RESEARCH ARTICLE

Computational Identification of Antigenicity-Associated Sites in the Hemagglutinin Protein of A/H1N1 Seasonal Influenza Virus

Xiaowei Ren[1]☯*, Yuefeng Li[2]☯, Xiaoning Liu[1], Xiping Shen[1], Wenlong Gao[1], Juansheng Li[1]

[1] Department of Epidemiology and Health Statistics, School of Public Health, Lanzhou University, Lanzhou, Gansu, China, [2] Center for Statistics and Information, National Health and Family Planning Commission, Beijing, China

☯ These authors contributed equally to this work.
* renxw@lzu.edu.cn

Abstract

The antigenic variability of influenza viruses has always made influenza vaccine development challenging. The punctuated nature of antigenic drift of influenza virus suggests that a relatively small number of genetic changes or combinations of genetic changes may drive changes in antigenic phenotype. The present study aimed to identify antigenicity-associated sites in the hemagglutinin protein of A/H1N1 seasonal influenza virus using computational approaches. Random Forest Regression (RFR) and Support Vector Regression based on Recursive Feature Elimination (SVR-RFE) were applied to H1N1 seasonal influenza viruses and used to analyze the associations between amino acid changes in the HA1 polypeptide and antigenic variation based on hemagglutination-inhibition (HI) assay data. Twenty-three and twenty antigenicity-associated sites were identified by RFR and SVR-RFE, respectively, by considering the joint effects of amino acid residues on antigenic drift. Our proposed approaches were further validated with the H3N2 dataset. The prediction models developed in this study can quantitatively predict antigenic differences with high prediction accuracy based only on HA1 sequences. Application of the study results can increase understanding of H1N1 seasonal influenza virus antigenic evolution and accelerate the selection of vaccine strains.

OPEN ACCESS

Citation: Ren X, Li Y, Liu X, Shen X, Gao W, Li J (2015) Computational Identification of Antigenicity-Associated Sites in the Hemagglutinin Protein of A/H1N1 Seasonal Influenza Virus. PLoS ONE 10(5): e0126742. doi:10.1371/journal.pone.0126742

Academic Editor: Zhiping Ye, US Food and Drug Administration, UNITED STATES

Received: October 16, 2014

Accepted: April 7, 2015

Published: May 15, 2015

Data Availability Statement: All relevant data are available in the paper and its Supporting Information files.

Funding: This study was supported by the National Natural Science Foundation of China (No.81102205) to XR and by the Fundamental Research Funds for the Central Universities (No.lzujbky-2012-155) to XR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Influenza is an infectious disease caused by the influenza virus. Seasonal influenza causes approximately 250,000 to 500,000 deaths per year worldwide [1]. Influenza A/H1N1, A/H3N2, and B viruses are currently the main circulating subtypes of seasonal influenza that are included in each year's influenza vaccine. After the 1918 Spanish influenza pandemic, the A/H1N1 influenza virus caused a number of outbreaks and epidemics until its disappearance from the human population in 1957 [2]. The H1N1 virus reappeared in the 1977 Russian influenza epidemic and began to exhibit seasonal circulation worldwide among the human population [3].

Vaccination is the most effective way to prevent influenza infection. The primary components of influenza vaccines have to be updated annually due to antigenic drift of influenza viruses. The influenza virus surface glycoprotein hemagglutinin (HA) is the primary target of neutralizing antibodies. The HA protein consists of two polypeptides, HA1 and HA2. The HA1 polypeptide plays a much more important role than HA2 in natural selection [4]. The gradual accumulation of amino acid substitutions in HA1 results in antigenic drift as antigenic properties change with time [5]. The seasonal influenza vaccine cannot offer effective protection against antigenically mismatched circulating strains. Antigenic differences between influenza strains are routinely measured by hemagglutination-inhibition (HI) assays that assess the ability of reference antiserum to prevent agglutination of red blood cells by influenza virus particles.

Smith *et al*. created an antigenic evolution map of influenza A/H3N2 virus using HI assay data and showed that antigenic evolution was more punctuated than genetic evolution, with a single amino acid change sometimes having a disproportionately large antigenic effect [6]. Li *et al*. also demonstrated that single HA mutations that alter both antigenicity and receptor binding avidity influenced influenza virus antigenic clustering [7]. Therefore, there are compelling reasons to identify antigenicity-associated sites in HA1. In turn, identification of these sites may offer insight into the antigenic drift of influenza viruses and accelerate the selection of vaccine strains. Sun *et al*. suggested that antigenicity-associated sites were not necessarily the same among subtypes of influenza A viruses because of different antibody binding sites [8]. For H3N2 seasonal influenza viruses, various studies have used statistical models to identify critical antigenic amino acid positions based on HI assays and HA1 sequences [9–12]. However, only a few studies have explored amino acid sites that drive the antigenic drift of H1N1 seasonal influenza virus using HI assay data. Recently, Huang *et al*. identified 41 H1N1 HA natural epitope residues based on 1572 HA sequences and 197 pairs of HA sequences with HI assays. Likelihood ratios used to correlate a single mutation at the amino acid position with an antigenic variant were utilized to establish the antigenic variant score [13]. However, recent studies have revealed that amino acid interactions play a key role in the process that antigenic site mutations modulate receptor-binding site (RBS) properties; hence, it is necessary to consider the joint effects of amino acid residues on antigenic drift when inferring the relationships between antigenic variants and amino acid changes [14].

In the present study, two widely used multivariate feature selection methods, Random Forest Regression (RFR) and Support Vector Regression based on Recursive Feature Elimination (SVR-RFE), were used to analyze the associations between amino acid changes in the HA1 polypeptide and antigenic variation based on HI assay data and HA1 sequences. These models identified the best combinations of amino acid sites that could quantitatively predict antigenic differences between different strains of H1N1 seasonal influenza with high prediction accuracy using HA1 sequences. Antigenic cartography methods were used to quantify and visualize the antigenic evolution of H1N1 seasonal influenza viruses from 1977 to 2008, and antigenic cluster transitions were found to be associated with important amino acid substitutions in the HA1 domain of HA. Finally, our computational approaches were validated with the H3N2 dataset to assess the performance when applied to H3N2 influenza viruses. We believe that the identification of antigenicity-associated sites will be helpful in better understanding the antigenic evolution of H1N1 seasonal influenza virus and facilitating influenza vaccine strain selection.

Materials and Methods

HI assay data

The HI assay data on the H1N1 seasonal influenza viruses used in this study were collected from related articles and documents published by the US Centers for Disease Control and Prevention and other collaborating centers of the World Health Organization (WHO) [15–22]. Under the premise of without changing the ratios between homologous and heterologous antibody titers, which were used to calculate antigenic distances, homologous antibody titers were standardized to equal 1280. The combined HI table was shown in S1 Table. The HI assay dataset contained 154 pairwise comparisons of 37 H1N1 viruses isolated from 1977 to 2008. The antigenic difference between two viruses was measured using the Archetti-Horsfall antigenic distance metric, which could control for receptor binding avidity variation between viral strains [7, 23]. The antigenic distance metric between strains A and B was defined using the following formula:

$$d(A,B) = \log_2\left(\sqrt{\frac{H^{AA}H^{BB}}{H^{BA}H^{AB}}}\right) \tag{1}$$

H^{XY} = HI titer of virus strain X (antigen) relative to antiserum raised against virus strain Y. H^{AA} and H^{BB} = homologous titers of two strains. H^{BA} and H^{AB} = heterologous titers against each other.

Antigenic distances greater than 2 were usually treated as antigenic variants [9].

HA sequence data

Thirty-seven HA1 amino acid sequences of H1N1 seasonal influenza viruses isolated from 1977 to 2008 were downloaded from the National Center for Biotechnology Information Influenza Virus Resource (Table 1) (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) [24]. All HA1 sequences were aligned with ClustalW and trimmed to the same length (327 amino acid residues) [25]. Each pairwise amino acid sequence alignment was converted to a 1 x 327 vector. For a specific HA1 sequence position, the value was denoted as 1 if the residue types of the two HA1 sequences were different (mutation); otherwise, its value was denoted as 0 (no mutation). One hundred and fifty-four pairwise amino acid sequence alignments were linked with pairwise antigenic distances for further analysis.

Validation data

The previously published A/H3N2 dataset by Smith *et al.* was used to validate the biological relevance of the results of our approaches [6]. The validation dataset consisted of 271 pairwise antigenic distances among 52 viruses isolated between 1968 and 2003. The HI table was shown in S2 Table.

Random forest regression

Random forests is a widely used machine learning algorithm that has been applied to classification and regression problems. It was introduced by Breiman in 2001 [26]. RFR is an ensemble of regression trees in which each tree is constructed on a bootstrap sample that is a subset of the original sample. At each splitting node of a tree, the candidate set of variables is a random subset of the explanatory variables. To reduce bias, no pruning step is performed; hence, all trees of the forest are maximal trees. The overall prediction of the forest is the average of predictions from all individual trees. For each tree, approximately one third of the data that are not included in the

Table 1. Full names and accession numbers of H1N1 seasonal influenza viruses from 1977 to 2008.

Full name	Accession number
A/WUHAN/371/95	CAC86625
A/BAYERN/7/95	CAD29944
A/BEIJING/262/95	ACF41867
A/BRAZIL/11/78	ABO38065
A/BRISBANE/59/07	ACA28844
A/BRISBANE/193/2004	ACD37424
A/CAMBODIA/0371/2007	ACI45444
A/CHILE/1/83	ABO38340
A/FLORIDA/13/07	ACF40117
A/FUKUSHIMA/141/2006	ACM17297
A/HONG_KONG/2652/2006	ACD37439
A/INDIA/6263/80	ABO38362
A/JIANGXI/160/2005	ACF76722
A/JOHANNESBURG/82/96	CAD29943
A/KENTUCKY/1/2005	ABI96135
A/KENTUCKY/02/2006	ABU86800
A/NEW_CALEDONIA/9/2004	ABQ09837
A/NEW_CALEDONIA/20/99	AFO65027
A/PHILIPPINES/673/2006	ACD37433
A/SHENZHEN/227/95	AAP34325
A/SICHUAN/4/88	AAA43231
A/SINGAPORE/6/86	ABO38395
A/SINGAPORE/14/2004	ABQ09838
A/SOLOMON_ISLANDS/03/2006	ABU50586
A/SOUTH_DAKOTA/06/2007	AFM72510
A/TAIWAN/1/86	CAA35097
A/TEXAS/36/91	ABD60955
A/USSR/90/77	AFM73477
A/VICTORIA/500/2006	ABQ09960
A/VIRGINIA/01/2006	ABI96152
A/ENGLAND/333/80	X00031
A/CHILE/4795/00	AFO66147
A/FUJIAN/156/00	AFQ90525
A/HONG KONG/1870/2008	AFM72587
A/MALAYSIA/100/2006	ABQ09959
A/MOSCOW/13/98	AFQ90529
A/NEIMENGGU/52/2002	AFQ90530

doi:10.1371/journal.pone.0126742.t001

bootstrap sample are termed the “out-of-bag” (OOB) sample. The OOB sample is used as a testing set for that tree to estimate the prediction performance and then to evaluate variable importance. There are three parameters that need to be determined for RFR: *ntree*—the number of trees in the forest; *nodesize*—the minimum size of the terminal nodes; and *mtry*—the number of variables randomly sampled as candidates at each split. In this study, the parameters *ntree* and *nodesize* were set to their default values (*ntree* = 500 and *nodesize* = 5 for regression). The most important parameter (*mtry*) was tuned to achieve optimal predictive performance and increase the statistical power of the algorithm to detect true antigenicity-associated sites.

Cases of *mtry* equaling *p* (equivalent to bagging), 2/3*p*, *p*/2, *p*/3, *p*/4, and sqrt(*p*) were considered, where *p* was the number of variables in the data set. The permutation-based "mean of squared residuals (MSE) reduction" was used as the random forest importance criterion to give a ranking of variable importance, and the variables with high ranks were considered as potentially associated with antigenicity. In this study, a variable reduction wrapper algorithm was employed to find the best putative combinations of amino acid sites [27]. This measure of importance was used to rank the variables from most to least important, and the RFR model's performance could be monitored as more and more of the least important variables were iteratively removed. A 5-fold cross-validation (CV) procedure was used to evaluate the predictive performance of the RFR model. In the variable selection process, variable importance was not recalculated at each step because Svetnik *et al. 2004* reported that severe overfitting results from recalculating variable importance. The CV test set MSE was used to assess the predictive performance of the RFR model. The smaller the MSE, the better the performance will be. The final RFR model was chosen by using the corrected Akaike information criterion (AICc) [28] to prevent overfitting. The RFR algorithm and the nested cross-validation procedure were implemented by the randomForest R package [29].

Support vector regression based on recursive feature elimination

A support vector machine is a supervised data mining method based on statistical learning theory used for classification and regression. The basic idea of support vector regression was formulated by Vapnik *et al.* in 1995 [30]. Instead of minimizing the observed training error in the traditional regression procedure, support vector regression attempts to minimize the generalized error bound to achieve performance in which the generalization error bound is the combination of the training error and a regularization term that controls the complexity of the hypothesis space [31]. SVM-RFE is a popular embedded feature selection method. It was first proposed by Guyon *et al.* in 2002 to perform gene selection for binary classification problems [32]. SVM-RFE uses weight magnitude as the ranking criterion and generates a ranking of features using the backward feature elimination procedure, which starts with all the features and then recursively removes one or more of the least important features at a time. In this study, the ε-SVR-RFE algorithm was used to select a subset of amino acid sites that could contribute the most to the antigenic variation of H1N1 seasonal influenza virus as an alternative machine learning method to the RFR model. The 5-fold cross-validated MSE was adopted to evaluate the prediction performance of SVR models as more and more of the least important amino acid sites were recursively removed. The SVR model with the smallest AICc value was considered as the final selected model. The parameter ε and kernel function in SVR-RFE were set to 0.01 and radial basis function (RBF) kernel, respectively. The cost parameter *C* and parameter γ of the RBF kernel were optimized by extensive grid search. The search was conducted in the following ranges: $C - 2^{-8}$ to 2^{8} with step size 0.1; $\gamma - 2^{-5}$ to 2^{5} with step size 0.1. A 5-fold cross-validation was used in our experiments to determine optimum parameters. The SVR algorithm was implemented by the Matlab version of LibSVM provided by Chang and Lin [33]. The parameter optimization function was implemented using the Matlab toolbox Libsvm-FarutoUltimate3.1 provided by Li [34]. SVR-RFE was written in Matlab.

Antigenic cartography

The quantitative analyses of the antigenic properties of H1N1 seasonal influenza viruses collected from 1977 to 2008 were performed using antigenic cartography methods as described previously by Smith *et al.* for human H3N2 viruses [6]. The HI assay data were used to construct two-dimensional (2D) antigenic maps in which the distance between points represents the antigenic distance as measured by a HI assay. One unit of antigenic distance on the antigenic map corresponds to a two-fold difference in the serological assay. The web-based software for Antigenic Cartography is available at http://www.antigenic-cartography.org/.

Results

Data set

Amino acid changes were identified in 77 positions among the 154 pairwise comparisons of H1N1 HA1 sequences by excluding 250 positions that had no mutations (S3 Table). The amino acid positions that underwent the same change (correlation coefficient = 1) were noted as one variable. These 77 positions were classified into 60 variables based on their collinearity. Among the 154 pairwise comparisons, 83 (54%) had an antigenic distance ≥ 2 (i.e., antigenic variant), and 71 (46%) had an antigenic distance < 2 (i.e., similar antigenicity). The H3N2 validation dataset was treated in the same way. Sixty-eight variables that were within positions 109 to 301 of the HA1 region according to Koel *et al.* were obtained [35].

Antigenicity-associated sites identified using the RFR algorithm

RFR was used to quantitatively correlate amino acid changes in the HA1 region with antigenic distances and to identify the best combinations of amino acid sites that contributed to antigenic drift of H1N1 seasonal influenza viruses. The CV test set MSE curve based on the variation of the number of variables selected was plotted for different choices of *mtry* (Fig 1). The plot showed that *mtry*, *p/4*, performed the best, but the other choices were nearly as robust. The results suggested that the RFR model with 18 variables (*mtry* = *p/4*) achieved the smallest AICc value of 0.8115. The predictive proportion of variance explained by the final RFR model was 0.8353, suggesting that the model achieved a good predictive performance. In Table 2, the 18 variables ranked in order of importance for the final RFR model were reported. Twenty-three amino acid positions were identified as responsible for the antigenic variation of H1N1 seasonal influenza viruses.

Antigenicity-associated sites identified using the SVR-RFE

The SVR-RFE algorithm was performed to find the subset of amino acid residues that contributed the most to antigenic variation of H1N1 seasonal influenza virus. Fig 2 shows the cross-validation MSE against the number of variables used at each step of removing the least important variable. The minimum AICc value of 0.5826 was obtained at the 17-variable level by the SVR model (RBF kernel with parameters *C* and γ optimized to 4.5948 and 0.1015, respectively). The predictive proportion of variance explained rose up to 0.8610, which was 0.0257 higher than the RFR model. The final variable subset selected by the SVR-RFE algorithm is listed in Table 3 in order of variable importance. Twenty sites were identified as the best combinations of amino acid sites that could drive the antigenic drift of H1N1 seasonal influenza viruses.

Antigenic map of human H1N1 seasonal influenza viruses from 1977 to 2008

Antigenic cartography methods were used to map the antigenic evolution of human H1N1 seasonal influenza viruses from 1977 to 2008 (Fig 3). The antigenic evolution of human H1N1 seasonal influenza viruses was significantly slower than the antigenic evolution of human H3N2 viruses when compared with the map published by Smith *et al.* The antigenic evolution of H1N1 seasonal influenza viruses appeared to be punctuated rather than gradual. Two large jumps were observed in the antigenic phenotype from the ‘80–83’ strains to the ‘86–98’ strains and from the '86–98' strains to the '99–08' strains, and one small jump was observed from the '77–78' strains to the '80–83' strains. Fig 3 summarizes the cluster-transition amino acid substitutions that are most likely responsible for the antigenic differences between antigenic clusters.

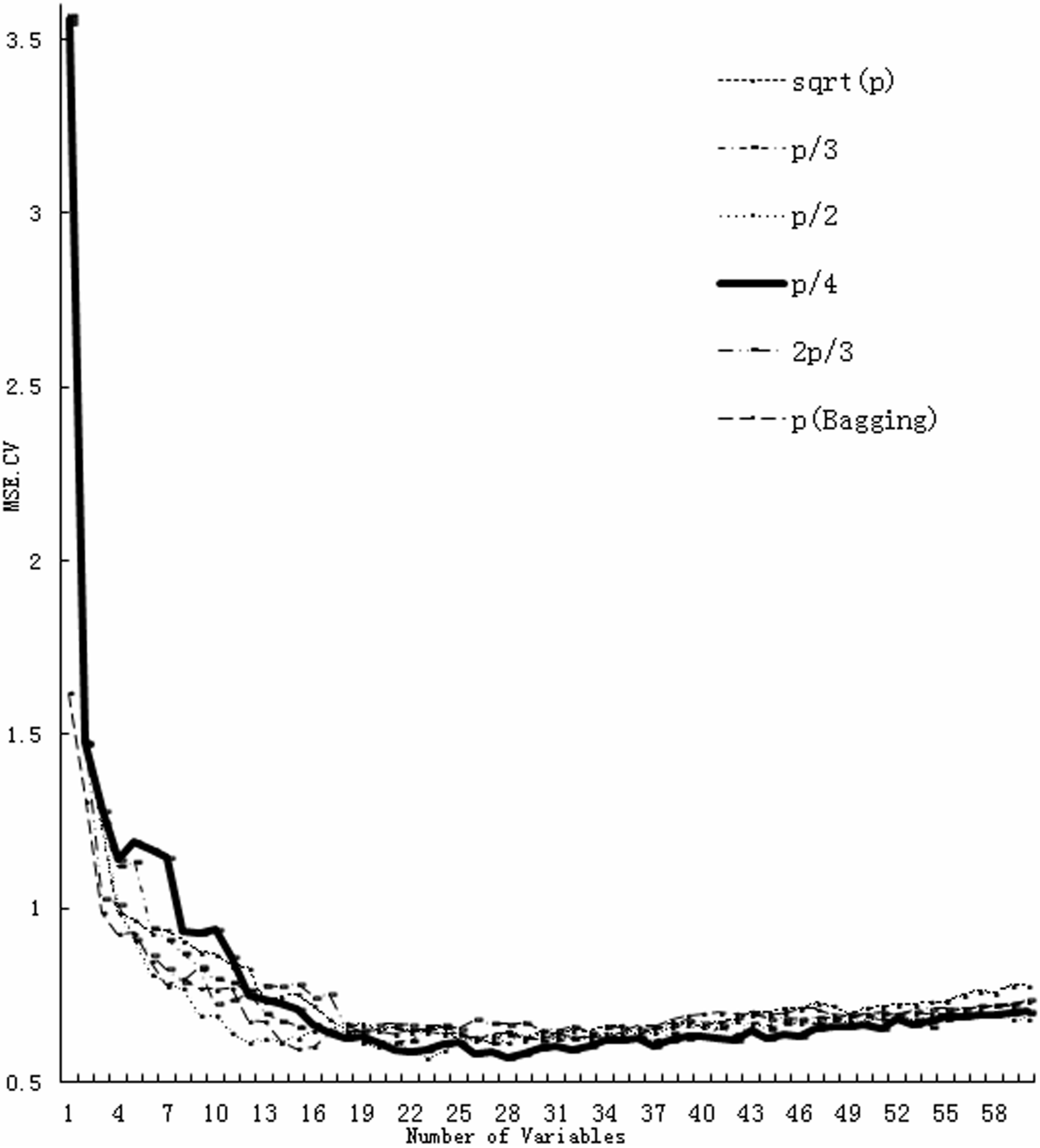

Fig 1. The cross-validation MSE of RFR against the number of variables selected using different *mtry* functions.

doi:10.1371/journal.pone.0126742.g001

Table 2. Antigenicity-associated sites of H1N1 identified using the random forest regression algorithm.

Order of variable importance	Amino acid position		
1	141		
2	130		
3	43		
4	54	127	193
5	186		
6	80	271	
7	71		
8	36		
9	190		
10	194		
11	163		
12	128		
13	187		
14	189		
15	125		
16	121	205	
17	321		
18	133	191	

doi:10.1371/journal.pone.0126742.t002

Model Validation with the H3N2 dataset

Our proposed RFR and SVR-RFE approaches identified 12 and 18 antigenicity-associated sites of H3N2 influenza viruses, respectively. The importance of each variable was assessed. The variables ranked in order of importance for the RFR model and the SVR-RFE model were reported in Tables 4 and 5, respectively. The amino acid substitutions at seven positions (positions 145, 155, 156, 158, 159, 189, and 193), which were responsible for A/H3N2 antigenic cluster transitions over the 35-year period, had been experimentally validated by Koel et al.[35]. The top ten ranked variables recognized by the RFR model included the seven experimentally validated sites, while there were six experimentally validated sites within the top ten ranked variables identified by the SVR-RFE approach. Only one experimentally validated site, position 156, was not recognized by the SVR-RFE algorithm. Both of positions 133 and 135 that had been experimentally validated as the accessory substitutions by Koel *et al* were also respectively identified by both our approaches. Then the experimentally validated seven positions were used to construct the RFR and SVR model. The predictive proportions of variation explained by the 7-features RFR and 7-features SVR model were, respectively, 0.8447 and 0.8496. The high rankings of experimentally validated positions in both subsets and their significant contribution to explaining variation confirmed the effectiveness of our proposed approaches in capturing important antigenicity-associated sites and predicting antigenic changes.

Discussion

The WHO annually recommends an influenza vaccine composition for the coming influenza season. Antigenic characterization based on the HI assay, which is labor-intensive and time-consuming, is the primary procedure for influenza vaccine strain selection. It is of considerable importance to detect new antigenic changes occurring in HA protein when updating vaccine compositions. The punctuated nature of antigenic evolution of influenza virus suggests that a

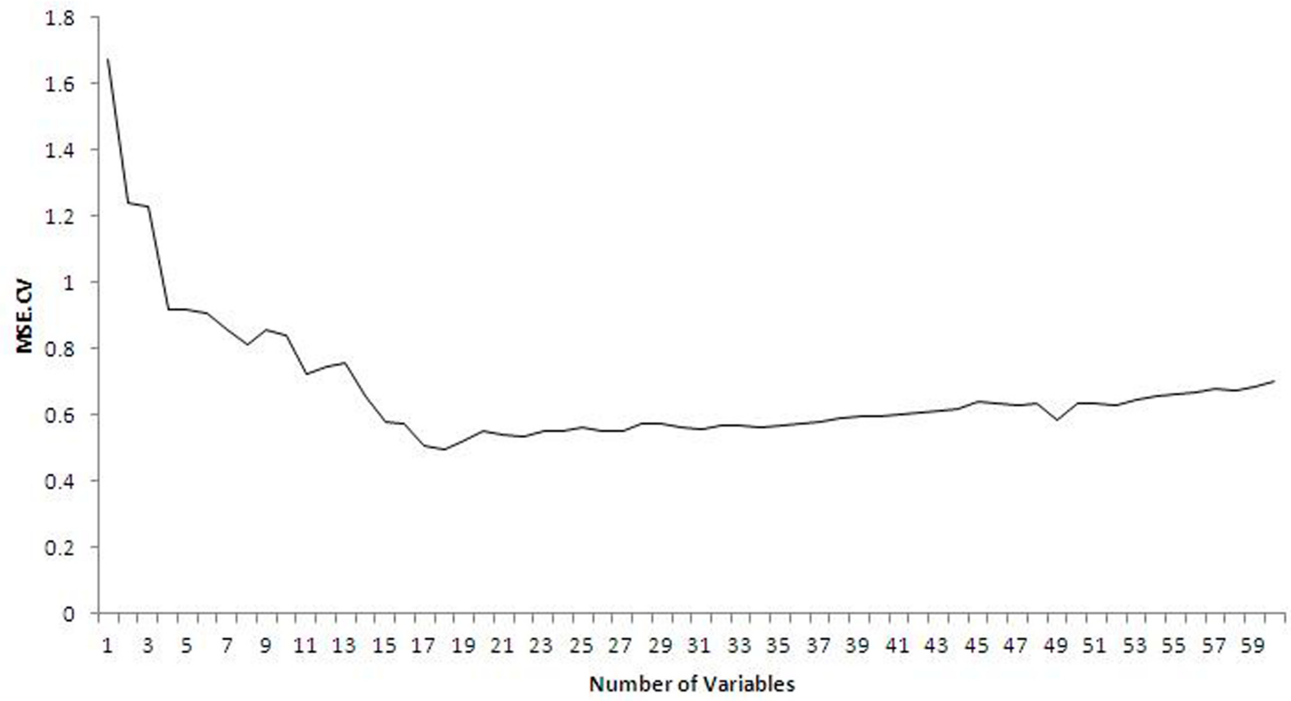

Fig 2. The cross-validation MSE of SVR against the number of variables used.

doi:10.1371/journal.pone.0126742.g002

relatively small number of genetic changes or combinations of genetic changes may drive changes in antigenic phenotype [35].

Table 3. Antigenicity-associated sites of H1N1 identified using the support vector regression based on recursive feature elimination.

Order of variable importance	Amino acid position		
1	130		
2	54	127	193
3	43		
4	141		
5	190		
6	160		
7	121	205	
8	71		
9	273		
10	321		
11	125		
12	96		
13	277		
14	69		
15	187		
16	269		
17	57		

doi:10.1371/journal.pone.0126742.t003

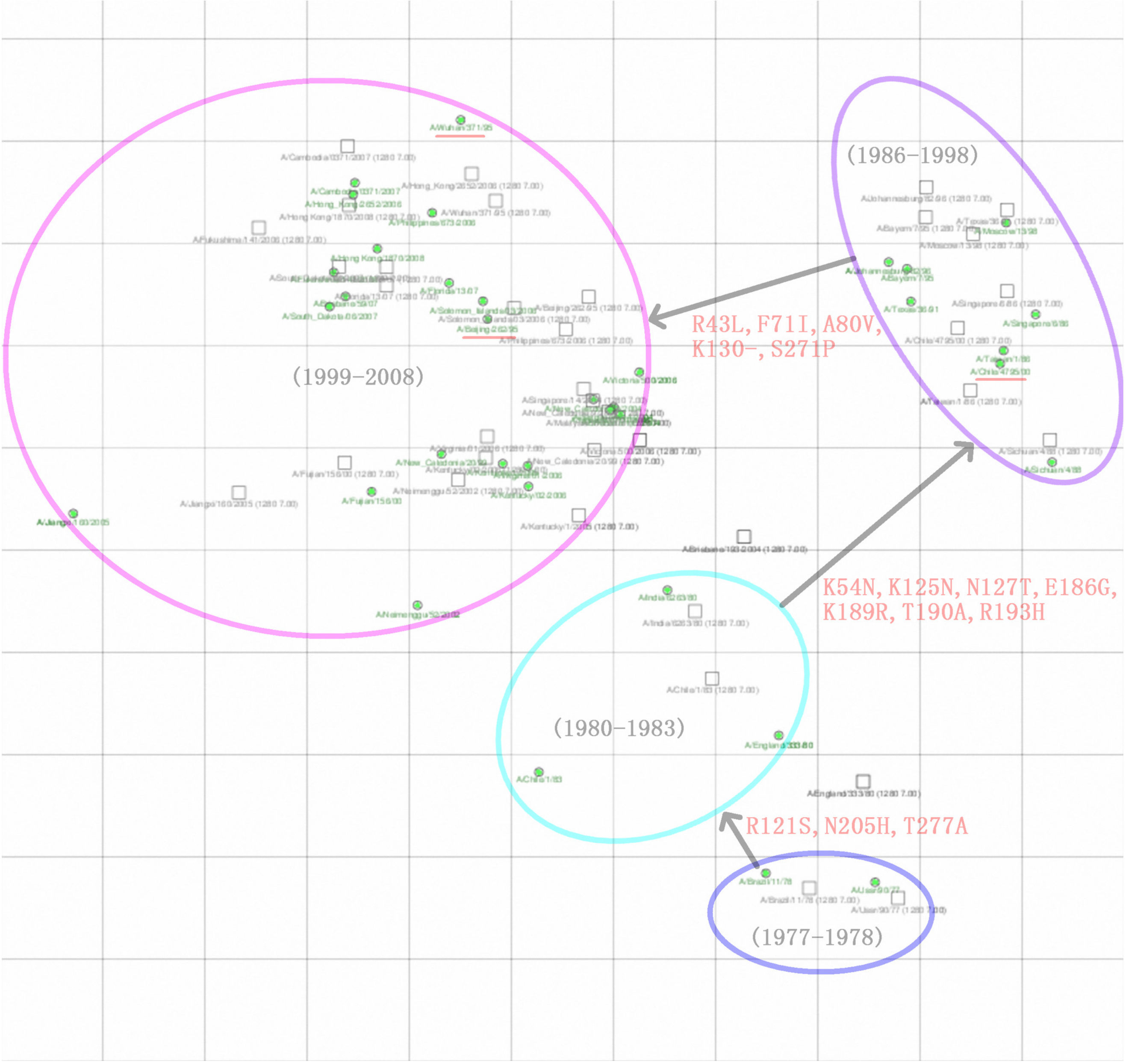

Fig 3. Antigenic map of human H1N1 seasonal influenza viruses from 1977 to 2008. The relative positions of strains (green circles) and antisera (uncolored squares) are adjusted such that the distances between strains and antisera on the map represent the corresponding HI measurements with the least error. One unit (grid) corresponds to a two-fold dilution of antiserum in the HI assay. The cluster-transition amino acid substitutions are shown in red.

doi:10.1371/journal.pone.0126742.g003

In this study, RFR and SVR-RFE were used to identify amino acid sites associated with HA antigenicity of H1N1 seasonal influenza viruses on the basis of HI assay data and HA1 sequences. The methods used in the study are different from those of previous studies because the algorithms consider the joint effects of amino acid residues on antigenic drift when deriving antigenicity-associated sites. The results showed that CV test set MSE curves decreased sharply at first and then increased slowly with respect to the number of variables used. This trend might be attributed to the combined effects of amino acid substitutions [36].

Table 4. Antigenicity-associated sites of H3N2 identified using the random forest regression algorithm.

Order of variable importance	Amino acid position
1	**189**
2	*133*
3	**193**
4	**145**
5	**155**
6	144
7	**159**
8	216
9	**156**
10	**158**
11	163
12	*135*

Experimentally validated sites were marked as bold. Accessory substitutions were marked as italic.

doi:10.1371/journal.pone.0126742.t004

The RFR and SVR-RFE algorithms identified 23 and 20 antigenicity-associated sites, respectively. Interestingly, thirteen amino acid residues overlapped between these two amino acid subsets. The first four ranked variables recognized by these two approaches contained the same six amino acid positions (43, 54, 127, 130, 141 and 193). With the RFR and SVR-RFE approaches, 2 and 5 new antigenicity-associated residues were identified, respectively, that were not covered by the natural epitope residues proposed by Huang *et al*. (Table 6). From the

Table 5. Antigenicity-associated sites of H3N2 identified using the support vector regression based on recursive feature elimination.

Order of variable importance	Amino acid position	
1	**189**	
2	*133*	
3	**155**	
4	**193**	
5	**158**	
6	216	
7	144	
8	285	
9	**159**	
10	**145**	
11	137	
12	196	
13	129	132
14	271	
15	131	
16	175	
17	*135*	

Experimentally validated sites were marked as bold. Accessory substitutions were marked as italic.

doi:10.1371/journal.pone.0126742.t005

Table 6. Comparison of amino acid positions related to antigenic variation of H1N1 seasonal influenza viruses identified by current and previous studies.

RFR[a]	SVR-RFE[b]	natural epitope residues[c]	antigenic cartography[d]	antigenic epitope regions[e]
		35		
36		36		
43	43	43	43	
		47		
54	54	54	54	
	57			
	69	69		
71	71	71	71	Cb
		73		Cb
80		80	80	
		82		
		94		
	96			
121	121	121	121	
125	125	125	125	Sa
127	127	127	127	
128		128		
130	130	130	130	
133		133		
141	141	141		Ca2
		146		
		153		Sa
	160	160		Sa
163		163		Sa
		183		
186		186	186	Sb
187	187			Sb
189		189	189	Sb
190	190	190	190	Sb
191		191		Sb
193	193	193	193	Sb
194		194		Sb
205	205	205	205	Ca1
		209		
		216		
		222		Ca2
		224		
		267		
	269			
271		271	271	
	273	273		
		274		
	277	277	277	
		295		
		310		

(*Continued*)

Table 6. (*Continued*)

RFR[a]	SVR-RFE[b]	natural epitope residues[c]	antigenic cartography[d]	antigenic epitope regions[e]
321	321			

[a] Twenty-three antigenicity-associated sites identified by random forest regression.

[b] Twenty antigenicity-associated sites identified by support vector regression based on recursive feature elimination.

[c] Forty-one natural epitope residues identified by Huang *et al*. [13].

[d] Fifteen cluster-difference substitutions revealed by antigenic cartography.

[e] Five antigenic epitope regions described by Brownlee and Fodor [37].

doi:10.1371/journal.pone.0126742.t006

antigenic map of human H1N1 seasonal influenza viruses from 1977 to 2008, 15 amino acid positions were found that were likely to contribute to the antigenic difference between clusters. Surprisingly, the 23 and 20 antigenicity-associated sites cover 93.33% (14/15) and 73% (11/15) of cluster-differentiating mutations, respectively. The high ratio overlap reveals that the models used in this study are reliable. However, only 7 of these 15 cluster-difference substitutions overlapped with the five antigenic epitope regions (Sa, Sb, Ca1, Ca2, and Cb) previously described by Brownlee and Fodor [37]. The low consistency reflects the different recognition mechanisms of antigenic epitopes between natural selection at the population level and neutralizing monoclonal antibodies (MAbs) in the laboratory. There were also limited antigenic differences between the '99–06' strains and the '06–08' strains, which might be due to a single amino acid substitution at position 141 [35, 38, 39, 40].

The H3N2 dataset was used to further validate the biological relevance of the results of our approaches. The RFR approach has been shown to be more effective in identifying the experimentally validated sites than the SVR-RFE algorithm, but the SVR model demonstrated better predictive performance compared to the RFR model. The final models developed in this study can perform quantitative prediction of antigenic differences between different strains of H1N1 based only on HA1 sequences and serve as an initial screening tool for antigenic variants before the labor-intensive and time-consuming HAI assay.

In conclusion, the proposed RFR and SVR-RFE approaches, which consider the joint effects of amino acid residues on antigenic drift, identified 23 and 20 antigenicity-associated sites of HA1 of H1N1 seasonal influenza viruses, respectively. The results obtained in this study can aid in better understanding of the antigenic evolution of H1N1 seasonal influenza viruses and accelerate the selection of vaccine strains.

Supporting Information

S1 Table. The combined H1N1 HI table.
(XLS)

S2 Table. The H3N2 HI table.
(XLS)

S3 Table. Amino acid changes in HA1 of H1N1 seasonal influenza viruses from 1977 to 2008.
(DOC)

Author Contributions

Conceived and designed the experiments: XR. Analyzed the data: XR YL. Contributed reagents/materials/analysis tools: JL XL XS WG. Wrote the paper: XR YL.

References

1. WHO. Influenza (Seasonal). Fact sheet N°211. Mar 2014. Available: http://www.who.int/mediacentre/factsheets/fs211/en/. Accessed 8 September 2014.
2. Kilbourne ED. Influenza pandemics of the 20th century. Emerg Infect Dis. 2006; 12: 9–14. PMID: 16494710
3. Nakajima K, Desselberger U, Palese P. Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. Nature. 1978; 274: 334–339. PMID: 672956
4. Chi XS, Bolar TV, Zhao P, Tam JS, Rappaport R, Cheng SM. Molecular evolution of human influenza A/H3N2 virus in asia and europe from 2001 to 2003. J Clin Microbiol. 2005; 43: 6130–6132. PMID: 16333111
5. Blackburne BP, Hay AJ, Goldstein RA. Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog. 2008; 4: e1000058. doi: 10.1371/journal.ppat.1000058 PMID: 18451985
6. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. Science. 2004; 305: 371–376. PMID: 15218094
7. Li Y, Bostick DL, Sullivan CB, Myers JL, Griesemer SB, Stgeorge K, et al. Single hemagglutinin mutations that alter both antigenicity and receptor binding avidity influence influenza virus antigenic clustering. J Virol. 2013; 87: 9904–9910. doi: 10.1128/JVI.01023-13 PMID: 23824816
8. Sun H, Yang J, Zhang T, Long LP, Jia K, Yang G, et al. Using sequence data to infer the antigenicity of influenza virus. MBio. 2013; 4: e00230–13. doi: 10.1128/mBio.00230-13 PMID: 23820391
9. Lee MS, Chen JSE. Predicting antigenic variants of influenza A/H3N2 viruses. Emerg Infect Dis. 2004; 10: 1385–1390. PMID: 15496238
10. Lee MS, Chen MC, Liao YC, Hsiung CA. Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses. Vaccine. 2007; 25: 8133–8139. PMID: 17950961
11. Liao YC, Lee MS, Ko CY, Hsiung CA. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. Bioinformatics. 2008; 24: 505–512. doi: 10.1093/bioinformatics/btm638 PMID: 18187440
12. Huang JW, King CC, Yang JM. Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. BMC Bioinformatics. 2009; 10: S41. doi: 10.1186/1471-2105-10-S1-S41 PMID: 19208143
13. Huang JW, Lin WF, Yang JM. Antigenic sites of H1N1 influenza virus hemagglutinin revealed by natural isolates and inhibition assays. Vaccine. 2012; 30: 6327–6337. doi: 10.1016/j.vaccine.2012.07.079 PMID: 22885274
14. Soundararajan V, Zheng S, Patel N, Warnock K, Raman R, Wilson IA, et al. Networks link antigenic and receptor-binding sites of influenza hemagglutinin: Mechanistic insight into fitter strain propagation. Sci Rep. 2011; 1:200. doi: 10.1038/srep00200 PMID: 22355715
15. Ndifon W, Dushoff J, Levin SA. On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness. Vaccine. 2009; 27: 2447–2452. doi: 10.1016/j.vaccine.2009.02.047 PMID: 19368786
16. McDonald NJ, Smith CB, Cox NJ. Antigenic drift in the evolution of H1N1 influenza A viruses resulting from deletion of a single amino acid in the haemagglutinin gene. J Gen Virol. 2007; 88: 3209–3213. PMID: 18024888
17. Centers for Disease Control and Prevention. Information for the vaccines and related biological products advisory committee, CBER, FDA. 2004. pp. 9.
18. Centers for Disease Control and Prevention. Preliminary information for the vaccines and related biological products advisory committee, CBER, FDA. 2006. pp. 9.
19. Centers for Disease Control and Prevention. Preliminary information for the vaccines and related biological products advisory committee, CBER, FDA. 2008. pp. 13.
20. Centers for Disease Control and Prevention. February 18–19 VRPBAC meeting briefing materials. 2009. pp. 12.
21. WHO. Recommended composition of influenza virus vaccines for use in 2000. WHO Wkly Epidemiol Rec. 1999; 74: 321–325. PMID: 10530085
22. WHO collaborating centre for reference and research on Influenza, Melbourne, Australia. Annual Report. 2006. pp. 27.
23. Archetti I, Horsfall FL Jr. Persistent antigenic variation of influenza A viruses after incomplete neutralization in ovo with heterologous immune serum. J Exp Med. 1950; 92: 441–462. PMID: 14778924

24. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the national center for biotechnology information. J Virol. 2008; 82: 596–601. PMID: 17942553

25. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-speciic gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22: 4673–4680. PMID: 7984417

26. Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32.

27. Svetnik V, Liaw A, Tong C, Wang T. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T, editors. June 2004: proceedings; Cagliari, Italy. New York: Springer; pp. 334–343.

28. McQuarrie ADR, Tsai CL. Regression and Time Series Model Selection. Singapore: World Scientific; 1998.

29. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2: 18–22. PMID: 12028823

30. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer; 1995.

31. Basak D, Pal S, Patranabis DC. Support Vector Regression. Neural Information Processing-Letters and Reviews. 2007; 11: 203–224.

32. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002; 46: 389–422.

33. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf. Accessed 8 September 2014.

34. Li Y. LIBSVM-farutoUltimateVersion a toolbox with implements for support vector machines based on libsvm. 2011. Available: http://www.matlabsky.com/thread-17936-1-1.html. Accessed 8 September 2014.

35. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GC, Vervaet G, et al. Substitutions near the receptor binding site determine major antigenic change during inluenza virus evolution. Science. 2013; 342: 976–979. doi: 10.1126/science.1244730 PMID: 24264991

36. Chen SH, Sun J, Dimitrov L, Turner AR, Adams TS, Meyers DA, et al. A Support Vector Machine Approach for Detecting Gene-Gene Interaction. Genet Epidemiol. 2008; 32: 152–167. PMID: 17968988

37. Brownlee GG, Fodor E. The predicted antigenicity of the haemagglutinin of the 1918 Spanish influenza pandemic suggests an avian origin. Philos Trans R Soc Lond B Biol Sci. 2001; 356: 1871–1876. PMID: 11779386

38. Ren XW, Ju LW, Yang JX, Lv XH, Jiang LF, Zhao NQ, et al. Antigenic and genetic variation in the hemagglutinins of H1N1 and H3N2 human influenza A viruses in the Shanghai area from 2005 to 2008. J Med Virol. 2011; 83: 1113–1120. doi: 10.1002/jmv.22078 PMID: 21520138

39. WHO collaborating centre for reference and research on Influenza, London. Interim Report. 2007. pp. 3.

40. Rossi T. Molecular epidemiology of influenza viruses in three consecutive epidemic seasons, 2005–2007. Ph.D. Thesis, The University of Trieste. 2008. Available: http://www.openstarts.units.it/dspace/handle/10077/2643?mode = full. Accessed 8 September 2014.