Virus Genes (2010) 41:329–340
DOI 10.1007/s11262-010-0526-z

Amino acid sequence analysis and identification of mutations under positive selection in hemagglutinin of 2009 influenza A (H1N1) isolates

Xiaofan Ding · Lifang Jiang · Changwen Ke · Zhan Yang · Chunliang Lei · Kaiyuan Cao · Jun Xu · Lin Xu · Xingfen Yang · Yonghui Zhang · Ping Huang · Weijun Huang · Xun Zhu · Zhenjian He · Liping Liu · Jun Li · Jie Yuan · Jueheng Wu · Xiaoping Tang · Mengfeng Li

Received: 4 March 2010 / Accepted: 17 August 2010 / Published online: 31 August 2010

Abstract The 2009 flu pandemic is caused by a new strain of influenza A (H1N1) virus, A/H1N1/09. With its high transmissibility, this novel virus has caused a pandemic and infected over 600,000 people globally. By comparing the hemaglutinin (HA) gene and protein sequences among over 700 A/H1N1/09 isolates, mutations in the receptor-binding sites and antigenic epitope regions were identified. Among these mutations, T220 and E/G239 were found to be strongly positively selected over the course of spreading of the A/H1N1/09 virus worldwide. Interestingly, both sites are located in the highly variable epitope regions of HA1, and residue 239 also plays an important role in the receptor-binding process. Further analyses demonstrated that the percentage of T220 mutants among all isolates increased rapidly during the evolution, and that an E/G239 mutation could decrease the binding affinity of the virus with its cellular receptor. Thus, due to a potential functional importance of residues 220 and 239, mutations at these sites, as well as the significant of positive selection on these sites deserves more attention, while new vaccines and therapeutic drugs are developed against this novel virus.

Keywords H1N1 influenza virus · Hemaglutinin · Mutation · Positive selection

Abbreviations

IAV	Influenza A virus
WHO	World Health Organization

Xiaofan Ding, Lifang Jiang, and Changwen Ke contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s11262-010-0526-z) contains supplementary material, which is available to authorized users.

X. Ding · L. Jiang · K. Cao · L. Xu · X. Zhu · Z. He · L. Liu · J. Li · J. Yuan · J. Wu · M. Li
Key Laboratory of Tropical Disease Control, Ministry of Education, Sun Yat-Sen University, Guangzhou, China

X. Ding · L. Jiang · K. Cao · X. Zhu · Z. He · L. Liu · J. Wu · M. Li (✉)
Department of Microbiology, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan Road II, Guangzhou, Guangdong 510080, China
e-mail: limf@mail.sysu.edu.cn

C. Ke · X. Yang · Y. Zhang · P. Huang
Guangdong Province Center for Disease Control and Prevention, Guangzhou, China

Z. Yang · C. Lei · X. Tang (✉)
The 8th People's Hospital of Guangzhou, 627 Dongfengdong Road, Guangzhou, Guangdong 510060, China
e-mail: xtang@21cn.com

K. Cao
Research Centre for Clinical Laboratory Standard, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

J. Xu
School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou, China

L. Xu
Department of Immunology, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

W. Huang
Department of Medical Genetics and Center for Genome Research, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

J. Li · J. Yuan
Department of Biochemistry, Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, China

HA	Hemagglutinin
NP	Nucleoprotein
NS	Nonstructural protein
NA	Neuraminidase
M	Matrix protein
CFR	Case fatality ratios
RBS	Receptor-binding sites
PDB	Protein data bank
S	Serine
T	Threonine
D	Aspartic acid
G	Glycine
E	Glutamic acid
MOE	Molecular operating environment

Introduction

In March 2009, a novel H1N1 swine-origin influenza A virus (IAV) was first detected in Mexico. With the ability to spread human-to-human, it sparked a growing outbreak of illness globally. The level of influenza pandemic alert was finally raised to Phase 6 by WHO on June 11, 2009. As of November 22, 2009, there have been more than 622,482 laboratory confirmed cases of infection of the pandemic influenza H1N1 2009 virus (A/H1N1/09) and over 7826 deaths reported to WHO [1], with the actual estimation of infections far exceeding the numbers of laboratory confirmed cases and deaths due to incomplete reporting. By May 14, 2010, more than 214 countries and overseas territories or communities have reported laboratory confirmed cases of pandemic influenza H1N1 2009 (A/H1N1/09), including at least 18,036 deaths reported to WHO [2].

IAVs cause epidemics and pandemics through antigenic drift and antigenic shift, respectively [3]. Antigenic drift results from an accumulation of point mutations leading to minor and gradual antigenic changes, while antigenic shift involves major antigenic changes by introduction of new HA and/or NA subtype into human population. Although the current A/H1N1/09 influenza virus remains to be of the H1N1 subtype, it is obvious that the viral changes have reached the level of intra-subtypic antigenic shift that gives rise to a pandemic.

Since 1918, three influenza pandemics, namely, the 1918–1919 H1N1, the 1957 H2N2, and the 1968 H3N2 pandemics, have emerged in human, all of which are thought to have originated from non-human reservoirs [4–6]. The current outbreak of A/H1N1/09 pandemic, as revealed by recent studies, is caused by a novel influenza virus containing a combination of gene segments from different sources. Sequence analyses have demonstrated that the hemagglutinin (HA); polymerases PB1, PB2, and PA; nucleoprotein (NP); and nonstructural protein (NS) gene segments of the A/H1N1/09 virus have the highest homologies with those derived from the swine triple reassortant lineage, which has been circulating in pigs in North America, and the neuraminidase (NA) and matrix protein (M) gene segments are most closely related to those of the Eurasian swine influenza viruses [7]. These analyses suggest that the novel virus might have derived from reassortment events occurring between the North American and the Eurasian lineages.

It is of note that the A/H1N1/09 pandemic initially has exhibited a relatively low mortality, with case fatality ratios (CFR) ranging from 0.3 to 1.5%, indicating that the currently widespread virus probably have not mutated to support a most virulent phenotype. Whereas, a relatively high transmissibility, found by the clinical surveillance of the pandemic influenza H1N1 2009 virus [8–12], suggests that the virus is able to escape protective immunity easily. Furthermore, it has been recently reported the A/H1N1/09 pandemic exhibits an unusual pattern of age-related morbidity and mortality, as it disproportionately affects children and young adults (ages 4–25), compared with seasonal influenza viruses [9–12]. The incidence of severe disease decreases with age, with the lowest occurrence in the population 65 years and older, suggesting preexisting immunity to the 2009 pandemic virus in people born before 1957 [13]. It is hence of great interest to identify the gene sites/mutations to which the highly transmissible feature of the novel virus is attributable.

The HA glycoprotein on the surface of IAV particles function as the receptor-binding ligand, mediating entry, and internalization of the virus into host cells and subsequent membrane-fusion events in the infected cells. The mature HA is a homotrimer of $\sim$220 kD containing several glycosylation sites. Each HA molecule is synthesised as a single polypeptide precursor (HA0) and subsequently cleaved into HA1 and HA2 subunits by an endoprotease that targets a specific cleavage site in HA0 [14]. This event is a prerequisite condition for successful infection, and the generated HA1 polypeptide bears the receptor-binding sites (RBS). Furthermore, major epitopes specific for protective immune response are also located in HA of IAV, as well as of influenza vaccines, as identified by previous studies as epitopes A through E [15]. Recently, experimental data have shown that this A/H1N1/09 influenza virus can bind to both 2,3- and 2,6-linked sialic acid receptors and replicate in the lower respiratory tracts of infected mammals [16]. Interestingly, Krause et al. reported that the naturally occurring human monoclonal antibodies neutralize both 1918 and 2009 pandemic influenza A (H1N1) viruses [17]. These features raised the concern that this new virus may possess virulence characteristics similar to those of the

highly pathogenic 1918 pandemic influenza viruses. Lately, Wei et al. defined the structural basis for cross-neutralization and protection between two distant pandemic influenza viruses of the 1918 and 2009 pandemics, suggesting that specific N-glycans in HA may play a key role in modulating immune recognition and influencing on viral evolution [18]. Accordingly, mutations in HA therefore can contribute to changes in virulence and transmissibility of influenza viruses.

Influenza virus is subject to genetic mutation, mainly due to the lack of proof-reading activity of its polymerase. Mutations in influenza viral genes accumulate over time and are under selection pressure during epidemics or pandemics. Thus, frequencies of mutations detected for a specific IAV strain may indicate a positive or negative selection, as well as the underlying biological or epidemiological factors, under which the virus evolves in the process of emergence and spreading among the host population. For example, previous analysis on the evolution of the M gene have found that one and ten sites in M1 and M2 regions, respectively, are under positive selection in human, and that the M1 and M2 regions are evolving independently under different selective pressure in different hosts. The study also identified potentially important sites that may be related to host tropism and immune responses [19].

In this study, we analyzed the sequences of A/H1N1/09 IAV with a focus on its HA protein using various analysis tools. Our results revealed the presence of mutations possibly relevant to the pathogenesis and evolution of A/H1N1/09 IAV, and therefore might be useful for further investigation of the pathogenic and immunogenic properties of this rapidly spreading virus and future design of more effective vaccines.

Methods

Point mutation analysis

All the non-redundant HA sequences of the A/H1N1/09 virus in the GenBank (deposited as of April 21, 2010), with a total number of 704 (full-length only, collapse identical sequences, and including the information of exact collection time), were downloaded and loaded into the ClustalW and the BioEdit programs for multiple alignment analysis, which led to the generation of a consensus sequence. Variations at each amino acid position along the HA protein were identified among 704 downloaded sequences.

The HA genes of four IAVs (H1N1) that were circulating in humans and recommended by WHO as an influenza vaccine component between 2001 and 2009 were used to create a consensus sequence for human seasonal H1N1 influenza viruses.

Meanwhile, the protein sequences were compared between the A/H1N1/09 and human seasonal IAV HA consensus sequences using multi-sequence alignment analysis. Individual A/H1N1/09 HA sequences, including A/California/04/2009 (H1N1), A/Beijing/01/2009 (H1N1), A/Sichun/1/2009 (H1N1) and A/Guangdong/1/2009 (H1N1), and seasonal IAV HA sequences, including A/New Caledonia/20/1999 (H1N1), A/Solomon Islands/03/2006 (H1N1), A/Brisbane/59/2007 (H1N1), and A/Washington/10/2008 (H1N1), were also compared with one another and with both consensus HA sequences, to identify variations along the HA protein. The glycosylation sites of both consensus HA sequences were analyzed by NetNGlyc 1.0 server [20].

Structural modeling

3-D structure of A/H1N1/09 HA protein was modeled using the Modeller program to modify previously known H1N1 HA protein, whose crystal structures have been determined, with altered amino acids identified in the HA of novel A/H1N1/09 virus. Briefly, the A/H1N1/09 HA consensus sequence was first used to search the PDB [21] using BLAST to find deposited IAV HA proteins with the highest similarity to it. This procedure led to the identification of three HA proteins in the PDB (PDB ID: 1RUY, 1RVT, and 1RVO), which were then downloaded as templates for further modeling procedures. Subsequently, a homology model was created between A/H1N1/09 HA and the selected templates (1RUY, 1RVT, and 1RVO), and the sequence of A/H1N1/09 HA protein was modeled 50 times onto the selected templates. The best model resulting from the above modeling procedure was obtained and visualized in Jmol [22].

Site-by-site analysis

All the 704 coding sequences of the A/H1N1/09 virus available in GenBank (deposited as of April 21, 2010) were taken for further site-by-site positive selection analysis using the HyPhy program [23] under the "MG94×HKY85×3_4×2_Rates" model (4 rate categories assigned). The ratios of non-synonymous (dN) and synonymous (dS) substitutions were calculated for each site in all codons. All the calculated dN/dS values were then further tested with the empirical Bayes method [24, 25], and when the Bayes factor (the ratio of posterior odds of an event and its prior odds) was significantly greater than 1, it was considered that the

hypothesis of dN/dS > 1 or dN/dS < 1 was true. Sites where dN/dS > 1 or dN/dS < 1 were considered positively selected or negatively selected, respectively.

Analysis for isolation time and geographic regions distribution of HA S220 and T220

For further analysis of the distribution of different residues on position 220 of the HA protein (pandemic 2009 H1 numbering), all the 704 non-redundant HA protein sequences of the A/H1N1/09 influenza virus in the NCBI influenza virus sequence database (deposited as of April 21, 2010) were downloaded and placed into 12 groups according to their isolation times. Multi-sequence alignment was performed. The frequency of each allele was calculated as function of isolation time to evaluate whether the frequency of a specific amino acid present at position 220 changed over time. The trends of such changes were then tested by employing the Kendall test and linear model in the R [26] program, and $P < 0.05$ was chosen to indicate that a trend was significant. The frequency of each allele was also calculated as function of geographic regions of isolation following grouping non-redundant A/H1N1/09 HA protein sequences according to the regions where they were isolated (Asia, Europe, North America, South America, and Oceania) and tested using the χ^2 test.

Binding energy analysis

The MOE (molecular operating environment) program [27] was used to calculate the binding energy between A/H1N1/09 HA and its cellular receptor based on a known binding structure (PDB ID: 3GBN, X-ray diffraction) composed of the A/South Carolina/1/1918 HA and a receptor analog. First, we partially minimized the complex by relaxing the ligand and the side chains within 10 nm from the ligand, while keeping all other atoms fixed.

Following calculation of energies, factor analysis (FA) and multiple regression analysis (MRA) were employed to generate an LRE-like equation [28, 29]:

$$\Delta G^{\mathrm{b}}(\mathrm{FEB}) = \omega_1 \Delta G^{\mathrm{b}}_{\mathrm{vdW}} + \omega_2 \Delta G^{\mathrm{b}}_{\mathrm{ele}} + \omega_3 \Delta G^{\mathrm{b}}_{\mathrm{solv}} + \omega_4 \Delta G^{\mathrm{b}}_{\mathrm{n}}$$

$$\Delta G^{\mathrm{b}} = \Delta G_{\mathrm{complex}} + \Delta G_{\mathrm{protein}} + \Delta G_{\mathrm{ligand}}$$

In this equation, ΔG (FEB) stands for the free energy of binding, ΔG_{vdW}, ΔG_{ele}, ΔG_{solv}, and ΔG_{n} stand for the van der Walls contribution, the electrostatic contribution, the polar solvation contribution, and the nonpolar solvation contribution to the binding process, respectively, where $w1$, $w2$, $w3$, and $w4$ are weight factors, and ΔG^{b} represents binding energy (i.e., energy difference between ligand/receptor complex and free protein and ligand).

Results

Point mutation analysis

To specify the differences between the HA molecules of the A/H1N1/09 virus and recently circulating seasonal IAVs, we generated a consensus sequence from 704 A/H1N1/09 HA amino acid sequences deposited in the GenBank (as of April 21, 2010) and a consensus sequence from the HA proteins of 4 seasonal H1N1 IAV strains that had been recommended by WHO for production of influenza vaccines, and compared these two groups of HA sequences by aligning both consensus sequences and several individual HA sequences from each group (Fig. 1).

As shown in Fig. 1 and Table 1, while the overall difference between the consensus sequences of seasonal H1N1 and A/H1N1/09 HA proteins was as high as 19.61% (111/566), the cleavage sites of HA, at which the HA0 protein is recognized by specific protease(s) and enzymatically cleaved into HA1 and HA2, thereby becoming activated in mediating the entry of IAVs into host cells, remain identical (PSIQSR↓GLFGAI) among the strains analyzed. Meanwhile, the Asp204, Asp239, Gln240, and Gly242 residues at the RBS that are responsible for the viral attachment to the host cell receptor, a critical step for viral entry, were also identical between the two consensus sequences. It is of note that the four positions (i.e., 204, 239, 240, and 242) have been found previously to be involved in the specific binding capacity of HA to the host cell receptor [30–32]. Whereas, amino acid variations were found at a number of positions at the HA RBS among different seasonal H1N1 IAVs and the A/H1N1/09 viruses, including positions 150, 152, 206, 207, 210, 211, and 212. Whether variations at these sites could impact the infectivity of an influenza virus to a specific host needs to be further investigated biologically.

To address the differences in the antigenic properties between A/H1N1/09 and human seasonal IAVs, previously proposed antigenic epitope regions were comparatively analyzed. In this context, two groups of epitope regions, namely, the highly conserved regions and the highly variable regions [33], were analyzed, respectively. As shown in Fig. 1 and Table 2, four highly conserved epitope regions (1–4), located in HA2, involving residues 345–354, 359–376, 394–411, and 436–453, were all identical between A/H1N1/09 isolates and seasonal H1N1 viruses. In contrast, the highly variable regions, which lie in the HA1 globular head and include sites (residues 86–91), Sa (residues 141–142, 170–174, and 176–181), Sb (residues 201–212), Ca1 (residues 183–187, 220–222, and 252–254) and Ca2 (residues 154–159, and 238–239) [34], showed dramatic changes in A/H1N1/09 isolates when compared with those in seasonal IAVs (H1N1). Such changes might

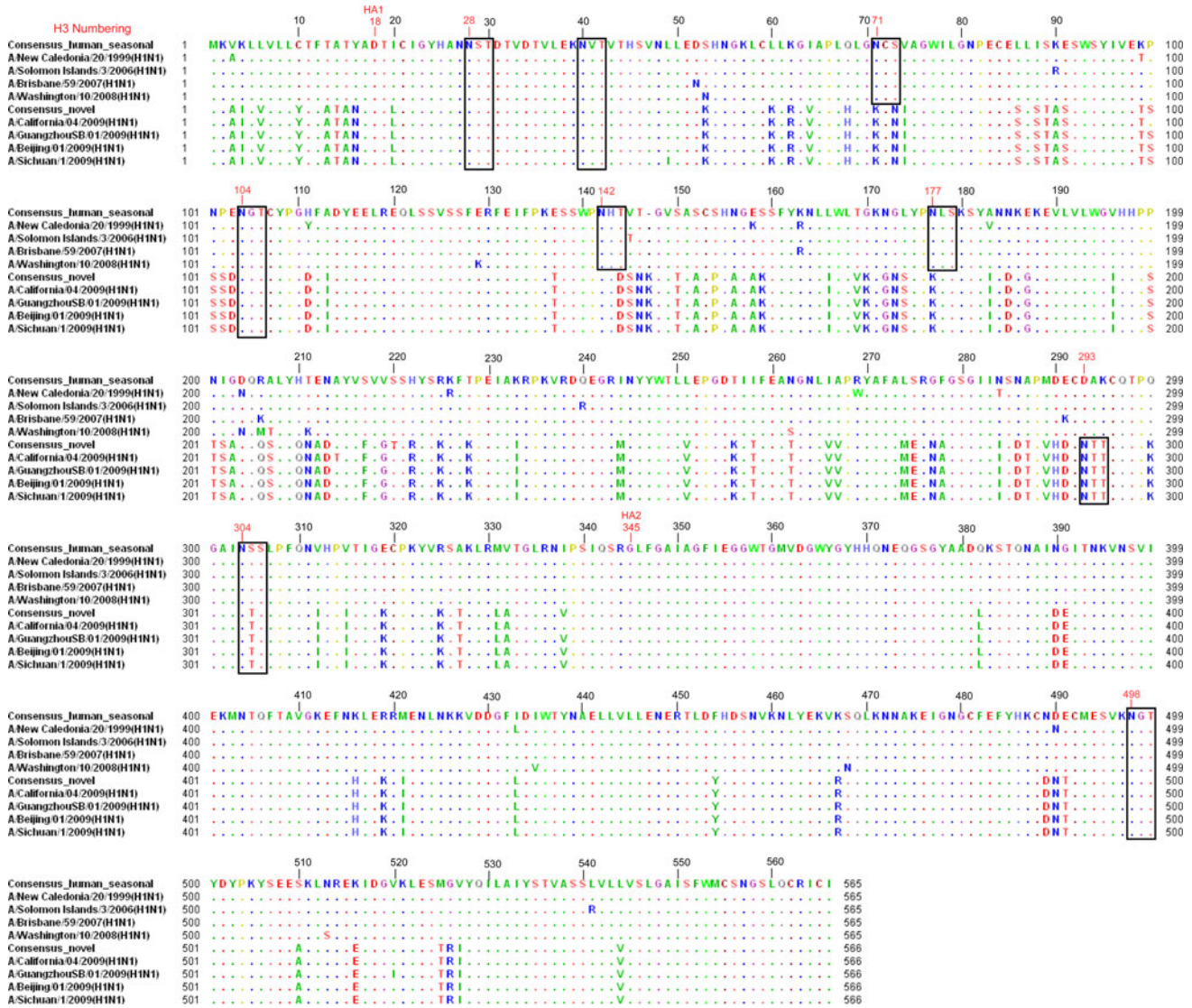

Fig. 1 Alignment of HA amino acid sequences of human seasonal IAVs and their consensus sequence in comparison with the pandemic A/H1N1/09 IAVs and their consensus sequence. Residues different from the consensus HA amino acid sequence (*top line*) of human seasonal H1N1 IAVs are shown, and *dots* stand for identical residues. The sequences were numbered according to pandemic 2009 H1 numbering. *Boxed* residues indicate the Asn-X-Thr/Ser motifs of glycosylation sites

Table 1 Comparison between the new A/H1N1/09 isolates and seasonal H1N1 IAVs at HA0 cleavage site and RBS

Functional sites	Virus	Residues		
The cleavage site	Human seasonal	PSIQSR↓GLFGAI		
	A/H1N1/09	PSIQSR↓GLFGAI		
		Difference: 0		
RBS (in HA1)		Residue positions (pandemic 2009 H1 numbering)		
		149–152	204–212	235–242
	Human seasonal	**V**S**A**S	DQ**RA**LY**HTE**	PKVRDQEG
	A/H1N1/09	**V**T**AA**	DQ**QS**LY**QNA**	PKVRDQEG
		Difference: 2/4	Difference: 5/9	Difference: 0

Table 2 Comparison between A/H1N1/09 isolates and seasonal H1N1 IAVs at antigenic epitope regions

	Epitope region	Virus	Residues and positions[b]		
Highly conserved regions (in HA2)	1		345–354 in HA2		
		Human seasonal	GLFGAIAGFI		
		A/H1N1/09	GLFGAIAGFI		
			Difference: 0		
	2		359–376 in HA2		
		Human seasonal	TGMVDGWYGYHHQNEQGS		
		A/H1N1/09	TGMVDGWYGYHHQNEQGS		
			Difference: 0		
	3		394–411 in HA2		
		Human seasonal	NKVNSVIEKMNTQFTAVG		
		A/H1N1/09	NKVNSVIEKMNTQFTAVG		
			Difference: 0		
	4		436–453 in HA2		
		Human seasonal	WTYNAELLVLLENERTLD		
		A/H1N1/09	WTYNAELLVLLENERTLD		
			Difference: 0		
Highly variable regions (in HA1)[a]	Cb		86–91 in HA1		
		Human seasonal	L**ISK(R)E**S		
		A/H1N1	L**STAS**S		
			Difference: 4/6		
	Sa		141–142 in HA1	170–174 in HA1	176–181 in HA1
		Human seasonal	PN	**GKNGL**	P**N**LSKS
		A/H1N1/09	PN	**KKGNS**	P**K**LSKS
			Difference: 0	Difference: 4/5	Difference: 1/6
	Sb		201–212 in HA1		
		Human seasonal	**NIG**D(N)Q**R(K/M)A(T)**LY**HT(K)E**		
		A/H1N1/09	**TSA**DQ**QS**LY**QNA**		
			Difference: 8/12		
	Ca1		183–187 in HA1	220–222 in HA1	252–254 in HA1
		Human seasonal	**A(V)**N**NKE**	SS**H**	EPG
		A/H1N1/09	**INDKG**	**TSR**	EPG
			Difference: 3/5	Difference: 1/3	Difference: 0
	Ca2		154–159 in HA1	238–239 in HA1	
		Human seasonal	S**HNGE(K)S**	RD	
		A/H1N1/09	**PHAGAK**	RD	
			Difference: 4/6	Difference: 0	

[a] Amino acid sequences for all highly variable epitope regions listed in this table are given using the consensus sequences generated for A/H1N1/09 isolates and seasonal H1N1 IAVs, respectively, as presented in Fig. 1. Variations in individual sequences as compared with the consensus sequence are given in *parenthesis*

[b] Numbering according to pandemic 2009 H1 numbering

provide a molecular explanation for the observed lack of cross-protection from previous infection or vaccination of seasonal IAVs against the novel A/H1N1/09 virus.

Since point mutations could cause emergence or loss of Asn-X-Thr/Ser motifs and thereby, attachment or loss of N-glycans, respectively, leading to alteration of the antigenicity and receptor specificity of HA, we analyzed the glycosylation sites on H1 HAs by further examining both consensus sequences of the seasonal H1N1 and the novel pandemic H1N1, which revealed five identical glycosylation sites (28, 40, 104, 304, 498, pandemic 2009 H1 numbering) between the two consensus sequences, as shown in Fig. 1. In marked contrast, the A/H1N1/09 stains contained amino acid mutations predicted to lose three

glycosylation sites at position 71, 142, and 177. It is of note that the two highly conserved glycosylation sites (i.e., 142 and 177) have been found previously to be involved in the antigenic properties of H1N1 IAVs, which was within or around the Sa site. Meanwhile, we found that the A/H1N1/09 stains carried amino acid mutations predicted to acquire one potential glycosylation site at position 293, raising the issues whether such glycosylation does occur at the site and alters the antigenicity and receptor specificity of 2009 influenza A (H1N1) virus.

Point-to-point analysis of mutations in the HA protein was performed among 704 isolates of the novel A/H1N1/09 virus. As shown in Tables 3 and 4, 13 out of 566 positions in the entire HA molecule displayed variations among 704 non-redundant A/H1N1/09 HA sequences (variation frequency was higher than 2), including 5 positions at RBS, and 11 positions in the highly variable epitopes, and all other variations were shown in the supplement data Table S1. Specifically, the amino acid change at position 342 (Gln342Leu) was at the cleavage site of one A/H1N1/09 isolate [A/Guangdong/03/2009 (H1N1)]. Whether such a change would influence the interaction of HA0 with proteolytic enzyme(s) and consequently, the cleavage activity, remains unclear. It requires further investigation to clarify whether mutation Asp239Glu in two isolates [A/Paris/2591/2009 (H1N1) and A/New Jersey/01/2009 (H1N1)] was relevant to host adaptation, since Glu239 had been previously found to be associated with acquisition of SA α-2,6Gal binding specificity [30, 35]. Most noteworthily is the high frequency (72.02%) of Ser220Thr mutation in the Ca1 epitope, strongly indicating existence of a positive selection or, at a lesser likelihood (due to lack of other high-frequency mutated positions in the HA), more than one origin of the new A/H1N1/09 viruses.

Structural modeling

To better characterize the variations in A/H1N1/09 HA, we sought to map the altered amino acids to a predicted three-dimensional (3-D) HA structure. We first BLAST searched the PDB database for deposited HA molecules with the highest sequence homologies to A/H1N1/09 HA. Three sequences, together with their previously X-ray determined 3-D structures, of similarities higher than 80% were selected, which were 1RUY [HA of A/swine/Iowa/15/30 (H1N1)], 1RVT [HA of A/swine/Iowa/15/30 (H1N1) complexed with receptor analog LSTC] and 1RV0 [HA of A/swine/Iowa/15/30 (H1N1) complexed with receptor analog LSTA] [32]. Using the selected molecules as templates, onto which the new A/H1N1/09 HA consensus sequence was modeled for 50 times, a predicted 3-D structure of the A/H1N1/09 HA was obtained, visualized with Jmol, and demonstrated in Fig. 2.

Table 3 Amino acid variations in the HA protein sequence among A/H1N1/09 isolates

	Positions[a]	Primary residue	Frequency (%)	Variation (s)	Variation frequency (%)
HA1	36	V	97.73	I	2.13
				L	0.14
	49	L	96.45	I	3.41
				X	0.14
	103	D	97.17	E	0.14
				G	2.41
				M	0.28
	114	D	95.74	N	4.12
				X	0.14
	145	S	95.32	M	4.40
				X	0.28
	220	**S**	**26.99**	**A**	**0.14**
				T	**72.02**
				X	**0.85**
	222	R	96.17	G	0.14
				K	3.13
				S	0.28
				T	0.14
				X	0.14
	239	**D**	**88.78**	**E**	**4.69**
				G	**3.55**
				N	**0.99**
				X	**1.99**
	310	Q	95.31	H	4.26
				X	0.43
	314	P	97.30	S	2.70
	338	V	94.88	I	4.69
				S	0.43
HA2	391	E	88.78	G	0.71
				K	10.51
	428	V	97.30	I	2.56
				X	0.14

[a] Positions listed according to pandemic 2009 H1 numbering

[b] "X" stands for any amino acid

As shown in Fig. 2, the structure of A/H1N1/09 HA monomer was similar to those of other published HAs, as expected, with a globular head containing the RBS, a cleavage site, a trans-membrane domain, and an α-helical stalk. Variations found in various isolates of the A/H1N1/09 HA, as well as differences between the HA molecules of human seasonal IAVs and the new A/H1N1/09 virus, and the glycosylation attachment sites were mapped on the 3-D structure as shown in distinct colors in Fig. 2. These variations and different residues were mainly located in the HA1 fragment, most of which lied in antigenic epitopes.

Table 4 Amino acid variations in functional regions of A/H1N1/09 isolates

Functional region	Residue positions[a]	Variations[a]
The cleavage site	Between HA1 and HA2	None
RBS (in HA1)	149–152	151
	204–212	204
	235–242	238, 239, 240
Highly conserved epitopes (in HA2)	345–354	None
	359–376	None
	394–411	None
	436–453	None
Highly variable epitopes (in HA1)		
Cb	86–91	90
Sa	141–142	None
	170–174	172
	176–181	179
Sb	201–212	202, 203, 204
Ca1	183–187	None
	220–222	220, 222
	252–254	252
Ca2	154–159	None
	238–239	238, 239

[a] Positions listed according to pandemic 2009 H1 numbering

Site-by-site analysis

To address the significance of the mutations identified in A/H1N1/09 HA, we conducted positive selection analysis on a site-by-site basis, as positive natural selection could drive the increase in prevalence of advantageous traits, which may lead to a new pandemic virus. In this study, the ratio between non-synonymous (dN) and synonymous (dS) substitutions were used to indicate selective pressure on each codon. According to Yang et al. [24], when the dN/dS value on a certain codon is greater than 1 and tested to be significant by the Bayes test, the site is considered to be under positive selection; and in contrast, the site would be recognized as being under negative selection in the case if the dN/dS value is significantly smaller than 1. Using the HyPhy program [23], our positive selection analysis revealed that 34 codon sites showed dN/dS > 1 with statistical significance (Bayes factor > 1) (Fig. 3). To minimize possible errors such as those caused by biased sampling, we chose to use a very conservative strategy in the identification of the site mutations under positive selection by selecting only those codon sites with the highest Bayes factors, which accounted for the leading 5% of all codon sites with a Bayes factor > 1. Consequently, these procedures led to identification of two sites, namely, positions 220 and 239 (pandemic 2009 H1 numbering, equivalent to 206 and 225, respectively, according to the H3 numbering) in the HA1 protein. Notably, the residue 220 lied in the Ca1 epitope region in the head of HA1, and residue 239 was found to be included in the RBS of HA1. It is of interest that previous studies have suggested that both regions are relevant to the determination of the severity and transmissibility of an IAV [35], raising the question whether the mutations at these sites could favor the prevalence of the novel virus.

Variation at position 220

Particularly noteworthy was the relatively high frequencies of S220 and T220 present in the HA of A/H1N1/09

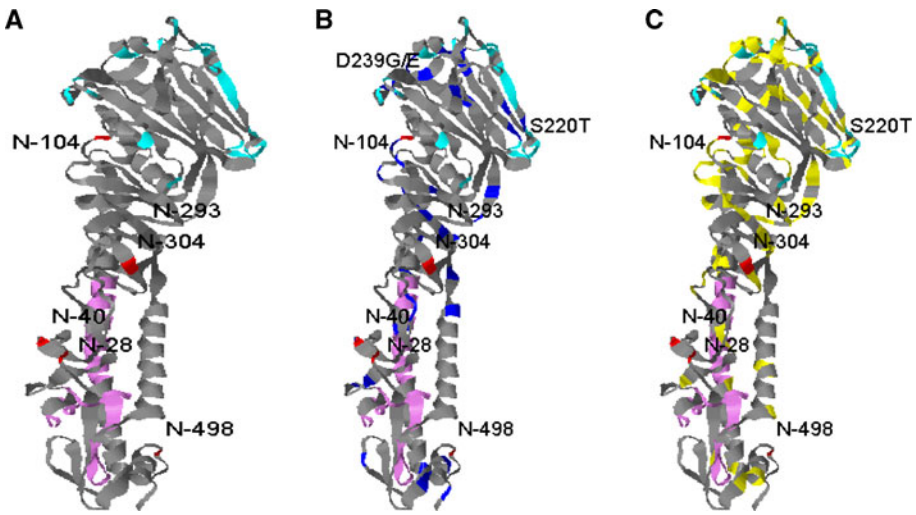

Fig. 2 Mapping of mutations onto the 3-D structure of HA monomer. Side views of the HA monomers are shown. Each has the highly conserved epitope regions in HA2 colored *violet* and the highly variable epitope regions in HA1 colored *cyan*. **a** the reference HA monomer, and **b**, **c** the HA monomers with mutations mapped. In **b**, regions colored *blue* stand for mutations detected among the A/H1N1/09 viruses; in **c**, regions colored *yellow* demonstrate differences between the human seasonal influenza virus and this novel virus, and the amino acid variations at residues 220 and 239 and the glycosylation sites were numbered (Color figure online)

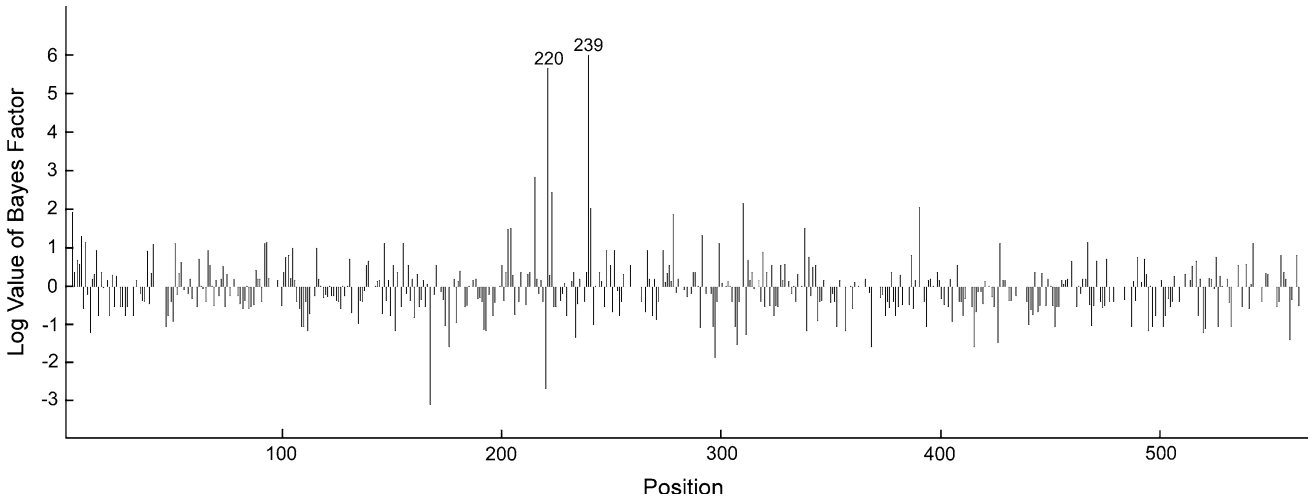

Fig. 3 Site-by-site mutation analysis. The selection profile of HA genes is shown. The abscissa indicates the codon position, and the ordinate stands for the log value of the Bayes factor for each position, which indicates differentials of selective pressures on sites

isolates, prompting us to ask whether they were a result of selection during the course of evolution. To address this issue, we further analyzed the frequency distribution of S220 and T220 in the HA among isolates of the novel A/H1N1/09 virus as a function of isolation time and geographic regions. When 704 HA sequences derived from A/H1N1/09 viruses isolated between the period of March 30, 2009 and April 21, 2010, were grouped according to their isolation time, a total of 507 isolates (72%) with T220 were identified, and the percentage of S present at position 220 was found to gradually decline, while the frequency of T220 increased over time (Fig. 4). To test the significance of the descending trend of S220, Kendall test and a linear model were used, and results revealed a Kendall test

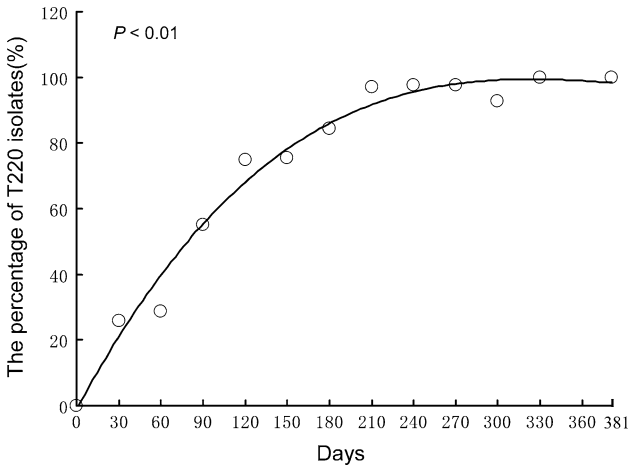

Fig. 4 The frequencies of threonine present at position 220 of HA changed over time. The abscissa indicates the virus collection time since March 30, 2009, while the ordinate indicates the percentage of T220 in all isolates. The trends of such changes were then tested by employing the Kendall test and linear model in the R program, and $P < 0.05$ was chosen to indicate statistical significance

P value of 2.503×10^{-5} and a descending rate of 0.06655 with a P value of 9.84×10^{-5} in the linear model analysis. Both tests confirmed the significant descending trend of the residue S at position 220, with T220 gradually becoming prevalent in the infected population. In contrast, no significant difference in the frequency distribution of S220 versus T220 among HA molecules derived from viruses isolated from Asia, Europe, North America, Oceania, and South America, with a χ^2 test P value of 0.2801, indicating that the changes in the frequencies of S220 versus T220 found in the above study probably were not geographic region-specific and might have been occurring worldwide.

We next sought to investigate whether S220 and T220 were present in the HA molecules of previously isolated swine IAVs and seasonal human H1N1 viruses. We downloaded 272 non-redundant HA protein sequences of North American classical swine flu virus (H1 subtype) from the NCBI influenza virus sequence database and analyzed the frequency of amino acids present at position 220. Interestingly, 251 out of the 272 sequences were found to have S220, as opposed to that only 20 isolates carried T at the position 220. Moreover, all of the 20 strains with T220 were H1N1 subtype swine influenza A viruses, among which 14 were isolated from Tennessee in 1976, 1977, and 1978, and the remaining six strains were isolated sporadically from several other US states. On the other hand, the frequency distribution in 1767 HA sequences derived from human-infected H1 subtype IAVs isolated before the 2009 pandemic showed an S220:T220 ratio of 1760:7.

Variation at position 239

Since our point mutation analysis found a notable variability of position 239 in HA, we analyzed the frequencies

of different residues at the position in HA of 704 A/H1N1/09 isolates. Our analysis found that aspartic acid (D) was present at a frequency of 88.78% (623/704), glycine (G) at 3.55% (25/704), and glutamic acid (E) at 4.69% (33/704), with the remaining residues undetermined. Possible impact of mutations at this site on receptor binding was examined by calculating the minimal energy using the MOE (Molecular Operating Environment) program, and the result showed that the binding energy between receptor analog and wild type HA (D239) was −29.506 kcal/mol, while for mutants G239 and E239, it was −17.027 and −16.445 kcal/mol, respectively, indicating possibly weakened receptor-binding capacities of the mutants.

Discussion

Our current study analyzed the site variations in the HA protein sequences among the novel A/H1N1/09 viral isolates, as well as those between human seasonal influenza viruses and A/H1N1/09. 3-D structure of the HA was also modeled, and identified point mutations were analyzed to predict their potential significance in molecular evolution and function of the pandemic virus.

In this study, the point mutation analysis showed that the A/H1N1/09 strains carry amino acid mutations that might lead to acquisition of one glycosylation site (position 293) and loss of three glycosylation sites (position 71, 142, and 177). It has been reported that variation in glycosylation is used by influenza viruses to interfere with surveillance by the host immune system. Acquisition of a glycosylation site masks the protein surface from antibody recognition because the glycans themselves are host-derived, and hence considered as "self" by the human immune system [36, 37] Meanwhile, upon the addition of glycosylation to this region has been thought to slow down yearly antigenic drift, presumably because glycosylation shielded this region from the antigenic pressure of antibodies [18, 36]. Recently, Wei et al. and Xu et al. have evaluated the cross-neutralization of pandemic 1918 and 2009 H1N1 influenza viruses and found that they were both resistant to antisera directed to a relatively recent seasonal influenza virus of the same subtype [18, 38]. Interestingly, pandemic 2009 H1N1 influenza virus (A/California/04/2009), like A/South Carolina/1/1918 virus, does not have any glycosylation in or around the Sa site in HA and hence, the epitope is exposed for antibody recognition. They suggest that these N-glycans of the RBD and antigenic epitope regions may play roles in evading the human immune response and viral evolution in humans [18]. The probability of glycosylation of the predicted asparagine residues 293 above mentioned in our current study, and the role of the acquisition (position 293) or loss of glycosylation (position 71) in modulating immune recognition and its influence on viral evolution, is under investigation in our laboratory.

By mapping the variations in the HA protein sequence, we identified that the distinction between the HA antigenic specificities of the human seasonal influenza virus and the novel A/H1N1/09 virus mainly lie in the HA1 epitope regions, which might explain the lack of effective cross-protection by previous seasonal IAV vaccines or infections and the highly transmissible property of the A/H1N1/09 virus.

A key notion derived from this study is the possibility that two HA1 sites in the novel A/H1N1/09 virus, namely, residues 220 and 239, might be positively selected during its evolution. Interestingly, residue 239 lies in a region possibly involved in both receptor binding and determination of antigenic specificity [30–32, 34]. It has been previously reported that when residue 239 changes from Gly to Glu, the virus becomes more adaptive to human host [30]. Recently, Chen et al. reported the D225G (H3 numbering, D239G according to 2009 pandemic H1 numbering) substitution in 7 (12.5%) of 57 patients with severe disease, and in 0 (0%) of 60 patients with mild disease, and the D225E mutation was identified in one patient with severe disease, by direct analysis of polymorphisms in 126 amino acids spanning the receptor-binding site in the hemagglutinin of pandemic H1N1 2009 virus from 117 clinical specimens in Hong Kong [39]. Our current study revealed that the vast majority of pandemic A/H1N1/09 isolates carried D239 (88.78%) and yet a small fraction of the A/H1N1/09 isolates had glutamic acid (E, 4.69%) and glycine (G, 3.55%) at this position. Free binding energy analysis suggested that a D239E/G mutation tended to decrease the affinity of the H1 subtype IAV to the sialic acid receptor. It remains to be determined, whether the identified positive selection of the D239E/G mutation is functionally significant during the spreading of the A/H1N1/09 virus.

Residue 220 lies in the Ca1 epitope region. In all A/H1N1/09 isolates obtained as of August 2009, T220 was present at a frequency of 72.02%, whereas the frequency of S220 was 26.99%. It is of note that both T220 and S220 were present in H1 subtype classical swine IAVs, from which the HA of the pandemic A/H1N1/09 was originated, and that the frequency of T220 was far lower than that of S220 in these swine IAVs. Our analysis on the dynamic change of T220 versus S220 demonstrated an ascending trend of the percentage of T220 and a descending curve of S220 during the course of the pandemic spreading, suggesting that T220 might have been positively selected over S220. Speculatively, T220 was then fixed through natural or other selection at the peak of this pandemic and favored the transmissibility of the new virus. Interestingly, compared with serine, amino acid threonine carries an extra

methyl group and therefore displays a different polarity than serine. Whether such a difference in the polarity could contribute to altering recognition and binding capacity between the antigenic epitope and specific antibodies and thereby should be taken into consideration when developing more effective vaccines and therapeutic drugs remains to be further investigated.

Acknowledgments This study was supported by the National Natural Science Foundation of China-Guangdong Province joint grant (U0632002), a grant from the State Major Infectious Disease Research Program (China Central Government, 2009ZX10004-213), a Key Science and Technique Research Project of Guangdong Province (2009A020101006, 2009B020600001), National High-Tech R&D Program (863 Program) (Ministry of Science and Technology, China, 2006AA02A223, 2007AA09Z448, 2007AA09Z431), National Science and Technique Research Program for public welfare applications (201005022), and a Key Project of Science & Technology Planning of Guangdong Province (2007A03260001).

References

1. WHO, http://www.who.int/csr/don/2009_11_27a/en/index.html. Accessed 27 Nov 2009
2. WHO, http://www.who.int/csr/don/2010_05_14/en/index.html. Accessed 14 May 2009
3. R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, Y. Kawaoka, Microbiol. Rev. **56**, 152–179 (1992)
4. J.K. Taubenberger, A.H. Reid, R.M. Lourens, R. Wang, G. Jin, T.G. Fanning, Nature **437**, 889–893 (2005)
5. S.E. Lindstrom, N.J. Cox, A. Klimov, Virology **328**, 101–119 (2004)
6. C. Scholtissek, W. Rohde, V. Von Hoyningen, R. Rott, Virology **87**, 13–20 (1978)
7. R.J. Garten, C.T. Davis, C.A. Russell, B. Shu, S. Lindstrom, A. Balish, W.M. Sessions, X. Xu, E. Skepner, V. Deyde, M. Okomo-Adhiambo, L. Gubareva, J. Barnes, C.B. Smith, S.L. Emery, M.J. Hillman, P. Rivailler, J. Smagala, M. de Graaf, D.F. Burke, R.A. Fouchier, C. Pappas, C.M. Alpuche-Aranda, H. Lopez-Gatell, H. Olivera, I. Lopez, C.A. Myers, D. Faix, P.J. Blair, C. Yu, K.M. Keene, P.D. Dotson, D. Boxrud Jr., A.R. Sambol, S.H. Abid, K. St George, T. Bannerman, A.L. Moore, D.J. Stringer, P. Blevins, G.J. Demmler-Harrison, M. Ginsberg, P. Kriner, S. Waterman, S. Smole, H.F. Guevara, E.A. Belongia, P.A. Clark, S.T. Beatrice, R. Donis, J. Katz, L. Finelli, C.B. Bridges, M. Shaw, D.B. Jernigan, T.M. Uyeki, D.J. Smith, A.I. Klimov, N.J. Cox, Science **325**, 197–201 (2009)
8. C. Fraser, C.A. Donnelly, S. Cauchemez, W.P. Hanage, M.D. Van Kerkhove, T.D. Hollingsworth, J. Griffin, R.F. Baggaley, H.E. Jenkins, E.J. Lyons, T. Jombart, W.R. Hinsley, N.C. Grassly, F. Balloux, A.C. Ghani, N.M. Ferguson, A. Rambaut, O.G. Pybus, H. Lopez-Gatell, C.M. Alpuche-Aranda, I.B. Chapela, E.P. Zavala, D.M. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, Science **324**, 1557–1561 (2009)
9. G. Chowell, S.M. Bertozzi, M.A. Colchero, H. Lopez-Gatell, C. Alpuche-Aranda, M. Hernandez, M.A. Miller, N. Engl. J. Med. **361**, 674–679 (2009)
10. S.A. Webb, V. Pettila, I. Seppelt, R. Bellomo, M. Bailey, D.J. Cooper, M. Cretikos, A.R. Davies, S. Finfer, P.W. Harrigan, G.K. Hart, B. Howe, J.R. Iredell, C. McArthur, I. Mitchell, S. Morrison, A.D. Nichol, D.L. Paterson, S. Peake, B. Richards, D. Stephens, A. Turner, M. Yung, N. Engl. J. Med. **361**, 1925–1934 (2009)
11. D.N. Fisman, R. Savage, J. Gubbay, C. Achonu, H. Akwar, D.J. Farrell, N.S. Crowcroft, P. Jackson, N. Engl. J. Med. **361**, 2000–2001 (2009)
12. B. Cao, X.W. Li, Y. Mao, J. Wang, H.Z. Lu, Y.S. Chen, Z.A. Liang, L. Liang, S.J. Zhang, B. Zhang, L. Gu, L.H. Lu, D.Y. Wang, C. Wang, N. Engl. J. Med. **361**, 2507–2517 (2009)
13. K. Hancock, V. Veguilla, X. Lu, W. Zhong, E.N. Butler, H. Sun, F. Liu, L. Dong, J.R. DeVos, P.M. Gargiullo, T.L. Brammer, N.J. Cox, T.M. Tumpey, J.M. Katz, N. Engl. J. Med. **361**, 1945–1952 (2009)
14. D.C. Wiley, J.J. Skehel, Annu. Rev. Biochem. **56**, 365–394 (1987)
15. W. Gerhard, J. Yewdell, M.E. Frankel, R. Webster, Nature **290**, 713–717 (1981)
16. R.A. Childs, A.S. Palma, S. Wharton, T. Matrosovich, Y. Liu, W. Chai, M.A. Campanero-Rhodes, Y. Zhang, M. Eickmann, M. Kiso, A. Hay, M. Matrosovich, T. Feizi, Nat. Biotechnol. **27**, 797–799 (2009)
17. J.C. Krause, T.M. Tumpey, C.J. Huffman, P.A. McGraw, M.B. Pearce, T. Tsibane, R. Hai, C.F. Basler, J.E. Crowe Jr., J. Virol. **84**, 3127–3130 (2010)
18. C.J. Wei, J.C. Boyington, K. Dai, K.V. Houser, M.B. Pearce, W.P. Kong, Z.Y. Yang, T.M. Tumpey, G.J. Nabel, Sci. Transl. Med. **2**, 24ra21 (2010)
19. Y. Furuse, A. Suzuki, T. Kamigaki, H. Oshitani, Virol. J. **6**, 67 (2009)
20. S.K. Saxena, N. Mishra, R. Saxena, M.L. Swamy, P. Sahgal, S. Saxena, S. Tiwari, A. Mathur, M.P. Nair, J. Infect. Dev. Ctries. **4**, 1–6 (2009)
21. RCSB:PDB, http://www.rcsb.org/pdb/home/home.do. Accessed 07 Jul 2009
22. Jmol: an open-source Java viewer for chemical structures in 3D, http://www.jmol.org/
23. S.L. Pond, S.D. Frost, S.V. Muse, Bioinformatics **21**, 676–679 (2005)
24. Z. Yang, R. Nielsen, N. Goldman, A.M. Pedersen, Genetics **155**, 431–449 (2000)
25. G.J. Smith, D. Vijaykrishna, J. Bahl, S.J. Lycett, M. Worobey, O.G. Pybus, S.K. Ma, C.L. Cheung, J. Raghwani, S. Bhatt, J.S. Peiris, Y. Guan, A. Rambaut, Nature **459**, 1122–1125 (2009)
26. Team RDC R: a language and environment for statistical computing (2006). R Foundation for Statistical Computing, http://www.R-project.org.
27. MOE, Chemical Computing Group, Montreal, Canada (2007)
28. A. Jacobo-Molina, J. Ding, R.G. Nanni, A.D. Clark, X. Lu Jr., C. Tantillo, R.L. Williams, G. Kamer, A.L. Ferris, P. Clark et al., Proc. Natl. Acad. Sci. USA **90**, 6320–6324 (1993)
29. H. Huang, R. Chopra, G.L. Verdine, S.C. Harrison, Science **282**, 1669–1675 (1998)
30. M. Matrosovich, A. Tuzikov, N. Bovin, A. Gambaryan, A. Klimov, M.R. Castrucci, I. Donatelli, Y. Kawaoka, J. Virol. **74**, 8502–8512 (2000)
31. E. Nobusawa, T. Aoyama, H. Kato, Y. Suzuki, Y. Tateno, K. Nakajima, Virology **182**, 475–485 (1991)
32. S.J. Gamblin, L.F. Haire, R.J. Russell, D.J. Stevens, B. Xiao, Y. Ha, N. Vasisht, D.A. Steinhauer, R.S. Daniels, A. Elliot, D.C. Wiley, J.J. Skehel, Science **303**, 1838–1842 (2004)
33. J.A. Greenbaum, M.F. Kotturi, Y. Kim, C. Oseroff, K. Vaughan, N. Salimi, R. Vita, J. Ponomarenko, R.H. Scheuermann, A. Sette, B. Peters, Proc. Natl. Acad. Sci. USA **106**, 20365–20370 (2009)
34. A.J. Caton, G.G. Brownlee, J.W. Yewdell, W. Gerhard, Cell **31**, 417–427 (1982)
35. S.J. Baigent, J.W. McCauley, Bioessays **25**, 657–671 (2003)

36. I.T. Schulze, J. Infect. Dis. **176**(Suppl 1), S24–S28 (1997)
37. S.Y. Mir-Shekari, D.A. Ashford, D.J. Harvey, R.A. Dwek, I.T. Schulze, J. Biol. Chem. **272**, 4027–4036 (1997)
38. R. Xu, D.C. Ekiert, J.C. Krause, R. Hai, J.E. Crowe Jr., I.A. Wilson, Science **328**, 357–360 (2010)
39. H. Chen, X. Wen, K.K. To, P. Wang, H. Tse, J.F. Chan, H.W. Tsoi, K.S. Fung, C.W. Tse, R.A. Lee, K.H. Chan, K.Y. Yuen, J. Infect. Dis. **201**, 1517–1521 (2010)