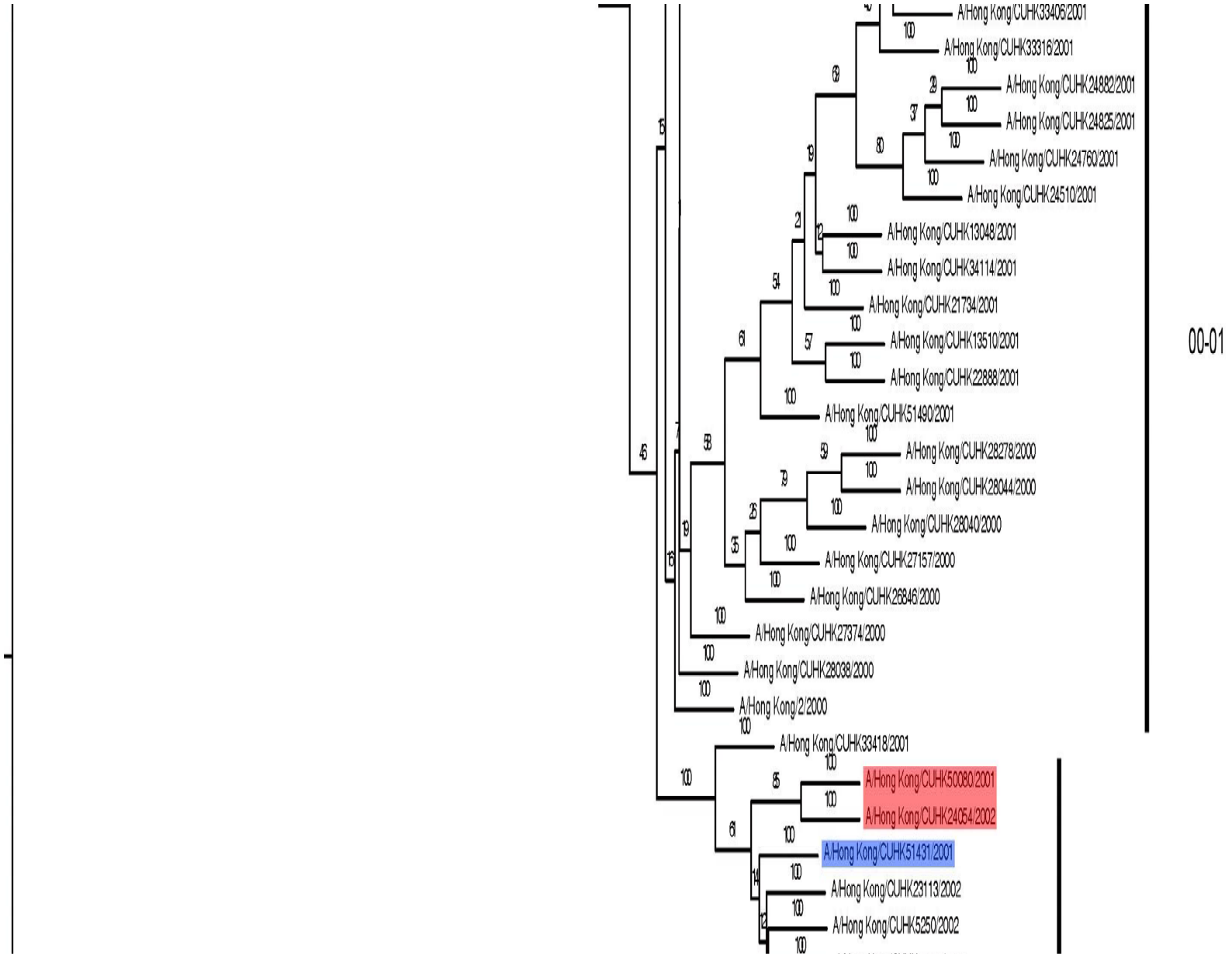


Figure 2. Contd.

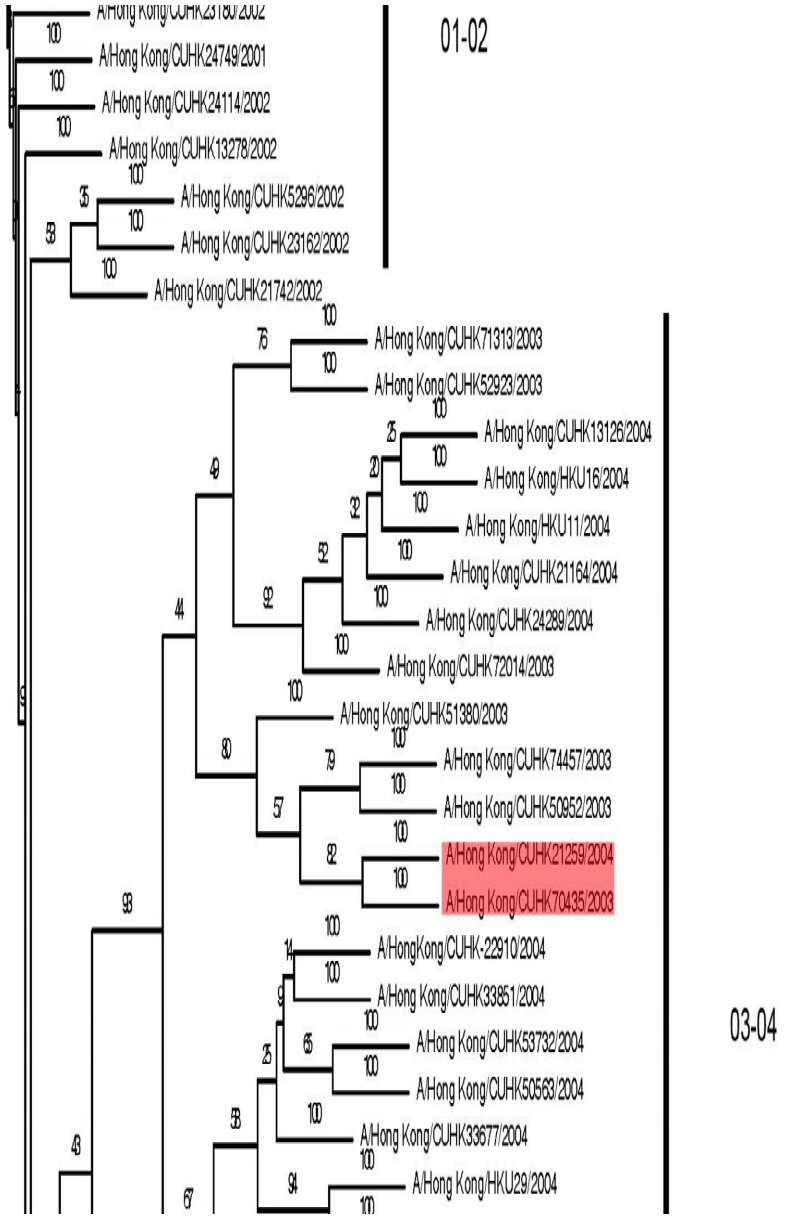


Figure 2. Contd.

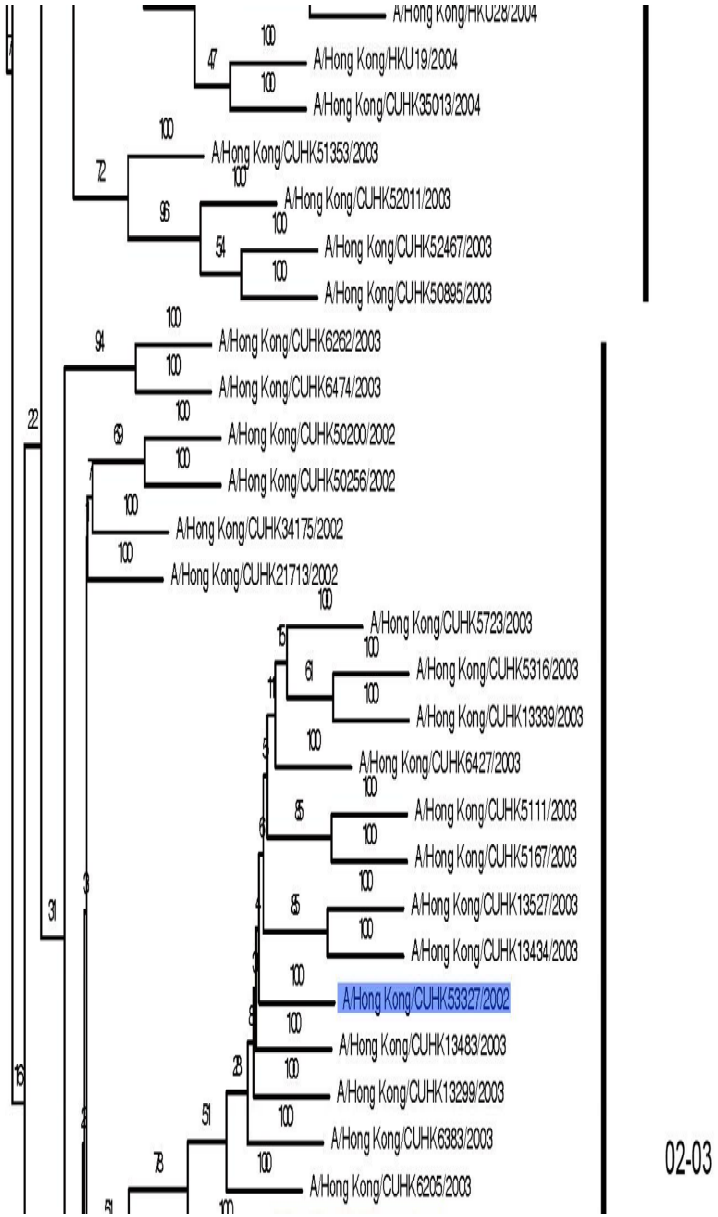


Figure 2. Contd.

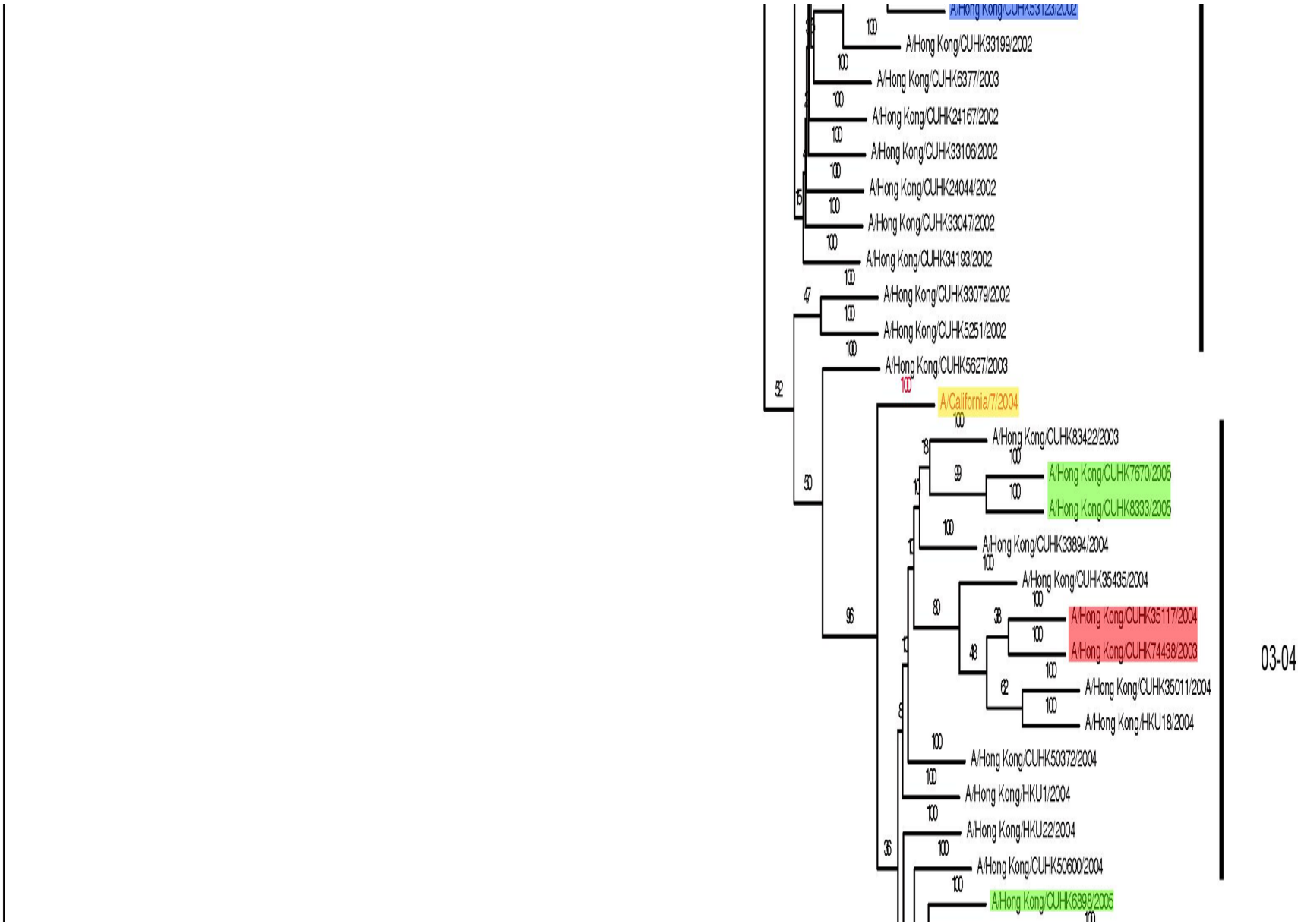
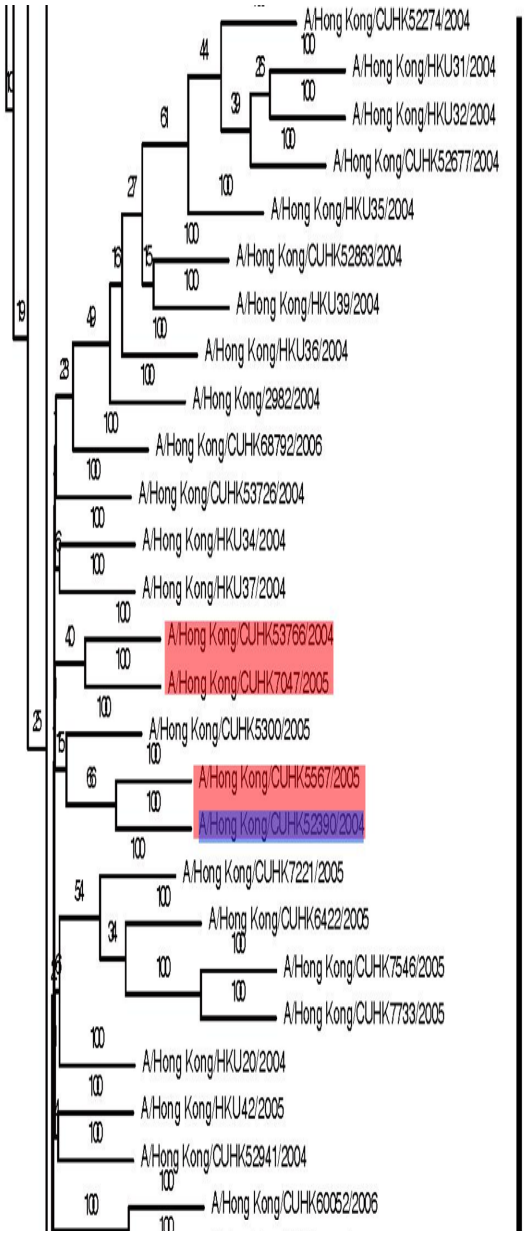


Figure 2. Contd.



04-05

Figure 2. Contd.

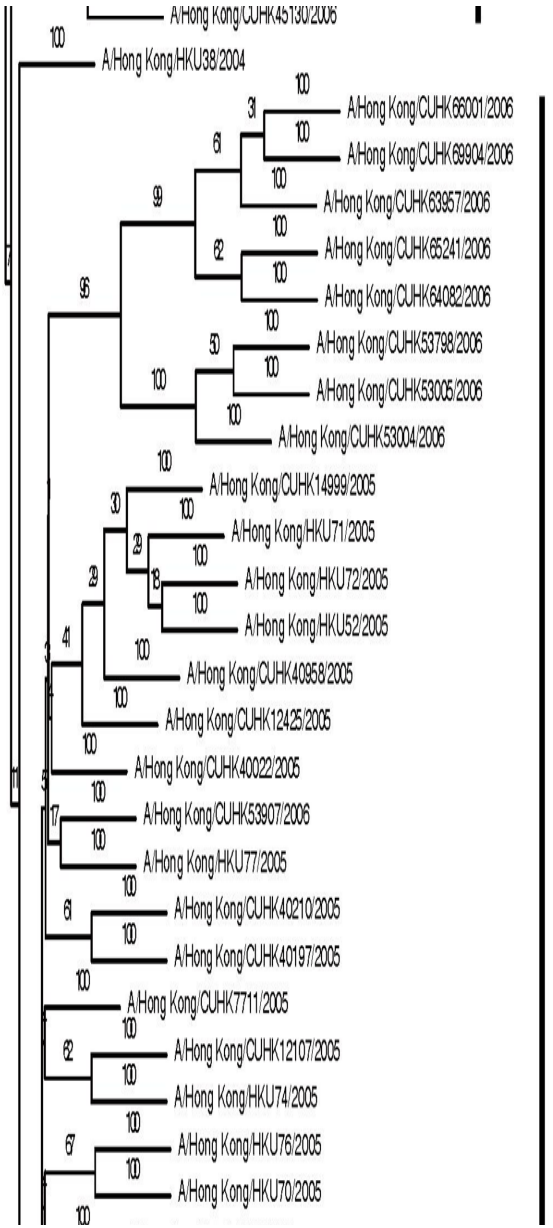


Figure 2. Contd.

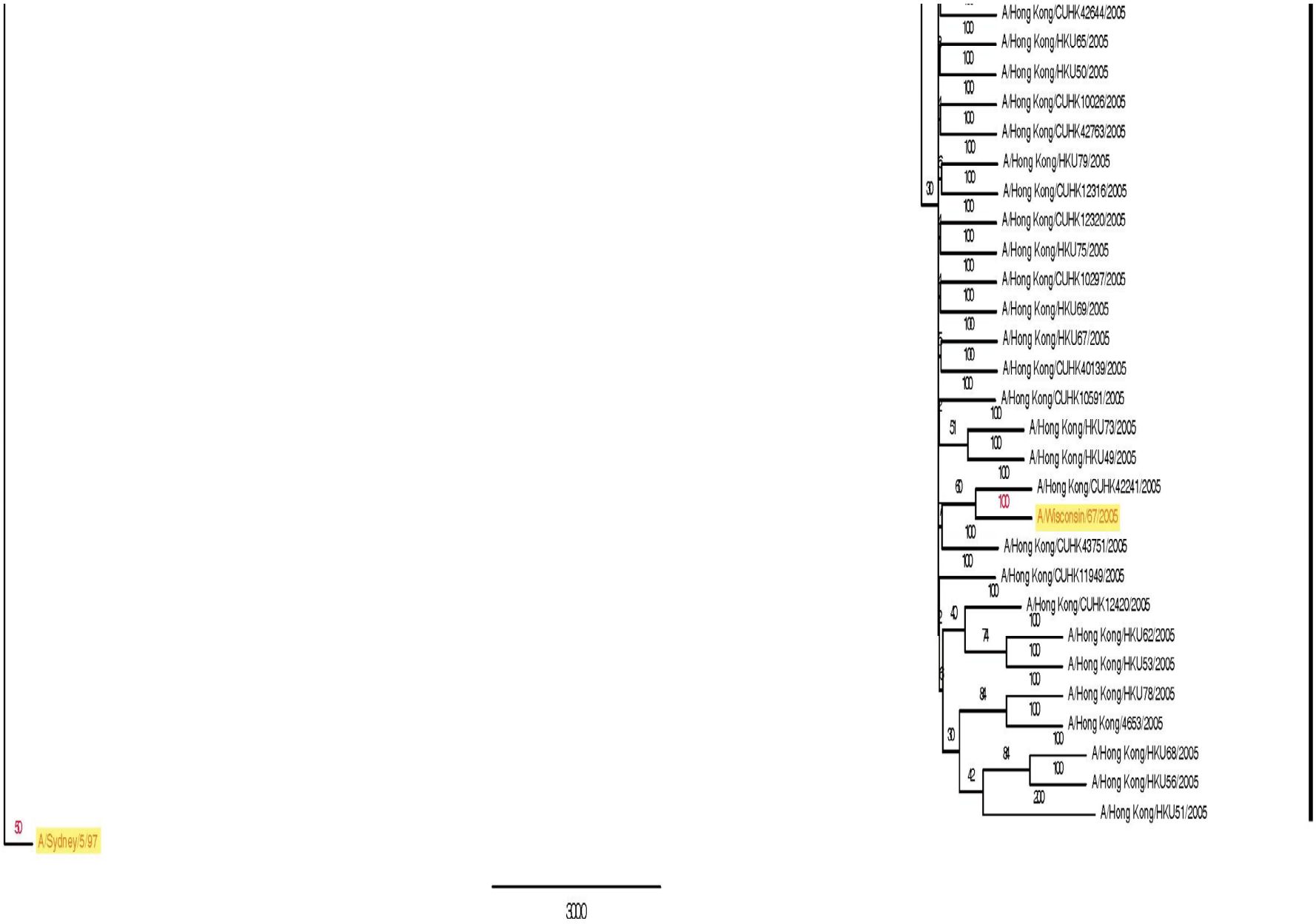


Figure 2. Contd.

Table 1. HA's Amino acid variations between sublineages of A/H3N2 viruses in Hong Kong.

Amino acid position	97-98	98-99	99-00	00-01	01-02	02-03	03-04	04-05	05-06
5	G		V	G					
25	L					L(I)	I		
33	Q		H	Q					
50C	R			G			G(ERV)	G	
75E	H					H(Q)	Q		
83E	E			E(K)	E	K			
131A	A					T(A)	T		
137A	Y(F)	S							
144A	I(V)	I		N	N/D	N(D)	N		
145A	K				K(N)	K	K(N)	N	
155B	H				H(P)	H(T)	T		
156B	Q(K)	Q					H		
159B	Y						Y(F)	F	
160B	K	R	K						
172D	D		E						
186B	S				G(S)	G			
189B	S						S(N)	N	
192B	T		I						
193B	S						S(N)	S	F
202	V					I			
222	W				R				
225	G			D	G	D			N
226D	V							I	
227D	S						S(P)	S(P)	P
271	D		N	D					
326	K							T	K
347	V	M(K)	V		M/V	V			
361	T								I
386	E	E		G					
530	V					A(V)	A		

Amino acids in brackets indicate less than half but more than two substitutions at the given amino acid position within a season. A single amino acid change in one position is not shown. Amino acids separated by '/' indicate equal substitutions of either amino acid at the given position. Letters in upper case on the right of an amino acid position indicate the antigenic site location of the residue.

Kong/CUHK33915/1999, A/Hong Kong /CUHK51431/2001, A/Hong Kong/CUHK53327/2002, A/Hong Kong/CUHK53123/2002, and A/Hong Kong/CUHK 52390/2004.

Amino acid variations

Patterns of antigenic site variations are observed by aligning the amino acid sequences of the HAs. Thirty amino acid sites changed, including 18 AA in the epitopes (Table 1). Between four and seven amino acid site season-to-season variations are observed, while two to four sites belong to epitopes. Most amino acid variation sites in the 1998/1999 and 2003/2004 seasons belong to epitopes, while in the 2004/2005 season, variations of four main amino acid sites are epitopes. A/H3N2 was also prevalent in these three seasons. Most amino acid variations in Table 1 were kept unchanged for two additional years. For example, the amino acid position S137 became stable in HAs after the 1998/1999 season; after the 2002/2003 season, positions K83, T131, I202 and A530 had also become stable. Between the seasons of 2003/2004 and 2005/2006, all HAs acquired I25, Q75 and T155. Positions 144 and 225 had the highest variability, which was illustrated by mutations more than twice in addition to a reverse mutation. Reverse mutations took place at several amino acid sites. For example, in the 1999/2000 season, the amino acid at site No. 5 mutated from G to V, but changed back to G in the next season. Reverse mutations took place at sites 225 and 347 twice.

Table 2. NA's amino acid variations between sublineages of A/H3N2 viruses in Hong Kong.

Amino acid position	97-98	98-99	99-00	00-01	01-02	02-03	03-04	04-05	05-06
18	A			S(A)	S				
23	L			F					
30	V				I				
42	C				F				
93	K	K(N)	N			N(D)	N	D	N
143	R		G		V				
197B	H		D						
199B	E						E(K)	K	
208	D	D(N)	N						
215	I						V	I	
216	G				V		V(I)	V	
221B	K							E	
249	R		K						
265	I	T(I)	T						
267	Q(P)	K	P(L)	T					S
307	V			I(V)	I				
385A	K				N				
399A	D	E(D)	D						
431	K						N	K	
432	Q							E	
437	W		L						

Amino acids in brackets indicate less than half but more than two substitutions at the given amino acid position within a season. A single amino acid change in one position is not shown. Amino acids separated by '/' indicate equal substitutions of either amino acid at the given position. Letters in upper case on the right of an amino acid position indicate the antigenic site location of the residue.

There are 21 amino acid variation sites in NAs, and only five sites belong to epitopes. This number is much smaller than that in HA's (Table 2). Five epitope sites were found in the 1998/1999, 1999/2000 and 2004/2005 seasons. The sites D197, N208, K249, D399 and L437 in amino acid sequence of NA had become fixed after the 1999/2000 season. All NAs possessed I30, F42, V143, V216, and N385 after 2001/2002, but not those amino acids which had been stable before 2001. After the 2004/2005 season, All NAs acquired K199, I215, E221, K431 and E432. Positions 93, 143 and 267 had the highest variability, shown by more than two different amino acids. Reverse mutations were also found in NAs, such as site 93, which mutated from N to D in the 1999/2000 season and reversed back to N in the next season. In addition, reverse mutations also took place at sites 399 and 431.

We found five HA major amino acid variation sites in the 1998/1999 season. The prevalence of A/H3N2 viruses increased dramatically in the 1998/1999 season in association with mutations in residues Y137S at site A, K160R at site B, and V347M in HA proteins. The preferred variable antigenic sites between strains in the successive seasons are A (Pepitope=0.280) and B (Pepitope=0.255). By comparing the sequences between the 2001/2002 and 2002/2003 seasons, we found seven major amino acids variation sites, including three epitopes, 83E, 144A, and 131A. The preferred antigenic variation site is A (Pepitope=0.202). By comparing the sequences between the 2003/2004 and 2004/2005 seasons, we found four major amino acids variation sites, all of which are epitopes: 145A, 159B, 189B and 226D. We further inferred that the preferred antigenic sites between the 2003/2004 and 2004/2005 seasons are B (Pepitope=0.543). In the 2003/2004 season, 25, 75 at site E, 15 at site B and 156 at site B changed from L25, H75, H155 and Q156 to I25, Q75, T155 and H156 in HA protein, respectively. Both H155T and Q156H are located at antigenic site B. T155 and H156 amino acids have been maintained in all Hong Kong isolates after the 2003/2004 season. Positions 5, 33 and 271 changed from G5, Q33 and D271 to V5, H33 and N271 in the 1999/2000 season and further to G5, Q33 and D271 in the 2000/2001 season, respectively. Since then, these residues (G5, Q33 and D271) have remained unchanged. Position 144 at the HA protein changed from I144 to D144 in 2001/2002 and further to N144 in 2002/2003.

There were three major sites that changed in the 1999/2000 season. Among them, 197B and 399A belong to epitopes. The preferred variable antigenic site between strains in the 1999/2000 season and previous strains was site B (Pepitope=0.172). We compared amino acids

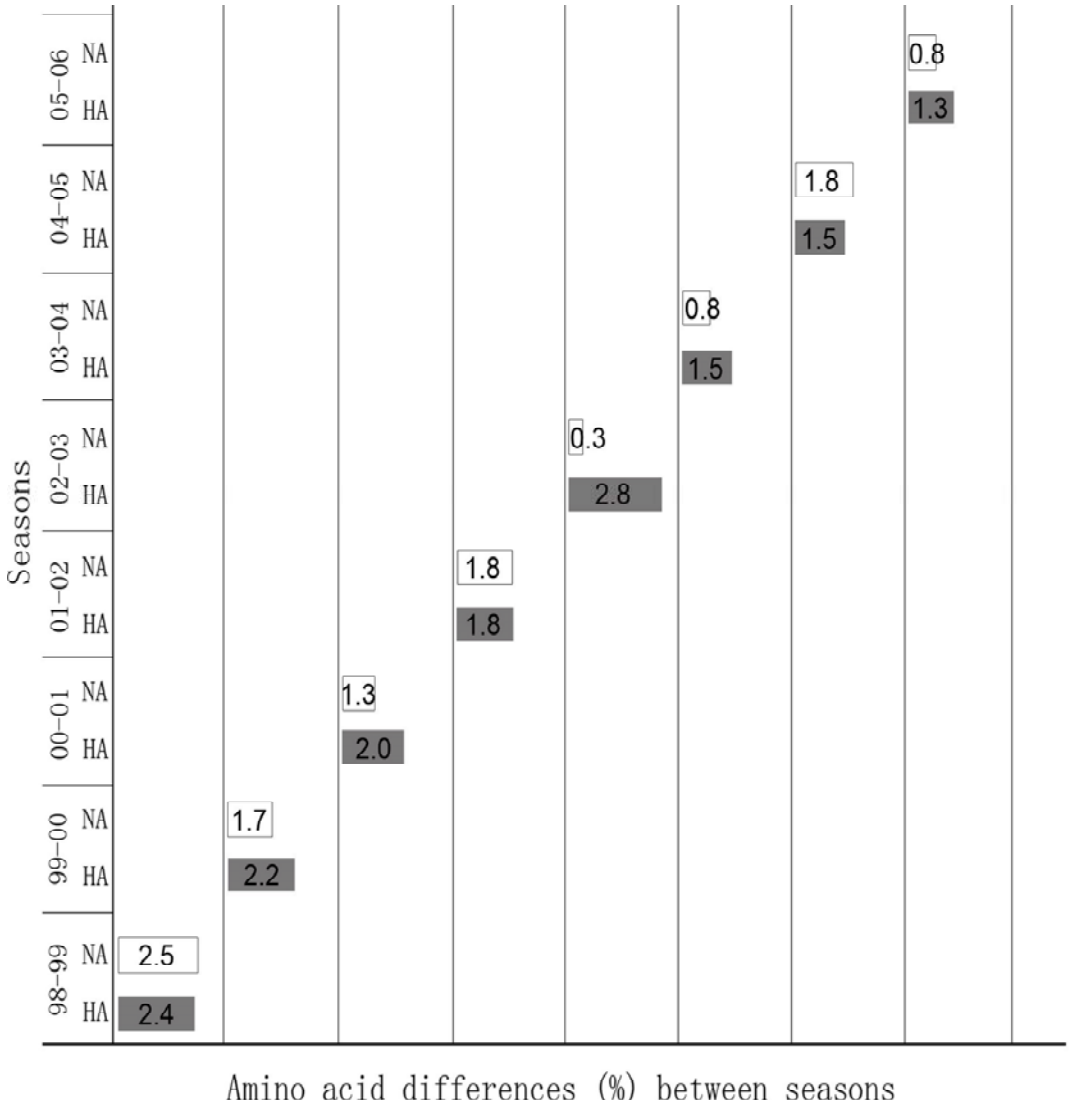


Figure 3. Amino acid distances of HAs and NAs between seasons. Sequences were grouped according to the seasons in Figure 1 and Figure 2. The amino acids distances between groups were computed by MEGA V4 software.

changed in the 2004/2005 season with those in the 2003/2004 season, showing that only 199B and 221B are epitope sites. The preferred variable antigenic site between strains in 2003/2004 and previous strains was site B (Pepitope=0.346). There were four major amino acids that changed from the 1999/2000 season to the 2000/2001 season, without epitope involvements. Residues I265T, Q267K, E399D at site A changed from the 1997/1998 season to the 1998/1999 season in NA proteins. Positions 215 and 431 at the NA proteins changed from I215, K431 to V215 and N431 in the 2003/2004 season and further to I215, K431 in the 2004/2005 season, respectively. Position 225 at the HA protein changed from G to D in the 2000/2001 season, further to G in the 2001/2002 season, then to D in the 2002/2003 season, and finally to N in the 2005/2006 season. Position 93 at the NA protein changed from K to N in the 1999/2000 season, further to D in the 2004/2005 season and finally to N in the 2005/2006 season.

Using seasonal evolutionary patterns of sequences as groups, we computed amino acid distances between groups using the software Mega version 4. The amino acids of HAs varied dramatically during the 1998/1999 and 2002/2003 seasons, in which A/H3N2 were prevalent (Figure 3). Generally, there were less amino acid variations when prevalence was lower, but more variations in the higher prevalence and turning seasons. By comparing with the previous season, HAs had a shorter distance in the 2004/2005 season, while NAs in the 2001/2002 season had a longer distance. The turning

Table 3. Log-likelihood values and parameter estimates for the selection analysis of HA and NA genes of A/H3N2 viruses in Hong Kong.

Gene	Model	P^a	L^b	*P* value for LRT	Estimates of parameters[c]	Avg ω	tS/tV
HA	M0 (one ratio)	1	-5614.03		0.31708	0.31708	4.62388
	M3 (discrete)	5	-5549.03	<0.0001	p_0=0.84944, p_1=0.13518, (p_2=0.01538); ω_0=0.11003 ω_1=1.27794, ω_2=3.73484	0.32366	4.54662
	M1a (Neutral)	2	-5556.62	0.0006	P_0=0.78733 (P_1=0.21267); ω_0=0.07346 (ω_1=1.00000)	0.27051	4.33857
	M2a (*P*ositive Selection)	4	-5549.14		P_0=0.81725, P_1=0.15479 (P_2=0.02796); ω_0=0.09758 (ω_1=1.00000), ω_2=3.11037	0.32150	4.53517
	M7 (beta)	2	-5559.50	<0.0001	p=0.10084, q= 0.25511	0.28331	4.41238
	M 8 (beta & w)	4	-5549.47		p_0=0.94592, p=0.40164, q=1.60989 (p_2=0.05408) ω_2=2.52155	0.32266	4.53860
NA	0	1	-5405.14	<0.0001	0.26784	0.26784	5.10390
	3	5	-5309.03		p_0=0.79903, p_1=0.19377 (p_2=0.00720); ω_0=0.05821, ω_1=0.94501, ω_2=8.14060	0.28824	5.36253
	1	2	-5331.60	<0.0001	P_0=0.81795 (P_1=0.18205); ω_0=0.05988 (ω_1=1.00000)	0.23103	5.03506
	2	4	-5309.07		P_0=0.81082, P_1=0.18203 (P_2=0.00714); ω_0=0.06291 (ω_1=1.00000), ω_2=8.21558	0.29170	5.38093
	7	2	-5333.81	<0.0001	p=0.11021, q= 0.36840	0.23035	5.02546
	8	4	-5309.33		p_0=0.99264, p=0.13993, q=0.48039 (p_2=0.00736) ω_2=7.98379	0.28270	5.33019

LRT: log likelihood ratio test for each comparison, **a.** The number of free parameters in each model, **b.** Log likelihood, **c.** The parameters in parentheses are not free parameters.

season in the 2002/2003 season had fewer variations than other seasons (such as in 1998/1999) in NAs, but HAs had the longest distance.

Positive selection

The average dN/dS (avg. ω) of HA's amino acid sequences of A/H3N2 virus ranges from 0.27 to 0.32 in all codon substitution models (Table 3). Thus, a non-synonymous substitution rate is approximately 27-32% that of synonymous mutations. The selections in M3, M2a and M8 models successfully detected positively selected sites and provided a much better fit to the data than with the alternative one ratio M0, neutral M1a and beta distribution M7 models, respectively, as determined by

LRT. At the 95% threshold level, only three amino acid sites, i.e., residues 50(antigenic site C), 144 (antigenic site A) and 500, are under positive selection in M3, M2a, M8 together. Residue 193 is also under positive selection in M3 and M8.

The average dN/dS (avg. ω) of NA's amino acid sequences of A/H3N2 virus ranges from 0.23 to 0.29 in all codon substitution models (Table 3). Thus, a non-synonymous substitution rate is 23-29% as likely being fixed as synonymous mutations. At the threshold 95%, only three amino acid sites, namely residues 113, 259 and 361, are under positive selection in M3, M2a, and M8 simultaneously. Unlike HAs, none of the three amino sites is at antigenic sites.

Glycosylation patterns

Point mutations can result in gain or loss of Asn-X-Thr/Ser motifs and therefore, gain or loss of N-glycans, which lead to the alteration of antigenicity and receptor specificity of HAs. We analyzed the glycosylation patterns. Nine putative N-glycosylation sites (residues 22, 38, 63, 122, 126, 133, 165, 246, and 285) are identified in the HAs (Table 4). These glycosylation sites have been conserved in our dataset from 1997 to 2006. An additional site at position 144, due to the I144N substitution within antigenic site A (Wilson et al., 1981), was acquired by the strains collected in the 2000/2001 season. The N-glycosylation site was temporarily lost in the 2001/2002 season, due to an N144D substitution, but later was retained back in subsequent seasons.

In the case of NA, five potential N-glycosylation sites (residues 61, 70, 86, 146, 200 and 234) were conserved throughout the study period. The 93 predicted sequon was partly seen from the 1998/1999 season due to K93N substitution, and was retained in subsequent seasons except 2004/2005 season due to N93D substitution. The 329 predicted sequon was only seen from 1997/1998 season, and lost in subsequent seasons. The potential glycosylation at positions 200 appeared or disappeared partly in 1997/1998 and 1998/1999 seasons.

DISCUSSION

The influenza subtype prevalence in Hong Kong was very similar to those found in other parts of the world where circulation of influenza A virus are often found. In the 1997/1998 season, Sydney-type influenza viruses replaced the dominant Wuhan-type strains (Bridges et al., 2000; Jong et al., 2000), whereas the Fujian-like viruses replaced Sydney strains in 2003/2004 (Organization, 2004). In the 2000/2001 season, while A/H1N1 was the dominating subtype in the world (Lin et al., 2004), the collection of A/H1N1 strains in Hong Kong was higher than in other years. All these showed the vital function of virus disseminations on influenza prevalence.

The monthly epidemics of H3N2 and H1N1 are similar, suggesting that the geographical conditions favorable for the spread of these two influenza types is similar, and in which influenza viruses can easily invade human body. It is interesting to note that when the influenza A/H3N2 strain dominated, influenza A/H1N1 was extremely weak. If a human is infected by one subtype of influenza viruses, he/she will be immune to other influenza virus subtypes (Rambaut et al., 2008).

Two virus strains from many successive seasons have been found in one clade (colored red) (Figures 1 and 2), and multiple sub-lineages in one season. The facts showing the continuity and latency of influenza viruses might be the reasons that different influenza viruses co-circulate within the same season. We also found several completely duplicated sequences from different seasons, which may indicate the continuity of influenza viruses.

The introduction of the strains in the 2001/2002 season caused a "jump" in the evolution of both HA and NA genes. Many of the substitutions in HAs introduced in the 2001/2002 season have become fixed. After the "jump" of influenza virus, the strains experienced a static period. We believe that this reflects a more adapted virus status.

It is worthwhile to note that the strain A/Hong Kong/1774/99 is much similar in antigenic and genetic characteristics to A/H3N2 viruses circulating in pigs in Europe during the 1990s. In addition, it is also closely related to viruses isolated from two children in the Netherlands in 1993 (Rambaut et al., 2008). This highlights the potential of pigs as a vehicle of novel human influenza viruses and the emergence of amantadine-resistant human viruses. The influenza viruses are possibly circulating continuously in the globe.

We found out that specific local epidemic strains in Hong Kong from a certain season could cluster phylogenetically with several strains. This supports the local persistence of influenza strains. However, the seasonal changes of influenza A/H3N2 may be mainly due to one of the two major global migration patterns, including 1) similar viruses appear in different countries at different times, or 2) while one virus is popular within a single location, it circulates continuously within this population, and re-emerges during the next influenza season with relatively little genetic change (Tang et al., 2008).

Generally, there were more amino acid variations of HAs and NAs in the seasons of influenza prevalence. Particularly, amino acids in epitopes changed much more frequently in prevalent seasons (Tables 1 and 2). Moreover, the majority changes of amino acids were observed during the early seasons of a lineage period. Sequence analysis of HA shows high variation in HA1, which may be due to its receptor-binding properties and the sequence being targeted by neutralizing antibodies

since it represents the membrane fusion glycoprotein of influenza virus. Amino acid variations in NA distributed uniformly. We found that some amino acids of the isolates had undergone variations in two successive seasons, demonstrating progressive evolution in each protein segment.

We found the same epitope sites never prevail in either different seasons or in the turning seasons of influenza. In HA, for the 1998/1999 season, the Pepitope values of epitopes A and B are higher than that for the 2002/2003 season, which explain that A/H3N2 strains caused more severe outbreaks in 1998 than in 2002. There were at most seven major amino acid variation sites in the 1999/2000, 2000/2001 and 2002/2003 seasons. We would conclude that the cause for an influenza pandemic may be either amino acid variations of epitopes surpassing a threshold, or the key epitope sites having changed.

Reverse mutations of amino acids show that they may be important for viral escape from the host immune system and for the overall adaptation of the virus. These reverse mutations may be caused by qualitative similarity between mutated amino acids and little effect of amino acid changes on protein structure and function. Amino acids changed in HAs and NAs are stochastic and scattered in the proteins. The total number of mutations in HAs is greater than that of NAs, suggesting that a higher selective pressure is being imposed on HAs.

As shown in Tables 1 and 2, the prevalence of influenza A/H3N2 in Hong Kong is closely related to the amino acid variations of HAs and NAs. High distances do not lead to high influenza prevalence. Random mutations contribute much more to amino acid variations of influenza genes, and we infer that the probable reason leading to influenza pandemic is the amino acid variations in epitopes surpassing a threshold. The preferred antigenic sites for mutation are sites A and B in HAs, and antigenic site B in NAs. Generally, HAs and NAs experienced a stasis-period after turning seasons, which might be an adaptive process between influenza virus and human body.

In HA, there are several amino acid mutation sites. Also, the year when the influenza viruses were prevalent in Hong Kong and in which these amino acids mutated was the same as that in Demark descript by Bragstad (2008). The accordant amino acids mutation sites and year would indicate the influenza viruses in the two areas may originate from the same ancestor through influenza circulation in the world. In NA, there are also many amino acid mutation sites that are the same between influenza viruses prevalent in Hong Kong and Demark, most of which mutated latter in one or two years in Demark than in Hong Kong. The reason may be that the evolution pressure of NA is less than that of HA. Compared with Demark, influenza viruses in Hong Kong lack the amino acid mutations of 92E, 126A, 128B, 173D, 304C in HA and 329C, 332C, 370C, 392A, 393A, 401A in NA, respectively. It is interesting to note that these amino acid sites in NA are adjacent. All of these differences indicate the genetic diversities of influenza viruses. The influenza viruses are still in evolution when they circulate in different regions, which may bring about the genetic differences between influenza viruses in different parts of the worlds.

The average non-synonymous-to-synonymous substitution ratios (dN/dS) for HA and NA genes in this study did not exceed 1 under any of the models that allow for positive selections. Hence, neither HAs nor NAs are directly impacted by positive selection. Instead, they are generally under purifying selection, which lowers the frequency of mutations that impose a negative effect on the fitness of the virus, and only certain sites are affected by adaptive selection. At the threshold 95%, only six amino acid sites, namely residue 50 (antigenic site C), 144 (antigenic site A) and 500 of HA, and 113, 29 and 361 of NA are under positive selection in M3, M2a, M8 together. Position 193 (antigenic site B) in HAs is under positive selection in M3 and M8. Position 500 is different from previous findings by Bragstad et al. (2008, 2009), which suggests that positively selected sites may vary within the dataset applied, method used and the significance level selected for a site to be classified as positively selected. In HAs, three (50C, 144A and 193B) of four positions under positive selection are at antigenic sites, which indicates the important role that epitopes play in pathogenicity of influenza virus. On the other hand, none of the three amino sites under positive selection is at antigenic sites in NA. Combining the less amino acid variation in NA than HA and the higher selection pressure of HA than NA, we would conclude that HAs play a more important role than NAs in the viral entry mechanism and immune recognition.

An important function of N-linked glycosylation of influenza virus proteins is to evade detection by the immune system. The loss or gain of N-glycosylation sites is an important mechanism in antigenic drift, which works by masking or unmasking of the antigenic sites (Sun et al., 2013; Tippmann, 2004).

In this study we predict there are 10-11 potential N-glycosylation sites in HAs and 5-6 potential N-glycosylation sites in NAs (Table 4). Most potential N-glycosylation sites are stable in HAs and NAs. The predicted N-linked glycosylation at position 144 of HAs at antigenic site A has been observed since the 2000/2001 season and half were lost in the following season due to a point mutation (N144D). It is interesting to note that the position 144 is under positive selection. However, this glycolsylation site was further retained in 2002/2003 season, and was conserved in sequent seasons. The predicted N-linked glycosylation at position 93 in NAs has been observed occasionally since the 1998/1999 season and was conserved in the following season. However, this

Table 4. N-Glycosylation sites predicted in the HA1 protein of influenza A isolates.

Gene	Season	Amino acid position[a]
HA	1997/1998	8, 22, 38, 63, ***122,126, 133, 165, 246,*** 285, 483
	1998/1999	8, 22, 38, 63, ***122, 126, 133, 165, 246,*** 285, 483
	1999/2000	8, 22, 38, 63, ***122,126, 133, 165, 246,*** 285, 483,
	2000/2001	8, 22, 38, 63, ***122,126, 133, 144, 165, 246,*** 285, 483
	2001/2002	8, 22, 38, 63, ***122,126, 133, (144)[b], 165, 246,*** 285, 483,
	2002/2003	8, 22, 38, 63, ***122,126, 133, 144, 165, 246,*** 285, 483,
	2003/2004	8, 22, 38, 63, ***122,126, 133, 144, 165, 246,*** 285, 483,
	2004/2005	8, 22, 38, 63, ***122,126, 133, 144, 165, 246,*** 285, 483
	2005/2006	8, 22, 38, 63, ***122,126, 133, 144, 165, 246,*** 285, 483
NA	1997/1998	61, 70, 86, 146, 234, ***329***
	1998/1999	61, 70, 86, (93)[b], 146, 234
	1999/2000	61, 70, 86, 93, 146, 234
	2000/2001	61, 70, 86, 93, 146, 234
	2001/2002	61, 70, 86, 93, 146, 234
	2002/2003	61, 70, 86, 93, 146, 234
	2003/2004	61, 70, 86, 93, 146, 234
	2004/2005	61, 70, 86, 146, 234
	2005/2006	61, 70, 86, 93, 146, 234

[a]Bold numbers represent an antigenic binding site. [b]Nearly half of the strains lost N-glycosylation at this position.

glycosylation site was lost in the 2004/2005 season due to a point mutation (N93D). The site was further retained in sequent seasons. These two positions may not play any major roles in escape from the immune system. There were also some occasional potential N-glycosylation sites in both HAs and NAs, for example, 198 in HAs and 200 in NAs (not shown in Table 4). We think these positions did not contribute significantly to the prevalence of influenza A/H3N2 in the past years.

The influenza outbreak is a complex phenomenon. The genetic make-up of influenza A viruses changes every year. Hence, continuous antigen and genome sequence surveillance of influenza A viruses is still a requirement. In this study, we performed amino acid sequence comparisons among Hong Kong's strains, vaccine strains provided by WHO, and some strains from other regions in the world. We detected significant amino acid substitutions in surface proteins from strains circulating in Hong Kong over a period of ten years (1997-2006). The accumulation of random mutation leads to quantitative variations. We infer that the probable cause leading to influenza pandemic is the amino acid variations in epitopes surpassing a threshold. It is likely that "jumps" in genetic distance rather than constant drift caused the virus evolution (Bragstad et al., 2008). One of the most important examples of influenza pandemic is gene rearrangement (including bird influenza), which resulted in several world-wide influenza pandemic in the 20th century. Random mutations and gene rearrangements in sequence between similar influenza strains may contribute greatly to regional influenza prevalence. The introduction of new influenza virus strains brought forth by gene rearrangements among significantly different strains may be the cause for their world-wide pandemics.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 91130009). The study sponsors had no influence in the study design, data collection, analysis or interpretation, the writing of the paper and the decision to submit this work for publication.

REFERENCES

Bragstad K, Nielsen LP, Fomsgaard A (2008). The evolution of human influenza a viruses from 1999 to 2006. Virol. J. 5(1): 40.

Bridges CB, Thompson WW, Meltzer MI, Reeve GR, Talamonti WJ, Cox NJ, Lilac HA, Hall H, Klimov A, Fukuda K (2000). Effectiveness and cost-benefit of influenza vaccination of healthy working adults. JAMA: J. Am. Med. Associ. 284(13):1655-1663.

Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999). Predicting the evolution of human influenza a. Science 286(5446):1921-1925.

Feng L, Shay DK, Jiang Y, Zhou H, Chen X, Zheng Y, Jiang L, Zhang Q, Lin H, Wang S (2012). Influenza-associated mortality in temperate and subtropical chinese cities, 2003-2008. Bull. WHO 90(4):279-288.

Guindon S, Gascuel O (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.System. Biol. 52(5):696-704.

Gupta R, Jung E, Brunak S (2004). Prediction of N-glycosylation sites in human proteins. Available at http://www.cbs.dtu.dk/services/NetNGlyc/.

Jong Jd, Beyer W, Palache AM, Rimmelzwaan GF, Osterhaus AD (2000). Mismatch between the 1997/1998 influenza vaccine and the major epidemic a (H3N2) virus strain as the cause of an inadequate vaccine-induced antibody response to this strain in the elderly. J. Med. Virol. 61:94-99.

Khor CS, Sam IC, Hooi PS, Quek KF, Chan YF (2012). Epidemiology and seasonality of respiratory viral infections in hospitalized children in Kuala Lumpur, Malaysia: a retrospective study of 27 years. BMC Pediatr. 12(1):32.

Kryazhimskiy S, Bazykin GA, Plotkin J, Dushoff J (2008). Directionality in the evolution of influenza a haemagglutinin. Proc. Roy. Soc. B: Biol. Sci. 275(1650):2455-2464.

Lin Y, Gregory V, Bennett M, Hay A (2004). Recent changes among human influenza viruses. Virus. Res. 103(1):47-52.

Mehle A, Dugan VG, Taubenberger JK, Doudna JA (2012). Reassortment and mutation of the avian influenza virus polymerase PA subunit overcome species barriers. J. Virol 86(3):1750-1757.

Munoz ET, Deem MW (2005). Epitope analysis for influenza vaccine design. Vaccine 23(9):1144-1148.

Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC (2007). Phylogenetic analysis reveals the global migration of seasonal influenza a viruses. PLoS Pathog. 3(9):e131.

Nelson MI, Simonsen L, Viboud C, Miller MA, Taylor J, St George K, Griesemer SB, Ghedin E, Sengamalay NA, Spiro DJ (2006). Stochastic processes are key determinants of short-term evolution in influenza a virus. PLoS Pathog. 2(12):1144-1151.

Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC (2008). The genomic and epidemiological dynamics of human influenza a virus. Nature. 453(7195):615-619.

Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC (2008). The global circulation of seasonal influenza A (H3N2) viruses. Science. 320(5874):340-346.

Shaman J, Goldstein E, Lipsitch M (2011). Absolute humidity and pandemic versus epidemic influenza. Am. J. Epidemiol. 173(2):127-135.

Stöhr K (2002). Influenza-who cares. The Lanc. Infect. Dis. 2(9):517.

Sun X, Jayaraman A, Maniprasad P, Raman R, Houser KV, Pappas C, Zeng H, Sasisekharan R, Katz JM, Tumpey TM (2013). N-linked glycosylation of the hemagglutinin protein influences virulence and antigenicity of the 1918 pandemic and seasonal H1N1 influenza a viruses. J. Virol. 87(15):8756-8766.

Tamura K, Dudley J, Nei M, Kumar S (2007). Molecular evolutionary genetics analysis (MEGA) software version 4.0.Molecul. Biol. Evol. 24(8):1596-1599.

Tang JW, Ngai KL, Lam WY, Chan PK (2008). Seasonality of influenza A (H3N2) virus: A Hong Kong perspective (1997-2006). PLoS One. 3(7):e2768.

Tippmann HF (2004). Analysis for free: comparing programs for sequence analysis. Brief. Bioinfo. 5(1):82-87.

Viboud C, Alonso WJ, Simonsen L (2006). Influenza in tropical regions. PLoS Med. 3(4):468-471.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y (1992). Evolution and ecology of influenza a viruses. Microbio. Rev. 56(1):152-179.

Wilson I, Skehel J, Wiley D (1981). Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature 289(5796):366-373.

Yang Z (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences: CABIOS. 13(5):555-556.

Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155(1):431-449.

Yang Z, Wong WS, Nielsen R (2005). Bayes empirical bayes inference of amino acid sites under positive selection. Molecul. Biol. Evol. 22(4):1107-1118.

Yewdell JW (2011). Viva la revolucion: rethinking influenza a virus antigenic drift. Curr. Opin. Virol. 1(3):177-183.